CONCEPTS OF ELECTRONICS

Heathkit

ZENITH

Educational Systems

CONCEPTS OF ELECTRONICS

Book 1

CONTENTS

INTRODUCTION

This is Heathkit Learning Publications Textbook, EB-6140, Concepts of Electronics. It is designed to give you a thorough foundation in the basic and intermediate principles of electronics. Starting at the very beginning, it answers the question: "What is an electron?" This is an important question because electronics is the science that studies and controls the behavior of electrons. Therefore, you must be familiar with the electron and its behavior to fully understand electronics.

As you progress through this learning program, you will examine the electron's behavior in direct current (DC) and alternating current (AC) circuits. You will study diodes, transistors, and integrated circuits. You will also study power supplies, amplifiers, oscillators, digital electronics, and digital computers.

You can get a good idea of the scope of this program by reviewing the "Course Outline" that follows. As you can see, this is a very comprehensive course with a substantial amount of material. To enliven and enrich your learning experiences, a total of twenty-six Experiments are included in the "Student Workbook." These experiments will reinforce the theory of electronics with practical applications.

How do you gauge your learning? Let the "**objectives**" be your guide. These carefully constructed objectives are the framework for the course. When you can meet all of the objectives, you have satisfied the requirements of the course. You'll find two types of objectives in this course: broad, "**Course Objectives**" are listed following this introduction. More specific "**Unit Objectives**" are listed near the front of each unit. When you can satisfy these unit objectives, you've learned everything that was intended from the units.

COURSE OBJECTIVES

When you have completed this course, you will be able to:

1. Discuss basic and intermediate electronics theory and techniques.

2. Construct complex electronic circuits using only a schematic diagram as a guide.

3. Use a multimeter and oscilloscope to measure electronic parameters.

4. Use transistors, diodes, and integrated circuits to construct both basic and intermediate electronic circuits.

5. Construct and analyze the operation of power supplies, amplifiers, and oscillators.

6. Construct and analyze digital electronic circuits.

7. Discuss computer concepts and principles.

COURSE OUTLINE

Unit 1 **Direct Current**
   I.   Introduction
   II. Unit Objectives
   III. The Basis of Electronics
        A.   Atoms.
        B.   Electrons, Protons, and Neutrons.
        C.   Electrostatics.
        D.   Coulomb's Law.
        E.   The Ion.
        F.   The Coulomb.
        G.   Powers of Ten and Scientific Notation.
   IV. Current
        A.   Freeing Electrons.
        B.   Conductors and Insulators.
        C.   The Battery.
        D.   Random Drift and Directed Drift.
        E.   The Electric Circuit.
        F.   The Ampere.
   V.   Voltage
        A.   Electromotive Force.
        B.   Potential Difference.
        C.   Voltage.
        D.   Battery Connections.
        E.   Voltage Rises and Voltage Drops.
        F.   Concept of Ground.
   VI. Resistance
        A.   The Ohm.
        B.   Resistors.
        C.   Variable Resistors.
        D.   Resistor Connections.
        E.   Conductance.
   VII. Ohm's Law
        A.   Finding Current.
        B.   Finding Current in Parallel Circuits.
        C.   Finding Voltage.
        D.   Finding Resistance.
        E.   Summary.
   VIII. Power
        A.   The Watt.
        B.   Power, Current, and Voltage.
        C.   Power Dissipation in Resistors.
        D.   The Wheel Diagram.

X.   RLC Circuits
     A.   Series RLC Circuits.
     B.   Parallel RLC Circuits.
     C.   Series Resonance.
     D.   Parallel Resonance.
     E.   Circuit Q.
     F.   Filters.
          1.   Bandstop Filter.
          2.   Bandpass Filter.
          3.   Low Pass Filter.
          4.   High Pass Filter.

XIV

Unit 1

DIRECT CURRENT

CONTENTS

INTRODUCTION

Electronics is that science which controls the behavior of **electrons** so that some useful function is performed. As this definition implies, the electron is vitally important to electronics. In fact, the word **electronics** is derived from the word **electron**. Electricity comes to our homes and offices by the movement of electrons through wires. Actually, electric current is nothing more than the movement of electrons. Obviously then, to understand electronics, you must first understand the nature of the electron. In this unit, you will see what the electron is, how it behaves, and how you can use it to perform useful jobs. You will also learn how to measure the flow of electrons.

The "Unit Objectives" listed on the next page state exactly what you are expected to learn from this unit. Study this list now and refer to it often as you study the text.

UNIT OBJECTIVES

When you complete this unit, you should be able to:

1. Define: atom, electron, proton, neutron, nucleus, and ion.

2. State the electrical charge associated with the: electron, proton, neutron, atom, and ion.

3. State Coulomb's law.

4. Define: insulator, conductor, current flow, schematic diagram, coulomb, and ampere.

5. Identify the schematic symbols of basic electronic components.

6. Define: electromotive force, potential difference, voltage, and volt.

7. State the effects of connecting batteries in series, parallel, and series-parallel.

8. Explain the difference between a voltage rise and voltage drop.

9. Define: resistance, ohm, and conductance.

10. Name three types of resistor construction.

11. Define potentiometer.

12. State how the total resistance is affected by connecting resistors in series, in parallel, and in series-parallel.

13. Write the three equation forms of Ohm's law.

14. Find current when given voltage and resistance.

15. Find voltage when given current and resistance.

16. Find resistance when given voltage and current.

17. Define power and watt.

18. Write the basic power formula.

19. Find power when given voltage and current.

20. Find voltage when given power and current.

21. Find current when given power and voltage.

22. Find power when given voltage and resistance.

23. Find power when given current and resistance.

24. State the basic law of magnetism.

25. Define magnetomotive force, ampere-turn, and permablility.

26. Name the instruments used to measure current, voltage, and resistance.

27. Demonstrate the correct way to use a ammeter, voltmeter, and ohmmeter.

28. Define ground.

THE BASIS OF ELECTRONICS

Electrons are tiny particles which carry the energy to light our homes, cook our food, operate our electronic devices, and do much of our work. Therefore, an understanding of the electron is vitally important to an understanding of electronics. Controlling electrons is what electronics is all about.

To understand what an electron is, we must investigate the make-up of **matter**. Matter is generally described as anything which has weight and occupies space. Thus, the earth and everything on it is classified as matter.

Atoms

All matter is made up of atoms. If we were to take any material, for example water, and start dividing it into parts over and over again, soon we would get a particle so small that it could not be divided further and still be water. The smallest particle of water which still retains the characteristics of water is called a **molecule**. This is shown in Figure 1-1. This molecule of water can be divided into still smaller particles, but these particles will not be water. These particles are called atoms. The **atom** is the basic unit of matter. The water molecule is composed of two hydrogen atoms and one oxygen atom.
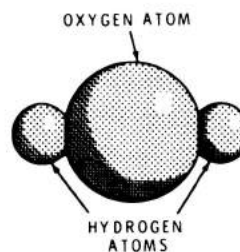
Figure 1-1
Molecule of water.

Electrons, Protons, and Neutrons

Atoms can be further broken down into particles called electrons, protons, and neutrons. These are the three basic building blocks which make up all atoms.

Figure 1-2 shows how electrons, protons, and neutrons are combined to form an atom. This is a helium atom. Two protons and two neutrons are bunched together near the center of the atom. This center part, which is composed of protons and neutrons, is called the **nucleus** of the atom.
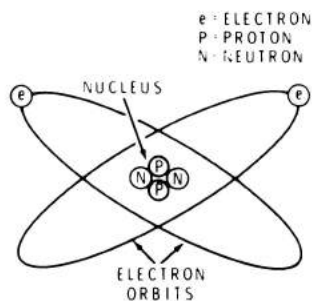
Figure 1-2
Structure of a helium atom.

Rotating around the nucleus are the electrons. Notice that the helium atom has two electrons. The electrons are extremely light and they travel at fantastic speeds. The atom can be compared to the solar system with the nucleus representing the sun and the electrons representing the planets. The electrons orbit the nucleus in much the same way that the planets orbit the sun.

Figure 1-3 shows the atomic structure of hydrogen, carbon, and copper. Notice that the number of electrons is always equal to the number of protons. This is normally true of any atom. When the number of electrons and protons are equal, the atom is said to be in its normal, balanced, or neutral state. As you will see later, this state can be upset by an external force. However, we normally think of the atom as containing equal numbers of electrons and protons.

HYDROGEN ATOM
(one electron, one proton)

CARBON ATOM
(6 electrons, 6 protons, 6 neutrons)

COPPER
(29 electrons, 29 protons, 35 neutrons)
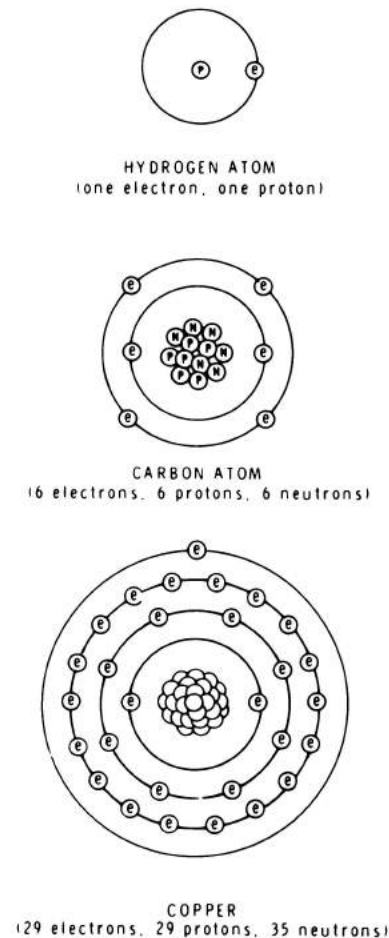
Figure 1-3
The difference between atoms is the number of electrons, protons, and neutrons they have.

Electrostatics

Electrostatics is the branch of physics dealing with electrical charges at rest, or static electricity. On the other hand, electronics deals largely with moving electrical charges. However, before you can fully understand the action of electrical charges in motion, you must first have some basic knowledge of their behavior at rest.

We have examined the structure of the atom and discussed some of the characteristics of the electron, proton, and neutron. However, we have not yet discussed the most important characteristic of these particles. This characteristic is their electrical charge. An **electrical charge** is a force associated with the electron and the proton. It is this electrical charge which makes the electron useful in electrical and electronic work.

The electrical charge is difficult to visualize because it is not a thing, like a molecule or an atom. Rather, it is a force in much the same way that gravity and magnetism are forces. It is a property which electrons and protons have that causes these particles to behave in certain ways.

There are two distinct types of electrical charges. Because these two types of charges have opposite characteristics, they have been given the names positive and negative. The electrical charge associated with the electron has been arbitrarily given the name **negative.** The electrical charge associated with the proton is considered **positive**. The neutron has no electrical charge at all. It is electrically neutral and, therefore, plays no known role in electricity.

The electron orbits around the nucleus at a fantastic speed. What force keeps the electron from flying off into space? It is not gravity because the gravitational force exerted by the nucleus is much too weak. Instead, the force at work here is caused by the charge on the electron in orbit and the charge on the proton in the nucleus. The negative charge of the electron is attracted by the positive charge of the proton. We call this force of attraction an **electrostatic force.** To explain this force, science has adopted the concept of an electrostatic field. Every charged particle is assumed to be surrounded by an electrostatic field which extends for a distance outside the particle itself. It is the interaction of these fields which causes the electron and proton to attract each other.

A.

PROTON

B.
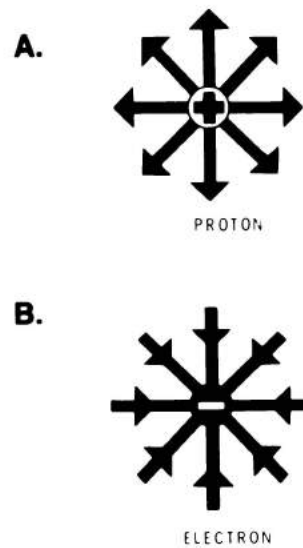
ELECTRON

Figure 1-4
Fields associated with protons and electrons.

Figure 1-4A shows a diagram of a proton. The plus sign represents the positive electrical charge. The arrows which extend outward represent the lines of force which make up the electrostatic field. Notice that the lines are arbitrarily assumed to extend outward away from the positive charge. Compare this to the electron shown in Figure 1-4B. The minus sign represents the negative charge while the arrows which point inward represent the lines of the electric field. Next, you will see how these two fields interact with one another.

Coulomb's Law

There is a basic law of nature which describes the action of electrical charges. It is called Coulomb's Law after Charles A. deCoulomb, who discovered this relationship. Coulomb's Law states that:

1. Like charges repel,

2. Unlike charges attract.

Because "like" charges repel, two electrons repel each other, as do two protons. Figure 1-5A illustrates how the lines of force interact between two electrons. The directions of the lines of force are such that the two fields cannot interconnect. The net effect is that the electrons attempt to move apart. That is, they repel each other. Figure 1-5B illustrates that the same is true of two protons. In Figure 1-5C, an electron and a proton (unlike charges) are shown. Here, the two fields do interconnect. As a result, the two charges attract and tend to move together.

A ELECTRONS REPEL

B PROTONS REPEL

C ELECTRON AND PROTON ATTRACT
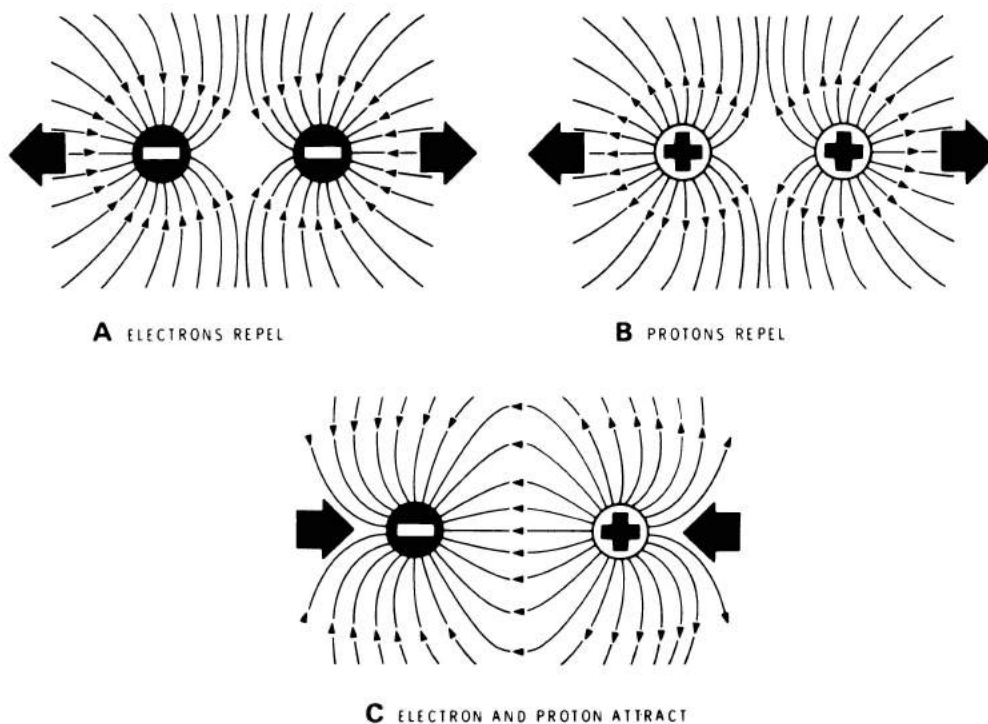
Figure 1-5
Action of like and unlike charges.

The magnitude of the negative charge on the electron is exactly equal to the magnitude of the positive charge on the proton. Figure 1-6 is a diagram of a hydrogen atom consisting of one electron in orbit around one proton. The negative charge of the electron is exactly offset by the positive charge of the proton. Thus, the atom as a whole has no charge at all. This atom has neither a negative nor a positive charge. It is electrically neutral.
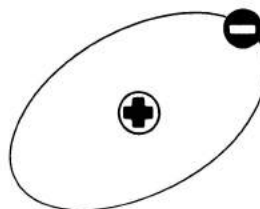
Figure 1-6
Hydrogen atom.

Atoms which are electrically neutral have no net charge. Therefore, they neither attract nor repel each other. By the same token, they are neither attracted nor repelled by charged particles such as electrons and protons. We have seen that atoms normally contain the same number of electrons (negative charges) as protons (positive charges). And, since the neutrons add no charge, all atoms are normally neutral as far as their electrical charges are concerned. However, this normal condition can be easily upset by external forces.

The Ion

Atoms are affected by many outside forces such as heat, light, electrostatic fields, chemical reactions, and magnetic fields. Quite often the balanced state of the atom is upset by one or more of these forces. As a result, an atom can lose or gain an electron. When this happens, the number of negative charges is no longer exactly offset by the number of positive charges. Thus, the atom ends up with a new charge. An atom which is no longer in its neutral state is called an **ion**. The process of changing an atom to an ion is call **ionization**.

There are both negative and positive ions. Figure 1-7 compares a neutral atom of carbon with negative and positive ions of carbon. Figure 1-7A shows the balanced or neutral atom. Notice that the six negative charges (electrons) are exactly offset by the six positive charges (protons). The neutrons are ignored in this example since they contribute nothing to the electrical charge.

Figure 1-7B shows the condition which exists when the carbon atom loses an electron. There are many forces in nature that can dislodge an electron and cause it to wander away from the atom. We will discuss this in more detail later. Notice that the carbon atom now has one more proton than electrons. Thus, there is one positive charge which is not cancelled by a corresponding negative charge. Therefore, the atom has a net positive charge. We call this **a positive ion**.

Figure 1-7C shows a carbon atom which has picked up a stray electron. In this case, there is one negative charge which is not offset by a corresponding positive charge. Hence, the atom has a net negative charge. This is called a **negative ion**.

A NEUTRAL ATOM

B POSITIVE ION CAUSED BY LOSING ONE ELECTRON

C NEGATIVE ION CAUSED BY PICKING-UP A STRAY ELECTRON
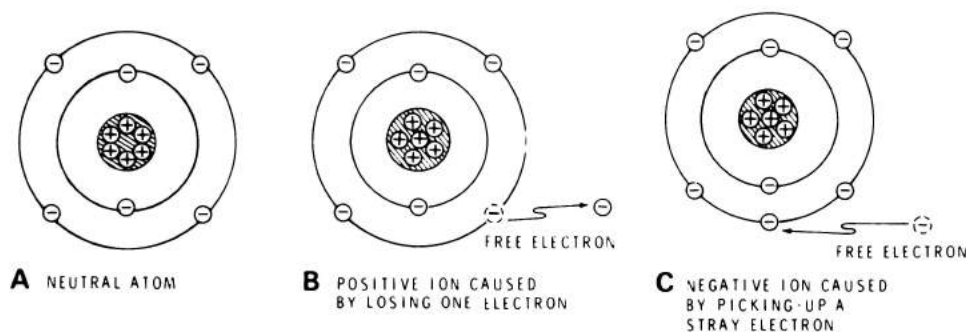
Figure 1-7
Carbon atom and ions.

The ion still has all the basic characteristics of carbon because the nucleus of the atom has not been disturbed. Therefore, an atom can give up or pick up electrons without changing its basic characteristics.

Changing atoms to ions is an easy thing to do and everything you see around you contains ions as well as atoms. The material around you also contains a larger number of free or stray electrons. These are electrons which have escaped from atoms, leaving behind a positive ion. As you will see later, the electrical characterisitics of different types of material are determined largely by the number of free electrons and ions within the material.

The Coulomb

The charge on an object is determined by the number of electrons it loses or gains. If it loses electrons, the charge is positive. If it gains electrons, the charge is negative.

The unit of electrical charge is called the **coulomb**. It is equal to the charge of 6.25×10^{18} electrons (6,250,000,000,000,000,000 or 6.25 billion billion. See "Powers of Ten and Scientific Notation," below). An object which has gained 6.25×10^{18} electrons has a negative charge of one coulomb. An object which has given up 6.25×10^{18} electrons has a positive charge of one coulomb.

Powers of Ten and Scientific Notation

Note above that the number 6,250,000,000,000,000,000 can also be expressed as 6.25×10^{18}. This number is read "six point two five times ten to the eighteenth power." The expression "ten to the eighteenth power" means that the decimal place in 6.25 should be moved 18 places to the right in order to convert to the proper number. The theory is that it is easier to write and remember 6.25×10^{18} than it is to write and remember 6,250,000,000,000,000,000. This shorthand method of expressing numbers is known as powers of ten or scientific notation. It is often used in electronics to express very large and very small numbers.

Very small numbers are expressed by using negative powers of ten. For example, 3.2×10^{-8} is scientific notation for the number 0.000000032. Here, "ten to the minus eighth power" means "move the decimal place in 3.2 eight places to the left." To be sure you have the idea, look at the following examples of both positive and negative powers of ten:

Positive Powers of Ten

$$7.9 \times 10^4 = 79{,}000$$

$$9.1 \times 10^8 = 910{,}000{,}000$$

$$1.0 \times 10^{12} = 1{,}000{,}000{,}000{,}000$$

Negative Powers of Ten

$$7.9 \times 10^{-4} = 0.00079$$

$$9.1 \times 10^{-8} = 0.000\ 000\ 091$$

$$1.0 \times 10^{-12} = 0.000\ 000\ 000\ 001$$

Study these examples until you get the idea of this system of writing numbers. If you feel you need additional explanation, read "Appendix A" at the end of this unit. This is a programmed instruction sequence designed to teach powers of ten and scientific notation in much greater detail.

Self-Review Questions

1. The _____ is the basic unit of matter.

2. The three basic particles which make up all atoms are the _____, _____, and _____.

3. The center of the of the atom, which is composed of protons and neutrons, is called the _____ .

4. State the electrical charge associated with each of the following:

    electron _____

    proton _____

    neutron _____

5. State Coulomb's law. _____

    _____

    _____

6. A normal atom has a _____ charge. However, if an atom loses an electron, it has a _____ charge. If an atom gains an electron it has a _____ charge. These charged atoms are called _____.

7. The unit of electrical charge is the _____.

Self-Review Answers

1. The **atom** is the basic unit of matter.

2. The three basic particles which make up all atoms are the **electron**, **proton**, and **neutron**.

3. The center of the atom, which is composed of protons and neutrons, is called the **nucleus**.

4. The electrical charge associated with each of the following is:

    electron − **negative**

    proton − **positive**

    neutron − **no charge or neutral**

5. Coulomb's law states — **Like charges repel, unlike charges attract**.

6. A normal atom has a **neutral** charge. However, if an atom loses an electron, it has a **positive** charge. If an atom gains an electron, it has a **negative** charge. These charged atoms are called **ions**.

7. The unit of electrical charge is the **coulomb**.

CURRENT

In electronics, **current** is defined as the flow of electrical charge from one point to another. Electrons have a negative charge. Therefore, if we can cause electrons to move from one point to another, we will have current. But before an electron can move from one point to another it must first be freed from the atom. Therefore, let's take a closer look at the mechanism by which electrons are freed from the atom.

Freeing Electrons

You have seen that electrons revolve around the nucleus at very high speeds. Two forces hold the electron in a precarious balance. The centrifugal force of the electron is exactly offset by the attraction of the nucleus. This balanced condition can be upset, causing the electron to be dislodged. But not all electrons can be freed from the atom with the same ease. Some are dislodged more easily than others. To see why, we must discuss the concept of **orbital shells**.

The orbits of the electrons in any atom fall into certain patterns. In all atoms which have two or more electrons, two of the electrons orbit relatively close to the nucleus. The area in which these electrons rotate is called a shell. This shell can support only two electrons. All other electrons must orbit in other shells, further from the nucleus.

A second shell, somewhat further from the nucleus, can contain up to eight electrons. A third shell can contain up to 18 electrons, and a fourth shell can hold up to 32 electrons. These first four shells are illustrated in Figure 1-8. Although not shown, there are also additional shells in the heavier atoms.
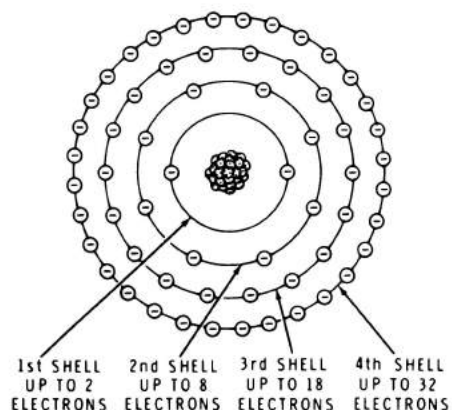
Figure 1-8
Arrangement of orbital shells in atoms.

1st SHELL   2nd SHELL   3rd SHELL   4th SHELL
UP TO 2   UP TO 8   UP TO 18   UP TO 32
ELECTRONS   ELECTRONS   ELECTRONS   ELECTRONS

Of particular importance to electronics is the outer electron shell of an atom. Hydrogen and helium atoms have one and two electrons respectively. In this case, the outer shell is the first (and only) shell. With atoms which have three to ten electrons, the outer shell is the second shell. Regardless of which shell it happens to be, the outer shell is called the **valence shell**. Electrons in this shell are called **valence electrons**.

Electrons are arranged in such a way that the valence shell never has more than eight electrons. This may be confusing since we have seen that the third shell can contain up to 18 electrons. An example shows why both statements are true. An atom of argon contains 18 electrons – 2 in the first shell, 8 in the second shell, and 8 in the third shell. It might seem that the next heavier atom, potassium, would have 9 electrons in its third shell. However, this would violate the valence rule stated above. Actually, what happens is that the extra electron is placed in a fourth shell. Thus, the 19 electrons are distributed in this manner: 2 in the first shell, 8 in the second shell, 8 in the third shell, and 1 in the fourth shell. Notice that the outer or valence shell becomes the fourth shell rather than the third. Once the fourth shell is established as the valence shell, the third shell can fill to its full capacity of 18 electrons.

The valence electrons are extremely important in electronics. These are the electrons which can be easily freed to perform useful functions. To see why the valence electrons are more easily freed, let's consider the structure of an atom of copper. Figure 1-9A shows how the electrons are distributed in the various shells. Notice that the valence shell contains only one electron. This electron is further from the nucleus than any of the other electrons. It is because of this greater distance that the valence electron experiences less attraction from the nucleus. For this reason, valence electrons can be easily dislodged from the atom.

Since we are concerned primarily with the valence electrons, we need not show the inner electrons. Instead, we can show the atom in the simplified form shown in Figure 1-9B. Figures 1-9C and 1-9D use this simplified form to illustrate one way in which a valence electron can be freed. Here, two copper atoms are shown as they might appear in a copper wire. Each valence electron is held in orbit by the attraction of the nucleus. However, the force of attraction is quite weak because the orbits are so far from the nucleus. If these two atoms are close together, the valence shells may be closer together than either electron is to its nucleus. At certain points in their orbits the two electrons may come very close together. When this happens, the force of repulsion between the two electrons is stronger than the force of attraction exerted by the nucleus. Thus, one or both of the electrons may be forced out of orbit to wander off as a free electron. Notice that, when the electron leaves, the atom becomes a positive ion.
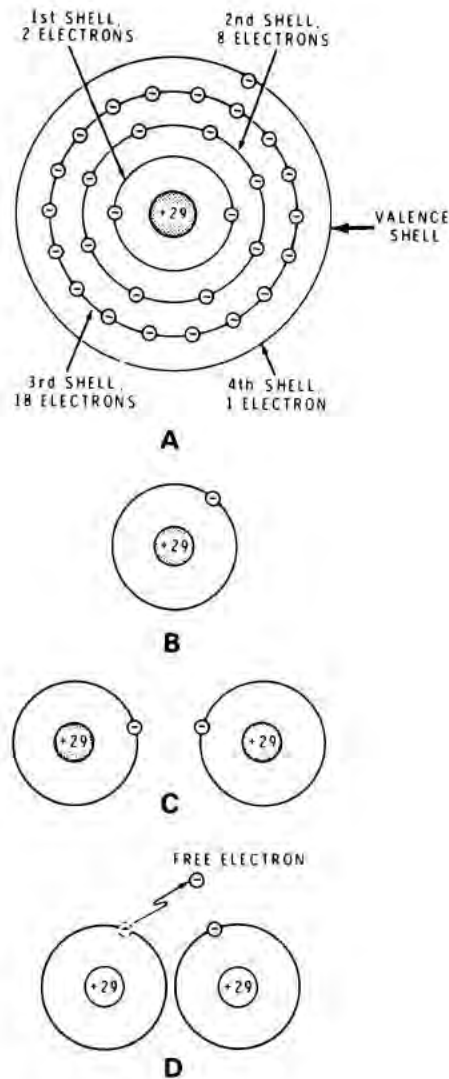
Figure 1-9
Freeing an electron from a copper atom.

As the free electron wanders around through the atomic structure, it may be eventually captured by another positive ion. Or, it may come close enough to other valence electrons to force them from orbit. The point is that events like these occur frequently in many types of material. Thus, in a piece of copper wire containing billions and billions of atoms, there are bound to be billions of free electrons wandering around the atomic structure.

Conductors and Insulators

The importance of valence electrons cannot be emphasized too strongly. The electrical characteristics of a substance depend on the action of the valence electrons. You have seen that the valence shell can hold up to eight electrons. Those substances which have their valence shells almost filled tend to be stable, and strive to fill their valence shell by capturing free electrons. Therefore, they have very few free electrons wandering around through their atomic structure. These substances which have very few free electrons are called **insulators**. By opposing the production of free electrons, they resist certain electrical actions. Insulators are important in electrical and electronics work for this reason. The plastic material on electrical wires, for example, is an insulator which protects us from electrical shock.

Materials in which the valence shell is almost empty have the opposite characteristics. Those with only one or two valence electrons tend to give up these electrons very easily. Copper, silver, and gold, for example, each have one valence electron, and valence electrons are very easily dislodged. Consequently, a bar of any one of these will have a very large number of free electrons. These substances which have a large number of free electrons are called **conductors**. Iron, nickel, and aluminum are also good conductors. Notice that all of these substances are metals. Most metals are good conductors. Conductors are important because they are used to carry electrical current from one place to another.

In some substances, like silicon and germanium, the valence shell is half filled, there are four valence electrons. These substances are called **semiconductors** because they are neither good conductors nor good insulators. Semiconductors are important in electronics because they are used to make transistors and integrated circuits.

The Battery

Current flow is the directed movement of free electrons from one place to another. Thus, to have current flow we must first have free electrons and positive ions. You have seen how valence electrons can be dislodged from atoms to form free electrons and positive ions. To perform a useful function, large numbers of electrons must be freed and concentrated in one area. One device for doing this is the ordinary battery. Figure 1-10 shows two familiar types of batteries, the dry cell (flashlight battery) and the wet cell (automobile battery).

While these two types of batteries are quite different in construction, they do have several points in common. Both have two terminals or poles to which an electrical circuit can be connected. Also, both employ a chemical reaction which produces an excess of electrons at one terminal and a deficiency of electrons at the other. The terminal at which the electrons congregate is called the negative terminal. It is indicated by the minus sign in Figure 1-10. The other terminal is indicated by the plus sign and has a deficiency of electrons. Next, you will see how a battery affects the free electrons in a conductor.
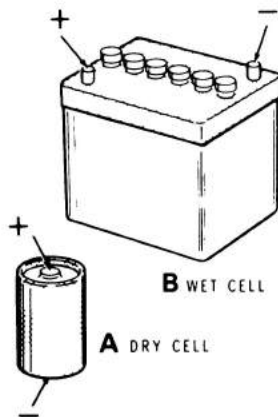
Figure 1-10
Types of batteries.

Random Drift and Directed Drift

A conductor is a substance that has a large number of free electrons. These free electrons do not stand still. Instead, they drift about in a random motion. Figure 1-11A represents a small section of a conductor containing many free electrons. At any instant, the free electrons are drifting at random in all directions. This is referred to as **random drift**. It occurs in all conductors and is of no practical use. To do useful work, the free electrons must be forced to drift in the same direction rather than at random.

By placing electrical charges at opposite ends of a conductor, you can influence the drift of electrons through it so that all or most electrons move in the same direction. Figure 1-11B shows a negative charge placed at one end of a conductor while a positive charge is placed at the other. The negative charge repels the free electrons while the positive charge attracts them. As a result, all of the free electrons move or drift in the same direction. The direction is from the negative charge to the positive charge.

Here, the application of the electrical charges at the end of the conductor has changed random drift to directed drift. This directed drift of free electrons is called **current flow**. We say that an **electric current** is flowing through the conductor. If the electrical charges shown in Figure 1-11B are isolated from one another, the flow of electrons will quickly cancel both charges and only a momentary current will flow. However, if the two electrical charges are caused by a battery, the chemical action of the battery can maintain the two charges for some time. Therefore, a battery can maintain a continuous current through a conductor for a long period.

A

B
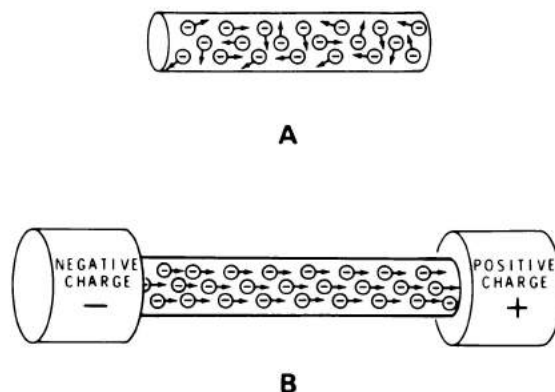
Figure 1-11
Comparison of random and directed drift.
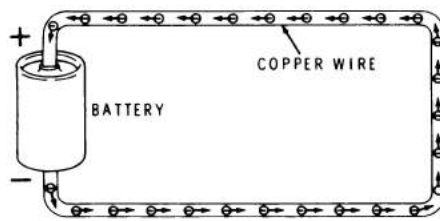
Figure 1-12
Current flows from the negative terminal to the positive
terminal of the battery.

A copper wire is a good example of a conductor. Figure 1-12 shows a length of copper wire connected from one terminal to the other of a battery. A heavy current will flow from the negative terminal of the battery to the positive terminal. Recall that the negative terminal is a source of free electrons. An electron at this point is repelled by the negative charge and is attracted by the positive charge at the opposite terminal. Thus, the electrons flow through the wire as shown. When they enter the positive terminal of the battery, they are captured by positive ions. The chemical reaction of the battery is constantly releasing new free electrons and positive ions to make up for the ones lost by recombination.

It should be pointed out that, in practice, you never connect a conductor directly across the terminals of the battery as shown in Figure 1-12. The heavy current would quickly exhaust the battery. This is an example of a "short circuit" and is normally avoided at all cost. This example is shown here merely to illustrate the concept of current flow.

The Electric Circuit

In its simplest form, an electric circuit consists of a power source, a load, and conductors for connecting the power source to the load. Often, the power source is a battery. The purpose of the power source is to provide the force necessary to direct the flow of electrons.

The **load** is generally some kind of electrical device that performs a useful function. It might be a lamp to produce light, a motor to produce physical motion, a horn to produce sound, or a heating element to create heat. Regardless of the type of load used, the load performs its useful function only when electric current flows through it.

The third part of the circuit is the conductors which connect the power source to the load. They provide a path for current flow. The conductor may be a length of copper wire, a strip of aluminum, the metal frame of an automobile, etc.
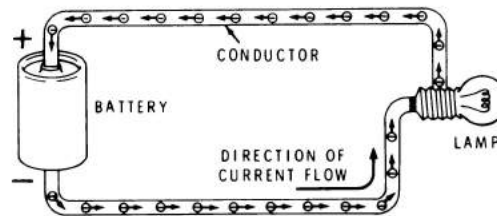
Figure 1-13
A simple electric circuit.

Figure 1-13 is a pictorial representation of an electric circuit consisting of a battery, a lamp, and copper wires. The battery produces the force necessary to cause the directed flow of electrons. The force it develops causes the free electrons in the conductor to flow through the lamp in the direction shown. These free electrons are repelled by the negative charge and attracted by the positive charge. Thus, the electrons flow from negative to positive. And since negative and positive charges in the battery are constantly being replenished by the chemical action of the battery, it can maintain a current flow for a long period of time.

As the electrons flow through the lamp in Figure 1-13, they heat up the wire within the lamp. As the wire becomes hotter, the lamp emits light. The lamp will glow as long as a fairly strong current is maintained.

You probably know from experiences with flashlights that a battery cannot maintain a constant current flow forever. As the battery is used, the chemical reaction within it slows down, the force it provides becomes weaker, and less current is provided. As a result, the lamp emits less light; it becomes dimmer and dimmer, and eventually it emits no light at all. At this time the battery is said to be dead, burned out, or run down. In this condition the battery can no longer push enough electrons through the lamp to make it glow.

The circuit in Figure 1-13 can be made much more practical if you add a switch. This provides a simple method of turning the lamp on and off.
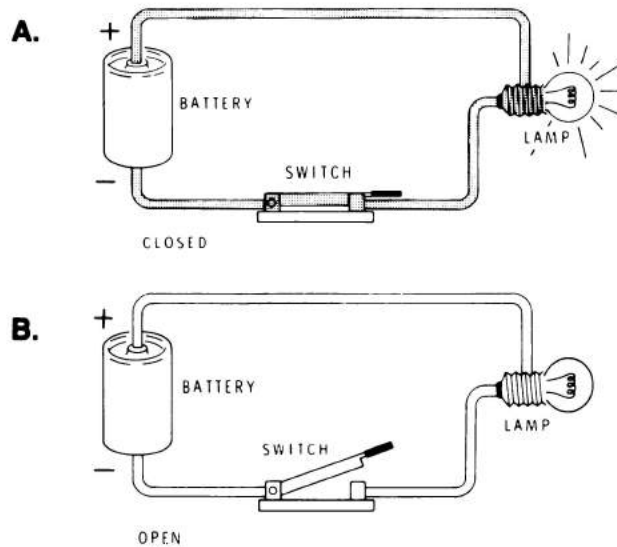
Figure 1-14
Circuit with switch.

Figure 1-14 shows the circuit after the switch has been added. For simplicity, a "knife" switch is shown. It consists of two metal contacts to which conductors may be connected, a metal arm that can be opened and closed, and a base. Current cannot flow through the base of the switch because an insulator material is used. Current can flow only through the arm, and then only if the arm is closed.

In Figure 1-14A, where the switch is closed, current can flow from the negative battery terminal through the switch and lamp to the positive terminal. The lamp lights because current flows through it. When the switch is opened, as shown in Figure 1-14B, the current path is broken and the lamp does not glow.

While simple circuits can be drawn as shown in Figures 1-13 and 1-14, it would be very difficult to draw complex circuits in this manner. For this reason, the schematic diagram was developed. A **schematic diagram**, or schematic, is a drawing that uses symbols to represent circuit components. Therefore, before you can understand a schematic diagram, you will have to learn the symbols for the various components.
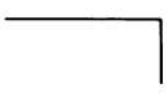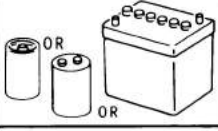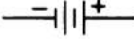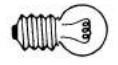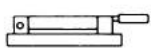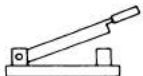
	PICTORIAL	SCHEMATIC SYMBOL
A CONDUCTOR OR WIRE		
B BATTERY OR CELL		
C LAMP		
D SWITCH (CLOSED)		
E SWITCH (OPEN)		

Figure 1-15
Pictorial representations compared with
the schematic symbols.

Figure 1-15 shows the schematic symbols for the circuit components we have used up to this point. The wire or conductor is represented by a single line. The battery is replaced by a series of long and short lines. Long lines represent the positive terminals while short lines represent the negative terminals. The same symbol is used for all types of batteries.
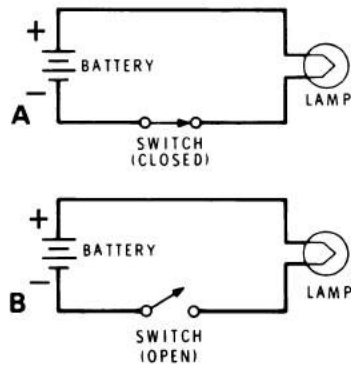
Figure 1-16
Schematic diagram of a simple circuit.

Figure 1-16 shows the symbols combined to form schematic diagrams. Figure 1-16A is the schematic diagram for the pictorial drawing of Figure 1-14A. Figure 1-16B represents Figure 1-14B.

The circuit shown in Figure 1-16 is the schematic diagram of a flashlight. It could also be a simplified schematic for the headlight system in an automobile. In fact, it can represent any system which contains a battery, a lamp, and a switch. If you replace the lamp with a motor, the circuit becomes that of the starter system in a car. In this case, the switch is operated by the ignition key. Other circuits which operate in a similar manner are the doorbell and the automobile horn. In the first case, the bell is the load while the switch is operated by a pushbutton at the door. In the second case, the horn is the load while the switch is located on the steering wheel.

The Ampere

The unit of current is the **ampere**. The ampere is the rate at which electrons move past a given point. An ampere is equal to 1 coulomb per second. That is, if 1 coulomb (6.25×10^{18} electrons) flows past a given point in 1 second, then the current is equal to 1 ampere. Coulombs indicate numbers of electrons; amperes indicate the rate of electron flow or coulombs per second.

When 6.25×10^{18} electrons flow through a wire each second, the current flow is 1 ampere. If twice this number of electrons flows each second, the current is 2 amperes. This relationship is expressed by the equation:

$$\text{amperes} = \frac{\text{coulombs}}{\text{seconds}}$$

If 10 coulombs flow past a point in two seconds, then the current flow is 5 amperes.

The name ampere is often shortened to amp and is abbreviated A. Many times the ampere is too large a unit. In these cases, metric prefixes are used to denote smaller units. The milliampere (mA) is one thousandth (.001) of an ampere. The microampere (μA) is one millionth (.000 001) of an ampere. In other words, there are 1000 milliamperes or 1,000,000 microamperes in an ampere. In electronics, the letter "I" is used to represent current. Therefore, in the shorthand of electronics you could write the statement "the current is 2 amps" as an equation "I = 2A."

We change from amperes to milliamperes by multiplying by 10^3 ($\times$ 1000). Thus, 1.7 amperes is equal to 1.7×10^3 milliamperes. Also, we change from amperes to microamperes by multiplying by 10^6 ($\times$ 1,000,000). Therefore, 1.7 amperes is equal to 1.7×10^6 microamperes.

For those who need it, a more detailed explanation of metric prefixes is given in "Appendix A."

Self-Review Questions

8. Substances that have very few free electrons are called _____.

9. Substances that have a large number of free electrons are called _____.

10. A directed drift of free electrons is called _____.

11. In an electrical circuit electrons flow from the _____ terminal of the battery to the _____ terminal.

12. A drawing that uses symbols to represent circuit components is known as a _____.

13. The schematic symbols for four components are shown in Figure 1-17. Identify each one:

    A. _____

    B. _____

    C. _____

    D. _____

14. The unit of current is the _____ .
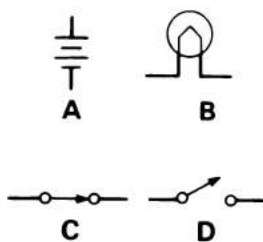
Figure 1-17
Identify these symbols.

Self-Review Answers

8. Substances that have very few free electrons are called **insulators**.

9. Substances that have a large number of free electrons are called **conductors**.

10. A directed drift of free electrons is called **current flow**.

11. In an electrical circuit electrons flow from the **negative** terminal of the battery to the **positive** terminal.

12. A drawing that uses symbols to represent circuit components is known as a **schematic diagram**.

13. Figure 1-17 shows the following schematic symbols:

    A. Battery,

    B. Light bulb,

    C. Closed switch,

    D. Open switch.

14. The unit of current is the **ampere**.

VOLTAGE

You have seen that current will not flow in a circuit unless an external force is applied. In the circuits discussed in the previous section, the force was provided by batteries, which change chemical energy to electrical energy by separating negative charges (electrons) from positive charges (ions). These charges produce the force, or pressure, which causes electrons to flow and do useful work. This force is given several different names, discussed below, which are used more or less interchangeably.

Electromotive Force

One popular name is **electromotive force**, which is abbreviated **emf**. This name is very descriptive since it literally means a force which moves electrons. Thus, emf is the force or pressure which sets electrons in motion. This force is a natural result of Coulomb's law.

You will recall that Coulomb's law states that like charges repel while unlike charges attract. The battery, by chemical action, produces a negative charge at one terminal and a positive charge at the other. The negative charge is simply an excess of electrons while the positive charge is an excess of positive ions. If a closed circuit is connected across the battery as shown in Figure 1-18A, a path for electron flow exists between the battery terminals. Free electrons are repelled by the charge on the negative terminal and are attracted by the charge on the positive terminal. The two opposite charges exert a pressure which forces the electrons to flow. Thus, the force or pressure is the result of the attraction of the unlike charges. To summarize, emf is the force which sets electrons in motion in a closed circuit.

A.
EMF CAUSES ELECTRONS TO FLOW IN CLOSED CIRCUIT

B.
POTENTIAL FOR CURRENT FLOW EXISTS EVEN THOUGH THERE IS NO CURRENT
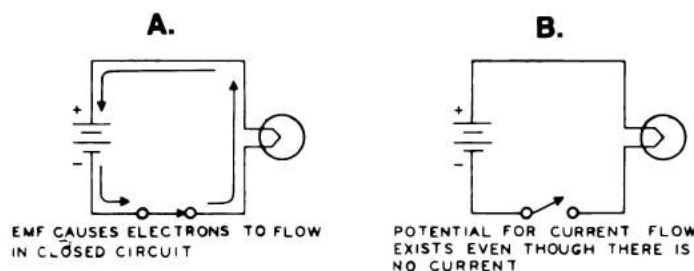
Figure 1-18
EMF and potential difference.

Potential Difference

Another name for this force is potential difference. This name is also very descriptive. It describes the characteristics of emf in an open circuit. Emf is the force which causes electrons to move as shown in Figure 1-18A. However, consider the situation shown in Figure 1-18B. Here, electrons cannot flow because the switch is open. Nevertheless, the battery still produces the same pressure or force as before. Thus, the **potential** for producing current flow exists even though no current is presently flowing. As used here, potential means the possibility of doing work. If the switch is closed, current flows, the lamp lights, and useful work is done. Therefore, whether a battery is connected into a circuit or not, it has the potential for doing work.

A single unit of charge, such as an electron, also has the potential for doing work. It can deflect other electrons, which is some small amount of work. However, it is the difference between two charges which is termed **potential difference**. An example is a battery. It has a negative terminal and a positive terminal. The electrons on the negative terminal are straining to move toward the positive terminal to equalize the charges. A potential difference exists between the negative terminal and the positive terminal of a battery.

Potential difference also exists between charges of the same polarity. Figure 1-19A shows two negative charges. Charge A is 4 coulombs. Charge B is 2 coulombs. When a conductor is connected between the two charges, electrons will flow from charge A to charge B. This is shown in Figure 1-19B. The electrons will flow until the two charges are equalized.

A

CHARGE A
4 COULOMBS
NEGATIVE

CHARGE B
2 COULOMBS
NEGATIVE
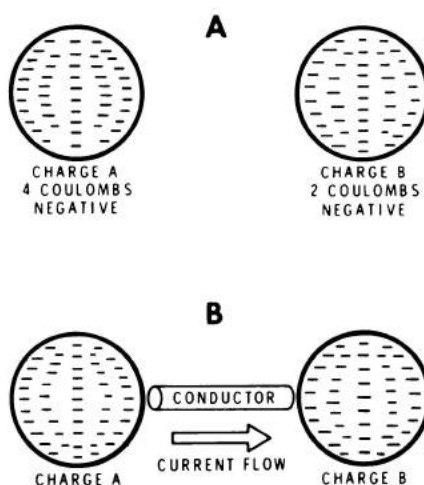
Figure 1-19
Potential difference exists
between unequal charges.

B

CONDUCTOR

CURRENT FLOW

CHARGE A   CHARGE B
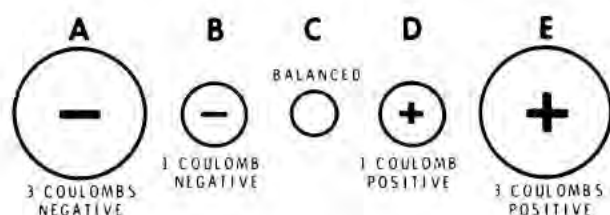
Figure 1-20
Five levels of charge.

Figure 1-20 shows five terminals at various levels of charge. Since no two are at the same charge level, a difference of potential exists between any two terminals. Consequently, if a conductor is placed between any two terminals, electrons will flow until those two charges are balanced. Notice that terminal C has no charge. That is, it contains the same number of electrons as positive ions. Nevertheless, if terminal C is connected to any other terminal, electrons will still flow. If it is connected to one of the negative terminals, electrons will flow into terminal C. If it is connected to one of the positive terminals, electrons will flow from terminal C. Remember that electrons always flow from the more negative to the more positive terminal.

Voltage

Another term which is often used interchangeably with emf and potential difference is **voltage**. However, strictly speaking, there is a difference between voltage and emf. Voltage is the measure of emf or potential difference. For example, the battery in your car has an emf of 12 volts. The emf supplied by wall outlets is 115 volts, while that required by most electric stoves is 220 volts. A large screen color TV receiver produces an emf at one point which may be 25,000 volts or higher. High tension power lines may have a difference of potential as high as 760,000 volts.

The unit of emf or potential difference is the **volt**. One volt is the pressure required to force a current of one ampere through a resistance of one ohm. We will be discussing resistance in detail later in this unit.

The abbreviation of volt is V. Thus, 1.5 volts is abbreviated 1.5 V. As with amperes, metric prefixes are attached to indicate smaller and larger units of voltage. Thus, one millivolt (mV) equals 1/1,000 volt, while one micro-volt (μV) equals 1/1,000,000 volt. Also, one kilovolt equals 1000 volts while one megavolt (MV) equals 1,000,000 volts. In electronics, the letter "E" is used to represent voltage. Therefore, in the shorthand of electronics you could write the statement "the voltage is 10 volts as an equation "E=10 V."

Four popular-size batteries. Courtesy RCA.

Battery Connections

You can obtain a voltage higher than the emf of a single battery by using the **series** connection shown in Figure 1-21. The total voltage of series connected batteries is equal to the sum of the individual battery voltages. This is shown in Figure 1-21 where three 1.5 V batteries are connected in series. The total voltage is 4.5 V. The schematic diagram for this connection is also shown in Figure 1-21.
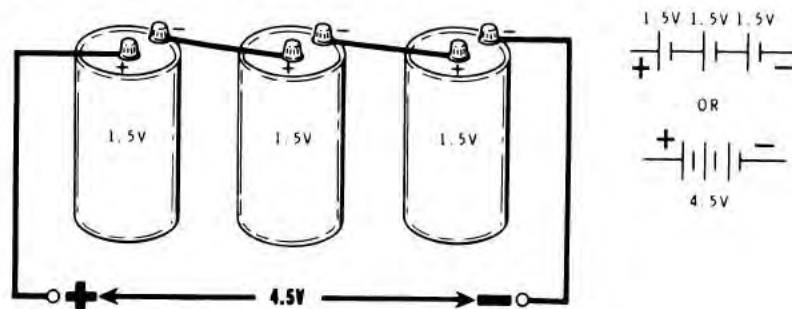
Figure 1-21
Series-connected batteries.

While the total voltage increases with the series connection, the current capacity does not. Since the total circuit current flows through each battery, the current capacity is the same as a single battery.

You can increase current capacity by connecting several batteries in **parallel**. This is shown in Figure 1-22A. In the parallel connection, each battery can contribute the same amount of current to the load. Therefore, if we connect three batteries in parallel, the current capacity will be three times that of a single battery. This is shown in Figure 1-22B. Notice that while the parallel connection increases the current capacity, the voltage remains the same.
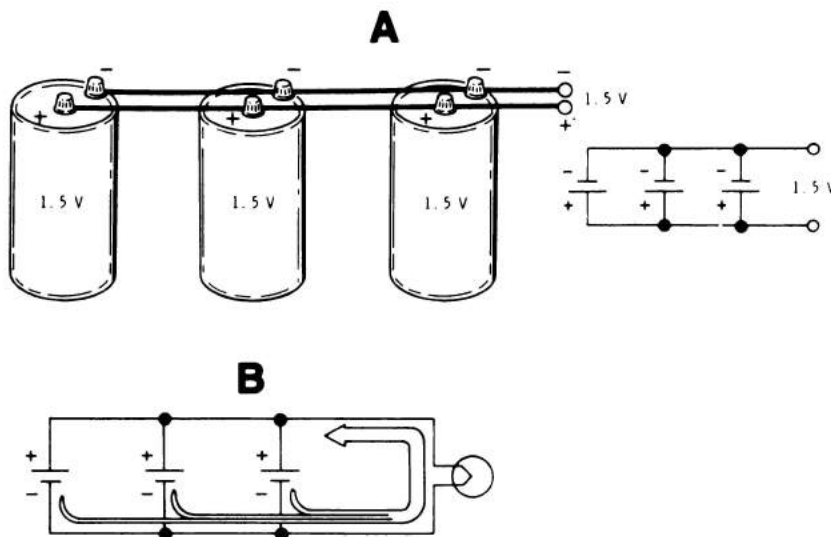
Figure 1-22
Parallel battery connection.

To obtain both a higher voltage and an increased current capacity, you can connect the batteries in **series-parallel**. An example of a series-parallel connection is shown in Figure 1-23. Here, two batteries are connected in series to increase the voltage to 3 V. Then the current capacity is doubled when a second series string is connected in parallel with the first.
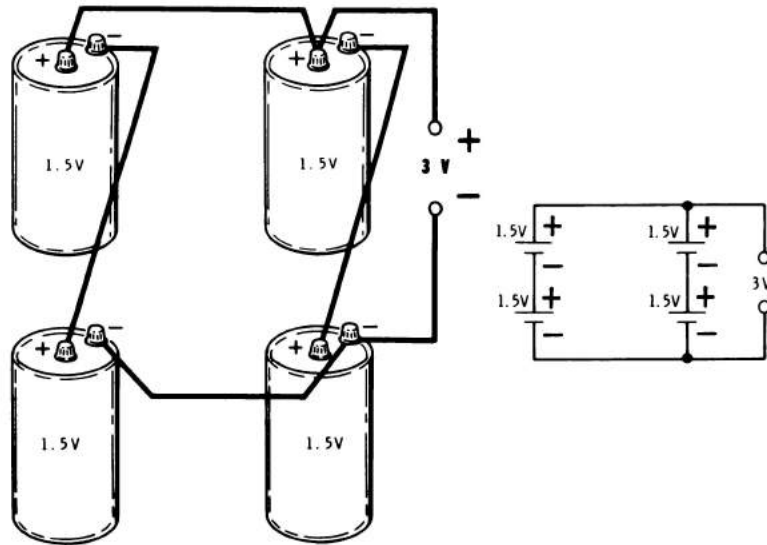
Figure 1-23
Series-parallel battery connection.

To be certain that you have the idea, consider the battery connection shown in Figure 1-24. Here, three batteries are connected in series. Therefore, the total voltage is 4.5 V. Connecting three series strings in parallel increases the current capacity by three. This series-parallel connection has tripled both voltage and current capacity.
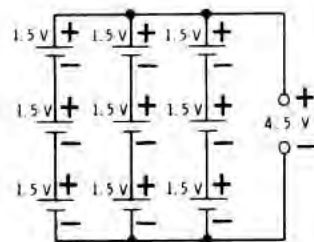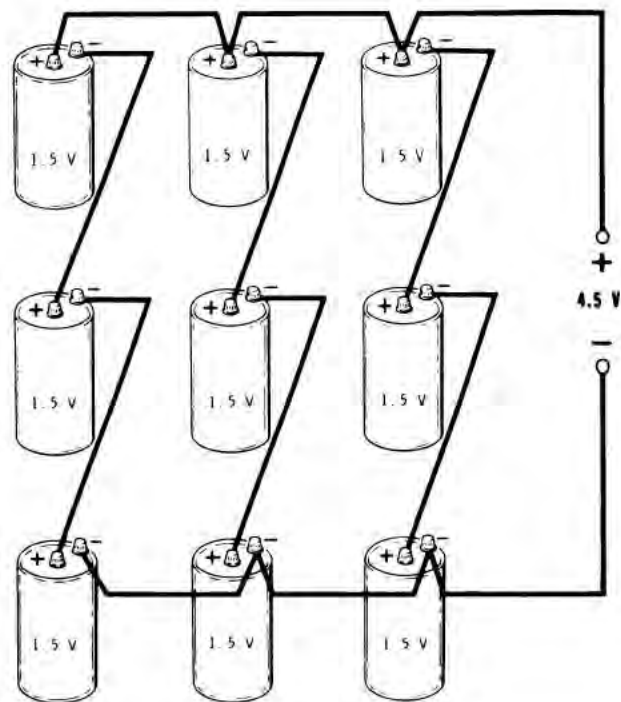
Figure 1-24
Nine batteries connected in series-parallel.

Voltage Rises and Voltage Drops

Any emf introduced into a circuit by a voltage source, such as a battery, is called a **voltage rise**. Thus, a 10 V battery has a voltage rise of ten volts.

Since the emf or energy introduced into the circuit is called a voltage rise, the energy removed from the circuit by the load is called a **voltage drop**. A voltage drop is expressed in volts just as is the voltage rise.

Figure 1-25A shows a 10-volt battery with a light bulb connected across it. The battery provides a voltage rise of 10 volts. As electrons flow through the lamp, a voltage drop is developed across it. Since the lamp consumes the same amount of energy that the battery provides, the voltage drop across the lamp is equal to the voltage rise across the battery. The voltage drop is 10 volts.

In Figure 1-25B, two light bulbs are connected in series across a 10 volt battery. Each bulb drops part of the 10 volts supplied. If the two lamps are identical, then each will drop half of the supplied voltage as shown. If the two lamps are not identical, one bulb will drop more voltage than the other. However, the sum of the voltage drops will always equal the sum of the voltage rises.
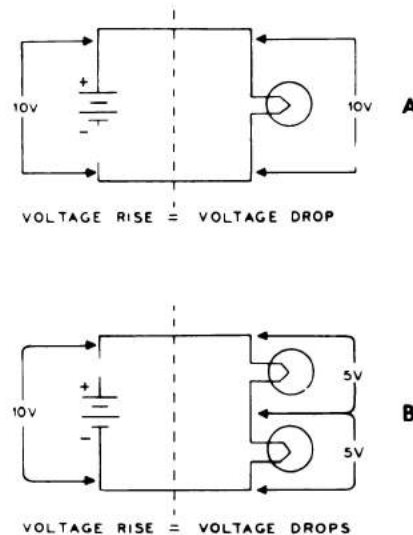
Figure 1-25
The voltage drops are equal to the voltage rises.

Concept of Ground

One of the most important points in the study of electricity and electronics is the concept of *ground*. Originally, ground was just what the name implies — the earth. In fact, in some countries the name **earth** is used instead of **ground**. Earth is considered to have zero potential. Thus, ground or earth is the reference point to which voltages are most often compared. Many electrical appliances in your home are grounded. This is especially true of air conditioning units, electric clothes dryers, and washing machines. Often this is done by connecting a heavy wire directly to a cold water pipe which is buried deep in the earth (ground). In other cases, a third prong on the power plug connects the metal frame to ground. The purpose of this is to protect the user in case a short circuit develops in the appliance. It also places the metal parts of different appliances at the same potential so that you are not shocked by a differece in potential between two appliances. This type of ground is sometimes called *earth ground*.

However, there is a slightly different type of ground used in electronics. For example, a certain point in a small transistor radio is called *ground* although the radio does not connect to earth in any way. This is the concept of ground with which we will be primarily concerned in this course. In this case, ground is simply a zero reference point within an electric circuit. In most larger pieces of electronic equipment the zero reference point or ground point is the metal frame or chassis on which the various circuits are constructed. All voltages are measured with respect to this chassis.

In your automobile, the chassis or metal body of the automobile is considered ground. If you look closely at the straps leaving the battery you will see that one cable connects directly to the metal frame of the car. This point is considered to be ground as is every other point on the metal frame.

In electronics, ground is important because it allows us to have both negative and positive voltages. Up to now we have been concerned only with relative voltages between two points. For example, a 6-volt battery has an emf between its two terminals of 6 volts. We do not think of this as +6 volts or −6 volts but rather simply 6 volts.

However, the concept of ground allows us to express negative and positive voltages. Remember ground is merely a reference point which is considered zero or neutral. If we assume that the positive terminal of a 6-volt battery is ground, then the negative terminal is 6 volts more negative. Thus, the voltage at this terminal with respect to ground is −6 volts.
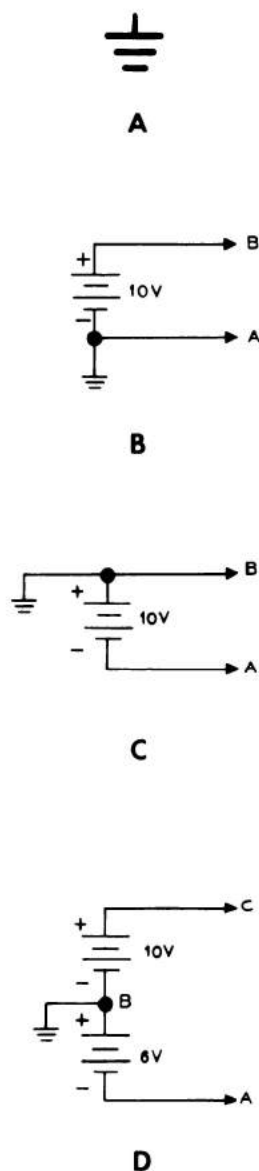
On the other hand, if we assume that the negative terminal of the battery is ground, then the positive terminal with respect to ground is +6 volts. Notice that the battery can produce −6 volts or +6 volts depending on which terminal we assign to ground.

Many small electronic devices such as calculators, transistor radios, etc. do not have metal frames. Instead, all components are mounted on a printed circuit board. Here ground is nothing more than an area of copper on the board. However, as before, all voltages are measured with respect to this point. In this case, ground is simply a common reference which is a handy starting point for measuring voltages.

The schematic symbol for ground is shown in Figure 1-26A. Figure 1-26B shows how it is used in the circuit. Point A is at ground or zero potential. Now, since this is a 10-volt battery, point B is at a plus 10-volt potential with respect to ground. We say that point B is ten volts above ground or that the voltage at this point with respect to ground is +10 volts.

Figure 1-26C shows why ground is so important. Here the same battery is shown but with the positive terminal connected to ground. That is, the positive terminal is the zero volt point in this circuit. Because the negative terminal is ten volts more negative, the voltage at point A with respect to ground is −10-volts. Thus, we can use the battery as a +10-volt source or as a −10-volt source depending on where we connect ground.

Another example is shown in Figure 1-26D. Here two batteries are connected in series, with the ground connection between them. Thus, the zero reference is at point B. Since the top battery has an emf of 10 volts, the voltage at point C with respect to ground is +10 volts. The lower battery has an emf of 6 volts. Because the positive terminal is connected to ground, the emf at point A with respect to ground is −6 volts.

We sometimes loosely speak of the voltage at a particular point. Actually, voltage is always the measure of the potential difference between *two* points. Thus, in Figure 1-26D, when we speak of the voltage at point A, what we really mean is the voltage between point A and ground.

Figure 1-26
The polarity of the voltage depends upon the connection to ground.

Self-Review Questions

15. Define emf. _____

16. What are two other names for emf? _____

17. What is the unit of emf? _____

18. What is the potential difference between the + and − terminals of the battery shown in Figure 1-27? _____
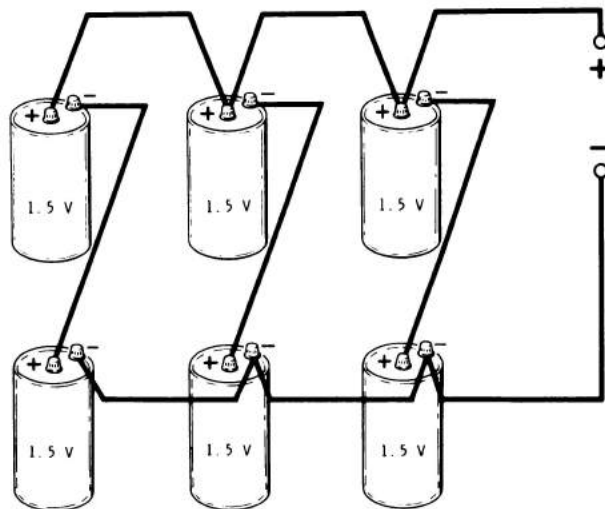
Figure 1-27
What is the output voltage?
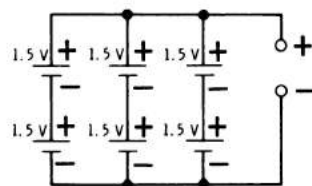
19. What is the current capacity of the battery connection shown in Figure 1-27? _____

20. What is the difference between a voltage rise and a voltage drop?
_____
_____

21. What is "ground"? _____
_____

Self-Review Answers

15. Emf is the abbreviation for electromotive force. It is the force which moves electrons.

16. The two other names for emf are potential difference and voltage.

17. The volt is the unit of emf.

18. The output voltage is 3 volts.

19. The current capacity is 3 times that of a single battery.

20. A voltage rise is the energy introduced into a circuit by a voltage source. A voltage drop is the energy removed from the circuit by the load.

21. Ground is an earth connection or a common reference point within a circuit.

RESISTANCE

Resistance is that property which opposes current flow. All materials have this property to some extent. Some materials, such as glass and rubber, offer a great deal of opposition to current flow. They allow almost no current to flow through them. Therefore, they are said to have a very high resistance. Other materials, such as silver and copper, offer very little opposition to current flow. Therefore, they have a very low resistance.

The Ohm

The unit of resistance is the **ohm**. An ohm may be defined in several different ways. Originally, it was defined as the resistance of a column of mercury which is 106.3 centimeters long and one square millimeter in cross sectional area. Unfortunately, this amount of resistance is hard to visualize.

It may be helpful to think in terms of something with which we are more familiar. For example, a length of number 22 copper wire that is 60 feet long has a resistance of about one ohm.

The most common way to define the ohm is in terms of voltage and current. One ohm is the amount of resistance which will allow one ampere of current to flow in a circuit to which one volt of emf is applied. Or stated another way, if one volt causes one ampere of current in a circuit, then the resistance of the circuit is one ohm. Figure 1-28 illustrates these three ways of defining the ohm.

MERCURY

106.3 cm

1 mm

1 mm

RESISTANCE OF COLUMN OF MERCURY WITH DIMENSIONS SHOWN

1V   1 OHM

1 AMP

RESISTANCE WHICH ALLOWS ONE AMPERE TO FLOW WHEN EMF IS 1 VOLT

60 FEET

NUMBER 22 COPPER WIRE

RESISTANCE OF 60 FEET OF NUMBER 22 COPPER WIRE
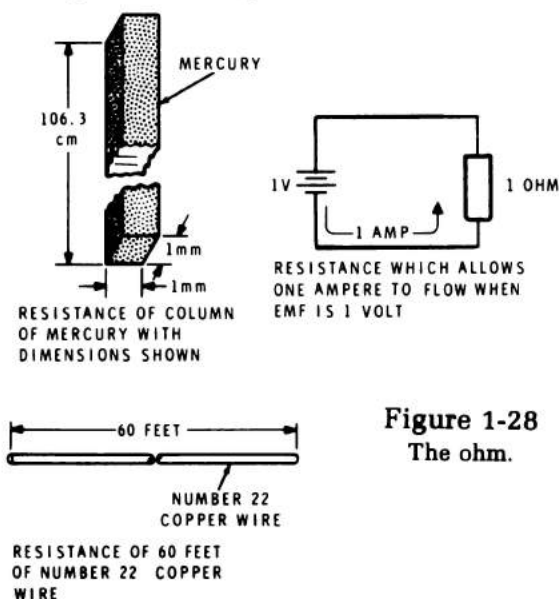
Figure 1-28
The ohm.

The Greek letter omega (Ω) is commonly used to represent ohms. Thus, 1 ohm may be written 1 Ω. Also, one thousand ohms may be written as 1000 Ω, 1 kilohms, or as 1 kΩ. Finally, one million ohms may be written as 1,000,000 Ω, 1 megohm, or as 1 MΩ. In electronics, the letter R is used to represent resistance. Thus, in the shorthand of electronics, the statement "the resistance is ten ohms" may be written as an equation: "R = 10 Ω."

Resistors

A **resistor** is an electronic component that has a certain specified resistance. It is designed specifically to introduce a desired amount of resistance into a circuit.

You have seen that a 60 foot length of number 22 copper wire has a resistance of one ohm. If we were to increase the length of the copper wire, the resistance would also increase. By varying the length of wire, we can obtain any value of resistance desired. However, making a resistor with copper wire would require large amounts of wire. Normally, a special resistance wire called nichrome is used that has a much higher resistance than copper. The correct length of nichrome wire is wound around a ceramic form. This type of resistor is a **wire-wound** resistor. The construction of a wire-wound resistor is shown in Figure 1-29.

Courtesy IRC Division of TRW Inc.

Figure 1-29
Wire-wound resistors are formed by wrapping resistance wire around a ceramic core. The unit is sealed by a hard plastic insulating material.

Carbon is neither a good conductor nor a good insulator. Instead, it falls in between and is called a semiconductor. This makes carbon ideal as a material for resistors. By combining carbon granules and a powdered insulating material in various proportions, a wide range of resistor values are possible. This type of resistor is called a **carbon-composition** resistor.

Figure 1-30A shows the construction of a carbon-composition resistor. Granules of carbon and a binder material are mixed together and shaped into a rod. Wire leads are inserted and the package is sealed with a nonconducting coating. The schematic symbol for a resistor is shown in Figure 1-30B.
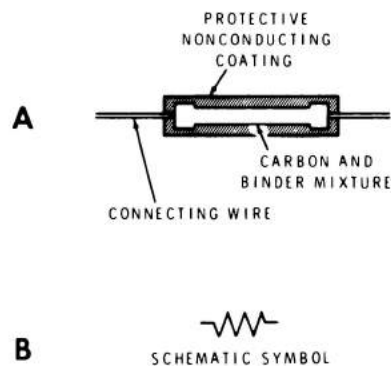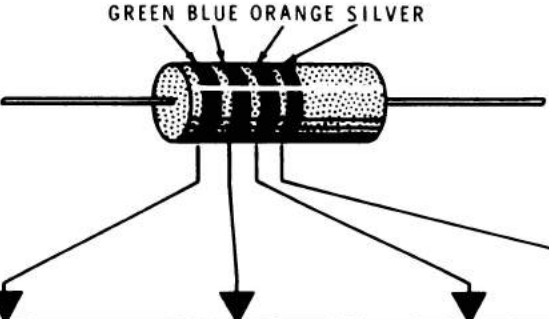
Figure 1-30
Carbon-composition resistor construction (A) and schematic symbol for a resistor (B).

GREEN BLUE ORANGE SILVER

COLOR	BAND 1 FIRST NUMBER	BAND 2 SECOND NUMBER	BAND 3 MULTIPLIER	BAND 4 TOLERANCE
BLACK	0	0	1	
BROWN	1	1	10	
RED	2	2	100	
ORANGE	3	3	1,000	
YELLOW	4	4	10,000	
GREEN	5	5	100,000	
BLUE	6	6	1,000,000	
VIOLET	7	7	10,000,000	
GRAY	8	8	100,000,000	
WHITE	9	9	1,000,000,000	
GOLD			0.1	±5%
SILVER			0.01	±10%
NONE				±20%

Color Code for Resistors

With the carbon composition resistor, the resistance value is indicated by the colored bands around the resistor. The first two bands indicate, by their color, the first two digits of the resistor value. The third band indicates a multiplier. The value of the resistor is determined by multiplying the first and second digits by the multiplier. For example, a resistor in which the first two bands were brown and the third band was red would have the value:

1st DIGIT   2nd DIGIT   MULTIPLIER
   1   1   100

11   X   100 = 1100Ω

The fourth band is the resistor's tolerance. It indicates the maximum amount that a resistor will vary from the value determined by the colors of the first three bands. If the fourth band of an 1100 ohm resistor was gold, then the maximum value of that resistor would be 5% greater than 1100 ohms while the minimum value would be 5% less than 1100 ohms. Therefore, the maximum value is 1155 ohms and the minimum is 1045 ohms.

Courtesy IRC Division of TRW Inc.

Figure 1-31
Deposited film resistor.

Another type of resistor that is becoming increasingly popular is the **deposited film** resistor. The construction of this type of resistor is shown in Figure 1-31. In these devices, a resistance film is deposited on a nonconductive rod. Then the value of resistance is set by cutting a spiral groove through the film. This changes the appearance of the film to that of a long, flat ribbon spiralled around the rod. The groove adjusts the length and width of the ribbon so that the desired value is achieved.

Several different types of deposited film resistors are available. The most common type is the carbon-film resistor. Here, a carbon film is deposited on a ceramic rod. Several metal-film types are also available. One uses a nickel-chromium (nichrome) film on an aluminum oxide rod. Another uses a tin-oxide film on a glass rod.

Variable Resistors

The volume controls on TV receivers and radios are examples of variable resistors. These are resistors whose values you can change simply by rotating a shaft.

Figure 1-32A shows the construction of a variable carbon resistor. Figure 1-32B is a rear-view of the inside of the device, which shows how the resistance is changed. A flat, circular strip of carbon is mounted between the two end terminals. A contact which moves along the resistance

element connects to the center terminal. This arm is attached to the movable shaft. If the arm is moved in the direction shown by the arrow, the resistance between terminals 1 and 2 increases. Notice that, simultaneously, the resistance between terminals 2 and 3 decreases. This type of variable resistor is called a **potentiometer** or simply a **pot**.

Wire-wound potentiometers are also common. Many have the same outward appearance as the potentiometer shown in Figure 1-32A. However, the internal construction is slightly different. As shown in Figure 1-32C, resistance wire is wound around an insulating core. A contact arm moves along the bare wire, changing the resistance between the center and outside terminals.

Another type of variable resistor is shown in Figure 1-32D. This type is sometimes called a sliding contact resistor. It is used in high power applications where the resistance value must be initially set or occasionally reset. The resistance value is changed when the sliding contact is moved along the bare resistance wire. Figure 1-32E shows the schematic diagram of the variable resistor.
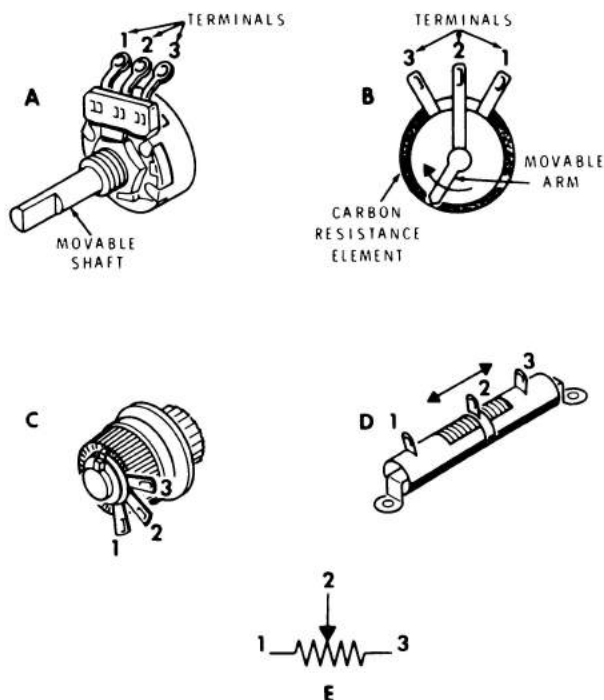
Figure 1-32
Variable resistors.

Courtesy Allen-Bradley.

Potentiometers are available in a wide
variety of shapes and sizes.

Resistor Connections

Resistors are often connected in series, in parallel, and in series-parallel combinations. In order to analyze and understand electronic circuits, you must be able to compute the total resistance of resistor networks.

A series circuit is one in which the components are connected end to end as shown in Figure 1-33A. Notice that the same current flows through all components. The current in the circuit must flow through all three resistors one after the other. Therefore, the total opposition to current flow is the total of the three resistances.

An example is shown in Figure 1-33B. Here, three resistors are connected in series. The total resistance is called R_T and is the resistance across the entire series circuit. You can find R_T by adding the individual resistor values together. That is:

$$R_T = R_1 + R_2 + R_3$$

$$R_T = 10\ \Omega + 20\ \Omega + 30\ \Omega$$

$$R_T = 60\ \Omega$$

These three resistors in series have the same opposition to current flow as a single 60 ohm resistor.

This example used three resistors. However, the same principle holds true for any number of series resistors. Thus, in Figure 1-33C, a 1 k ohm resistor in series with a 5 k ohm resistor has a total resistance of 6 k ohms or 6000 ohms. The circuit shown in Figure 1-33D uses six resistors. Some of the values are given in kilohms while others are given in ohms. To avoid confusion, we can convert all values to ohms, R_1, R_3, and R_5 are already expressed in ohms. R_2 is equal to 1 k ohms or 1000 ohms. R_4 is 1.2 k ohms or 1,200 ohms. And R_6 is 3.3 k ohms or 3300 ohms. Thus, you would find the total resistance (R_T) by adding:

$$R_T = R_1 + R_2 + R_3 + R_4 + R_5 + R_6$$

$$R_T = 500\ \Omega + 1000\ \Omega + 750\ \Omega + 1200\ \Omega + 600\ \Omega + 3300\ \Omega$$

$$R_T = 7350\ \Omega \text{ or } 7.35\ k\Omega$$
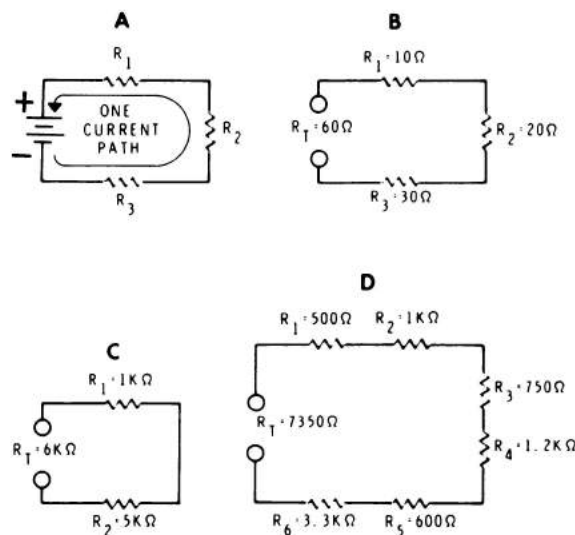
Figure 1-33
Resistors in series.
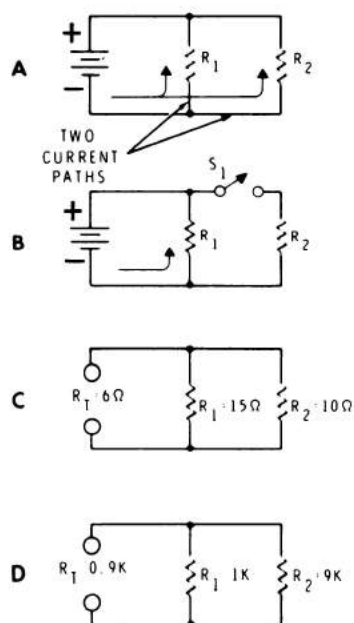
Figure 1-34
Resistors in parallel.

In **parallel** circuits components are connected across each other so that there are two or more paths for current flow. Figure 1-34A is an example.

To see how resistors in parallel act, let's add switch S_1 in series with R_2 so that we can switch R_2 in and out of the circuit. The resulting circuit is shown in Figure 1-34B. With S_1 open, a certain current will flow through R_1. The amount of current is determined by the resistance of R_1 and the applied voltage. Because there is only one path for current flow, the current through R_1 is the total circuit current.

Now let's close the switch and see what happens. The current through R_1 remains unchanged since neither the resistance of R_1 nor the applied voltage has changed. However, an additional current now flows through R_2. Thus, the total current provided by the battery has increased. If R_2 has the same resistance as R_1, both resistors will offer the same amount of opposition to current flow. Thus, the current through R_1 will equal the current through R_2. In this case, the current provided by the battery doubles when R_2 is switched in parallel with R_1.

Regardless of the size of R_1 and R_2, the total current provided by the battery will always increase when R_2 is placed in parallel with R_1 because a second current path is created. Obviously, then, the total opposition to current flow decreases since more current now flows. Thus, when one resistor is placed in parallel with another, the total resistance decreases. Let's assume that R_1 and R_2 are equal. When the switch is closed, the total current doubles. Thus, the total resistance has dropped to one half its former value.

The following simple formula allows you to find the total resistance of two resistors in parallel:

$$R_T = \frac{R_1 \times R_2}{R_1 + R_2}$$

Figure 1-34C shows an example. Here, a 15 ohm resistor (R_1) is in parallel with a 10 ohm resistor (R_2). Here's how you would find the total resistance, R_T.

$$R_T = \frac{R_1 \times R_2}{R_1 + R_2}$$

$$R_T = \frac{15\,\Omega \times 10\,\Omega}{15\,\Omega + 10\,\Omega}$$

$$R_T = \frac{150}{25}$$

$$R_T = 6\,\Omega$$

The total resistance of this circuit is the same as that of a 6 ohm resistor.

Figure 1-34D shows another example. Here, a 1 k ohm resistor (R_1) is in parallel with a 9 k ohm resistor (R_2).

$$R_T = \frac{R_1 \times R_2}{R_1 + R_2}$$

$$R_T = \frac{1000\,\Omega \times 9000\,\Omega}{1000\,\Omega + 9000\,\Omega}$$

$$R_T = \frac{9,000,000}{10,000}$$

$$R_T = 900\,\Omega \text{ or } 0.9\,\text{k}\Omega.$$

A R_T ? $R_1 = 100\Omega$ $R_2 = 100\Omega$ $R_3 = 400\Omega$ $R_4 = 400\Omega$
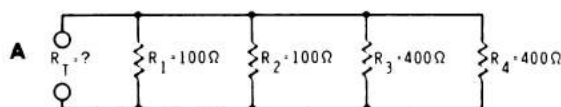
Figure 1-35
Compute the total resistance.

B R_T ? $R_1 = 500\Omega$ $R_2 = 500\Omega$ $R_3 = 250\Omega$
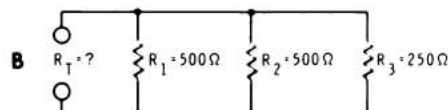
This formula is normally used when two resistors are in parallel. However, it can also be used with three or more resistors. For example, Figure 1-35A shows four resistors in parallel. Using the above formula, an equivalent resistance for R_1 and R_2 can be found. Then, using the formula again, an equivalent resistance for R_3 and R_4 can be found. Finally, the formula can be applied to the two equivalent resistances so that the total resistance can be found. However, this method involves using the formula three different times.

There is a better formula to use when more than two resistors are connected in parallel. The formula is:

$$R_T = \cfrac{1}{\cfrac{1}{R_1} + \cfrac{1}{R_2} + \cfrac{1}{R_3} + \cfrac{1}{R_4} + \cdots \cfrac{1}{R_n}}$$

Using this formula for the circuit shown in Figure 1-35A, you can find the total resistance (R_T) as follows:

$$R_T = \cfrac{1}{\cfrac{1}{100\ \Omega} + \cfrac{1}{100\ \Omega} + \cfrac{1}{400\ \Omega} + \cfrac{1}{400\ \Omega}}$$

$$R_T = \cfrac{1}{.01 + .01 + .0025 + .0025}$$

$$R_T = \cfrac{1}{.025}$$

$$R_T = 40\ \Omega$$

Figure 1-35B shows an example using only three resistors. Here's how you would find R_T.

$$R_T = \cfrac{1}{\cfrac{1}{R_1} + \cfrac{1}{R_2} + \cfrac{1}{R_3}}$$

$$R_T = \cfrac{1}{\cfrac{1}{500\ \Omega} + \cfrac{1}{500\ \Omega} + \cfrac{1}{250\ \Omega}}$$

$$R_T = \frac{1}{.002 + .002 + .004}$$

$$R_T = \frac{1}{.008}$$

$$R_T = 125\ \Omega$$

Sometimes two or more resistors that have the same value are placed in parallel. There is a simple rule which covers this situation. When all the resistors in parallel have the same value, you can find the total resistance by dividing that value by the number of resistors. For example, Figure 1-36A shows two 20 ohm resistors connected in parallel. Our rule states that:

$$R_T = \frac{\text{value of one resistor}}{\text{number of resistors in parallel}}$$

$$R_T = \frac{20\ \Omega}{2}$$

$$R_T = 10\ \Omega$$

Likewise, when three resistors of the same size are connected in parallel, the total resistance is one third the value of one resistor. Thus, in Figure 1-36B:

$$R_T = \frac{330\ \Omega}{3}$$

$$R_T = 110\ \Omega$$

In Figure 1-36C:

$$R_T = \frac{16\ k\Omega}{4}$$

$$R_T = 4\ k\Omega$$
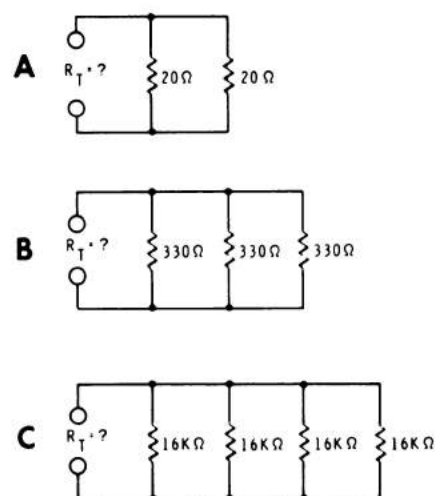
Figure 1-36
When all the resistors have the same value, R_T equals this value divided by the number of resistors in parallel.

Note that in any parallel circuit the total resistance is always less than the smallest resistor in the circuit.

In many circuits, a parallel circuit is connected in series with one or more resistors as shown in Figure 1-37A. This is known as a **series-parallel** circuit. The total resistance of this circuit is calculated using the formulas shown earlier. The procedure is to compute an equivalent resistance for the parallel circuit. Then this equavilent resistance is added to the series resistance values.

For example, you would first find the equivalent resistance for the parallel network made up of R_2 and R_3. Let's call this equivalent resistance R_A. Using the formula for a two-resistor parallel network, you will find:

$$R_A = \frac{R_2 \times R_3}{R_2 + R_3}$$

$$R_A = \frac{200\ \Omega \times 300\ \Omega}{200\ \Omega + 300\ \Omega}$$

$$R_A = \frac{60,000}{500}$$

$$R_A = 120\ \Omega$$

Now you can substitute R_A for the parallel network as shown in Figure 1-37B. Once you have the circuit in this simplified form, you use the formula for finding the total resistance in a series circuit.

$$R_t = R_1 + R_A + R_4$$

$$R_T = 100\ \Omega + 120\ \Omega + 300\ \Omega$$

$$R_T = 520\ \Omega$$
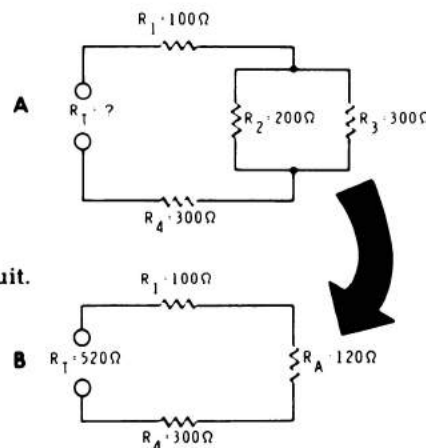
Figure 1-37
Simplifying the series-parallel circuit.

Conductance

Sometimes it is more convenient to think in terms of how well a material conducts current rather than how well it opposes current. Because of this, a property called conductance is often used. Conductance is just the opposite of resistance. It is defined as the ease with which a material passes current flow. Mathematically, conductance is the reciprocal of resistance. Stated as an equation:

$$\text{Conductance} = \frac{1}{\text{resistance}}$$

The letter G is used to represent conductance. Therefore, the equation can be written:

$$G = \frac{1}{R}$$

The unit of conductance is the **mho**. Notice that this is ohm spelled backwards.

Self-Review Questions

22. Define resistance _____

_____

_____

23. The unit of resistance is the _____.

24. Define conductance. _____

_____

_____

25. Name three types of resistor construction. _____

_____
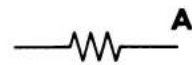
26. Figure 1-38A is the schematic symbol for a _____ .

Figure 1-38B is the schematic symbol for a _____.

27. A variable resistor is called a _____.

28. Find R_T for the series circuit shown in Figure 1-39A.
R_T = _____.

29. Find R_T for the parallel circuit shown in Figure 1-39B.
R_T = _____.

30. Find R_T for the parallel circuit shown in Figure 1-39C.
R_T = _____.

31. Find R_T for the series-parallel circuit shown in Figure 1-39D.
R_T = _____.

Figure 1-38
What do these schematic symbols represent?
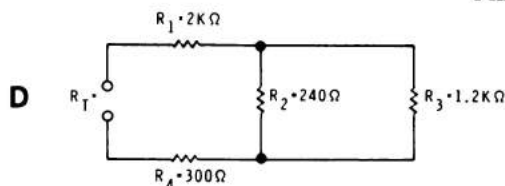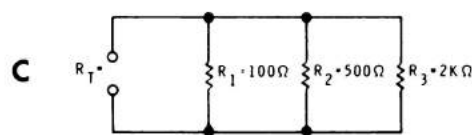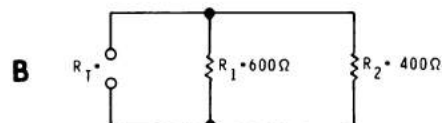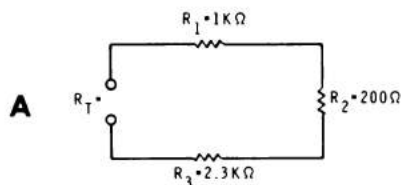
Figure 1-39
Find R_T for each of these circuits.

Self-Review Answers

22. Resistance is that property which opposes current flow.

23. The unit of resistance is the ohm.

24. Conductance is the ease with which a material passes current flow. It is the reciprocal of resistance.

25. Three types of resistor construction are wire-wound, carbon-composition, and deposited-film.

26. Figure 1-38A is the schematic symbol for a **resistor**.
    Figure 1-38B is the schematic symbol for a **potentiometer**.

27. A variable resistor is called a **potentiometer**.

28. R_T for the series circuit shown in Figure 1-39A is:

    $R_T = R_1 + R_2 + R_3$
    $R_T = 1 \text{ k}\Omega + 200 \text{ } \Omega + 2.3 \text{ k}\Omega$
    $R_T = 1000 \text{ } \Omega + 200 \text{ } \Omega + 2300 \text{ } \Omega$
    $R_T = 3500 \text{ } \Omega$
    $R_T = 3.5 \text{ k}\Omega$

29. R_T for the parallel circuit shown in Figure 1-39B is:

    $$R_T = \frac{R_1 \times R_2}{R_1 + R_2}$$

    $$R_T = \frac{600 \text{ } \Omega \times 400 \text{ } \Omega}{600 \text{ } \Omega + 400 \text{ } \Omega}$$

    $$R_T = \frac{240,000}{1,000}$$

    $$R_T = 240 \text{ } \Omega$$

30. R_T for the parallel circuit shown in Figure 1-39C is:

$$R_T = \cfrac{1}{\cfrac{1}{R_1} + \cfrac{1}{R_2} + \cfrac{1}{R_3}}$$

$$R_T = \cfrac{1}{\cfrac{1}{100\ \Omega} + \cfrac{1}{500\ \Omega} + \cfrac{1}{2000\ \Omega}}$$

$$R_T = \cfrac{1}{0.01 + 0.002 + 0.0005}$$

$$R_T = \cfrac{1}{0.0125}$$

$$R_T = 80\ \Omega$$

31. You must first find the resistance of the parallel network made up of R_2 and R_3. We will designate this resistance R_A.

$$R_A = \frac{R_2 \times R_3}{R_2 + R_3}$$

$$R_A = \frac{240\ \Omega \times 1200\ \Omega}{240\ \Omega + 1200\ \Omega}$$

$$R_A = \frac{288{,}000}{1440}$$

$$R_A = 200\ \Omega$$

Now substitute R_A for the parallel network as shown in Figure 1-40.

$$R_T = R_1 + R_A + R_4$$

$$R_T = 2000\ \Omega + 200\ \Omega + 300\ \Omega$$

$$R_T = 2500\ \Omega$$

$$R_T = 2.5\ k\Omega$$
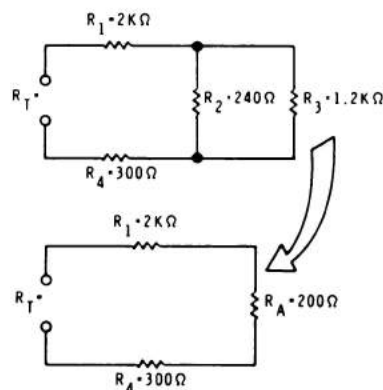
Figure 1-40
Solution for Self-Review Question 30.

From this complex transceiver to a simple flashlight, the operation of all electronic equipment is based on Ohm's law.

OHM'S LAW

Voltage, current, and resistance are all interrelated quantities in an electric circuit. The law that relates these three quantities is Ohm's law.

Finding Current

Ohm's law states that current is directly proportional to voltage and inversely proportional to resistance. Figure 1-41 will help illustrate this point. The source of voltage is the battery. Voltage is the force which causes current to flow. Therefore, the higher the voltage, the higher the current will be. Conversely, the lower the voltage, the lower the current will be. This assumes that the resistance remains constant. However, as we have seen, the current is also determined by the resistance (remember: resistance is the opposition to current flow). Assuming that the voltage is constant, the higher the resistance, the lower the current will be. Also, the lower the resistance, the higher the current will be.
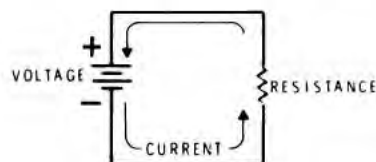
Figure 1-41

Ohm's law states that current is directly proportional to voltage and inversely proportional to resistance.

Mathematically, Ohm's law can be expressed like this:

$$I = \frac{E}{R}$$

Where: E = voltage, expressed in volts.
I = current, expressed in amperes.
R = resistance, expressed in ohms.

The equation states that the current in a circuit is equal to the voltage divided by the resistance. This formula may be used to find current in any circuit in which the voltage and resistance are known.

Figure 1-42 shows a circuit in which the values of voltage and resistance are given. To determine the current, substitute the known values into the formula:

$$I = \frac{E}{R}$$

$$I = \frac{10V}{5\,\Omega}$$

$$I = 2A$$

Thus, the current in the circuit shown in Figure 1-42 is 2A or 2 amperes.
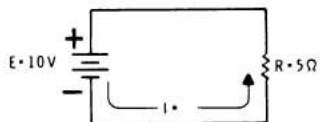
Figure 1-42
Current is determined by voltage and resistance.
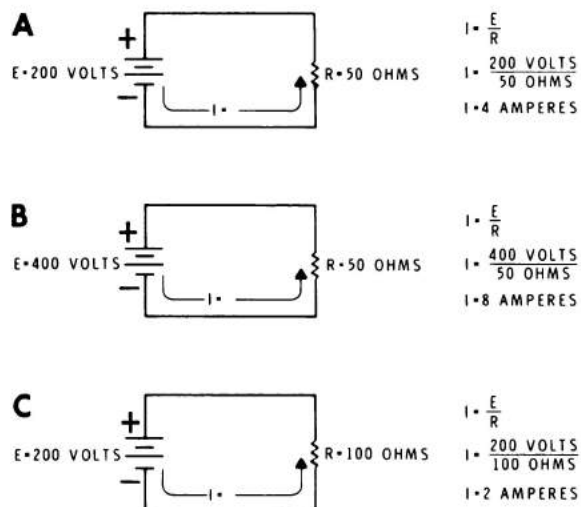
Figure 1-43
Using Ohm's law to find current.

Another example is shown in Figure 1-43A. The current in this circuit is:

$$I = \frac{E}{R}$$

$$I = \frac{200 \text{ V}}{50 \text{ }\Omega}$$

$$I = 4 \text{ A}$$

Figure 1-43B shows what happens when the applied voltage is doubled.

$$I = \frac{E}{R}$$

$$I = \frac{400 \text{ V}}{50 \text{ }\Omega}$$

$$I = 8\text{A}$$

Notice that when the voltage is doubled, the current also doubles. This is because current is directly proportional to voltage.

What happens if the resistance is doubled and voltage remains the same? This situation is shown in Figure 1-43C.

$$I = \frac{E}{R}$$

$$I = \frac{200\ V}{100\ \Omega}$$

$$I = 2A$$

When the resistance is doubled, the current is reduced to one-half its former value. This is because current is inversely proportional to resistance.
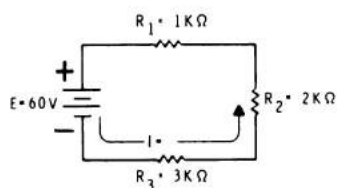
Figure 1-44
What is the current in this series circuit?

What is the current in the circuit of Figure 1-44? You will have to find total circuit resistance (R_T) before you can apply Ohm's law.

$$R_T = R_1 + R_2 + R_3$$

$$R_T = 1\ k\Omega + 2\ k\Omega + 3\ k\Omega$$

$$R_T = 6\ k\Omega$$

We can now use Ohm's law to find the current. The easiest way to solve this problem is to convert 6 kΩ to 6,000 Ω.

$$I = \frac{E}{R}$$

$$I = \frac{60\ V}{6000\ \Omega}$$

$$I = 0.01\ A$$

$$I = 10\ mA$$

Finding Current in Parallel Circuits

In parallel circuits, components are connected across each other so there are two or more paths or branches for current flow. This is shown in Figure 1-45. Notice that the total current is equal to the sum of the branch currents. This can be expressed mathematically as:

$$I_T = I_1 + I_2$$
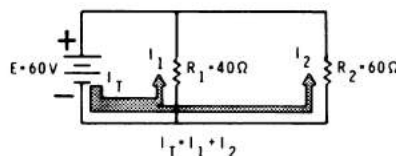
Figure 1-45
The total current in a parallel circuit is equal to the sum of
the branch currents.

This formula can be extended to include any number of parallel branches.

Notice that both R_1 and R_2 are connected directly across the battery. This means that the voltage across both resistors is equal to the battery voltage. This is true for any parallel circuit. The voltage is the same across parallel branches.

You can find the branch current using Ohm's law. To find the current through R_1, divide the voltage by R_1. Thus,

$$I_1 = \frac{E}{R_1}$$

$$I_1 = \frac{60\ V}{40\ \Omega}$$

$$I_1 = 1.5\ A$$

The current through R_2 can be found in a similar manner.

$$I_2 = \frac{E}{R_2}$$

$$I_2 = \frac{60\ V}{60\ \Omega}$$

$$I_2 = 1\ A$$

The total current is equal to the sum of the branch currents. Therefore:

$$I_T = I_1 + I_2$$
$$I_T = 1.5 \text{ A} + 1 \text{ A}$$
$$I_T = 2.5 \text{ A}$$

The total current in the circuit of Figure 1-45 is 2.5 amperes.

Finding Voltage

You have seen that the formula for current is:

$$I = \frac{E}{R}$$

You can transpose this equation to find voltage. (Transpose means to change the equation from one form to another.) Therefore, you must change the formula so that E is on one side of the equation by itself. This is easy to do if you remember a basic algebra rule for transposing equations. This rule states that you can multiply or divide both sides of the equation by any quantity without changing the equality. The equation for current is:

$$I = \frac{E}{R}$$

Multiplying both sides of the equation by R, we have:

$$I \times R = \frac{E \times R}{R}$$

Notice that R appears in both the numerator and the denominator of the fraction on the right side of the equation. The two R's in the fraction can be canceled like this:

$$I \times R = \frac{E \times \cancel{R}}{\cancel{R}}$$

This leaves:

$$I \times R = E$$

You can now reverse the two sides of the equation without changing the equality. Therefore:

$$E = I \times R$$

This equation states that voltage is equal to current times resistance. Generally, the times sign ($\times$) is omitted so the equation is written:

$$E = IR$$
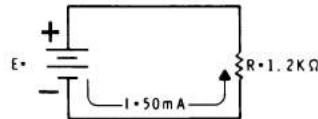
Figure 1-46
Ohm's law states that current times resistance equals voltage.

Figure 1-46 shows a circuit in which the values of resistance and current are given. To determine the voltage, substitute the known values into the formula:

$$E = IR$$

$$E = 50 \text{ mA} \times 1.2 \text{ k}\Omega$$

Before solving the equation, you must convert the values to amperes and ohms:

$$E = 0.05 \text{ A} \times 1200 \ \Omega$$

$$E = 60 \text{ V}$$

Figure 1-47
The sum of the voltage drops will always equal the voltage rise.
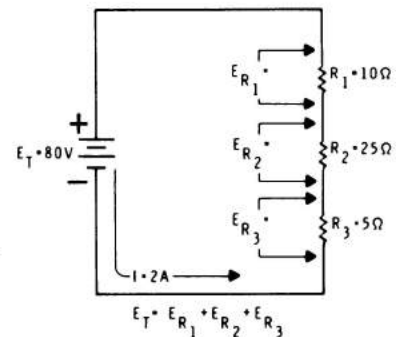
Figure 1-47 shows a slightly different problem. Here you must find the voltage drops across R_1, R_2, and R_3. We have called these voltages E_{R1}, E_{R2}, and E_{R3}. Since you know that $R_1 = 10 \ \Omega$ and that the current through R_1 is 2 A, you can find E_{R1} using Ohm's law.

$$E_{R1} = I \times R_1$$

$$E_{R1} = 2\ A \times 10\ \Omega$$

$$E_{R1} = 20\ V$$

You can find E_{R2} and E_{R3} the same way:

$$E_{R2} = I \times R_2$$

$$E_{R2} = 2\ A \times 25\ \Omega$$

$$E_{R2} = 50\ V$$

$$E_{R3} = I \times R_3$$

$$E_{R3} = 2\ A \times 5\ \Omega$$

$$E_{R3} = 10\ V$$

In an earlier section you saw that the sum of the voltage drops will always equal the voltage rise. This can be expressed mathematically for the circuit of Figure 1-47 as:

$$E_T = E_{R1} + E_{R2} + E_{R3}$$

Where

$$E_T = \text{voltage rise}$$

$$E_{R1-3} = \text{voltage drops}$$

We can prove that the sum of the voltage drops equals the voltage rise by substituting the values of Figure 1-47 into the equation:

$$E_T = E_{R1} + E_{R2} + E_{R3}$$

$$E_T = 20\ V + 50\ V + 10\ V$$

$$E_T = 80\ V$$

The voltage rise is 80 volts and the sum of the voltage drops is 80 volts.

Finding Resistance

You can find the formula for resistance by transposing the current or voltage formula. For example, the voltage formula is:

$$E = IR$$

Dividing both sides by I you get:

$$\frac{E}{I} = \frac{IR}{I}$$

The I's in the fraction on the right cancel:

$$\frac{E}{I} = \frac{\cancel{I}R}{\cancel{I}}$$

This leaves:

$$\frac{E}{I} = R$$

Reversing the equation, you have the formula for resistance:

$$R = \frac{E}{I}$$

This states that resistance is equal to voltage divided by current. Using this formula, you can find the resistance in any circuit in which the voltage and current are known.
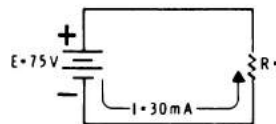
Figure 1-48

Ohm's law states that voltage divided by current equals resistance.

Figure 1-48 shows a circuit in which the voltage and current are known. Let's find the resistance.

$$R = \frac{E}{I}$$

$$R = \frac{75 \text{ V}}{30 \text{ mA}}$$

You must first convert from milliamperes to amperes before solving the equation. Therefore:

$$R = \frac{75 \text{ V}}{0.03 \text{ A}}$$

$$R = 2500 \text{ }\Omega$$

$$R = 2.5 \text{ k}\Omega$$
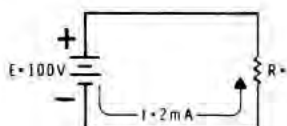
Figure 1-49
What is the resistance of this circuit?

Figure 1-49 shows another example. Solving for R, you find that:

$$R = \frac{E}{I}$$

$$R = \frac{100 \text{ V}}{2 \text{ mA}}$$

Converting milliamperes to amperes, you have:

$$R = \frac{100 \text{ V}}{0.002 \text{ A}}$$

$$R = 50,000 \text{ }\Omega$$

$$R = 50 \text{ k}\Omega$$

Summary

Ohm's law may be expressed by three different formulas:

$$I = \frac{E}{R}$$

$$E = IR$$

$$R = \frac{E}{I}$$

A convenient way to remember these three formulas is the Ohm's law circle shown in Figure 1-50A. To use this circle, cover the unknown quantity with your thumb; the desired equation will be the remaining terms. For example, if the unknown quantity is current, cover the I as shown in Figure 1-50B. The quantity covered is the left hand side of the equation. The remaining two symbols represent the right hand side of the equation. Therefore,

$$I = \frac{E}{R}$$

Figures 1-50C and D show how to find the formulas for resistance and voltage.
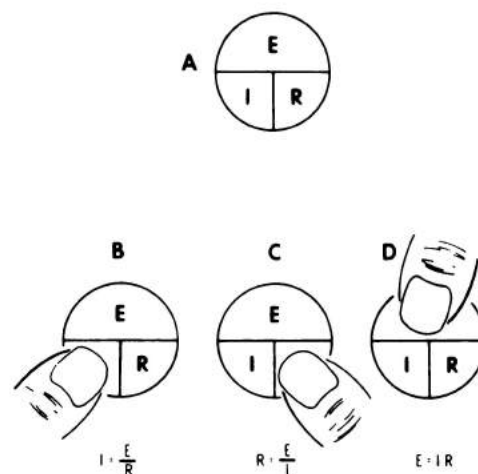
Figure 1-50
The Ohm's law circle.

Self-Review Questions

32. Write the three equation forms of Ohm's law.

33. Find the current in the circuit of Figure 1-51A. I = _____.

34. Find the applied voltage in the circuit of Figure 1-51B. E = _____.

35. Find the resistance of the circuit shown in Figure 1-51C. R = _____.

A

$E \cdot 160V$ $\quad$ $R_1 \cdot 200\Omega$

$I \cdot$

$R_2 \cdot 3K\Omega$

B

$E \cdot$ $\quad$ $R \cdot 4K\Omega$

$I \cdot 3mA$

C

$E \cdot 12V$ $\quad$ $R \cdot$

$I \cdot 40mA$
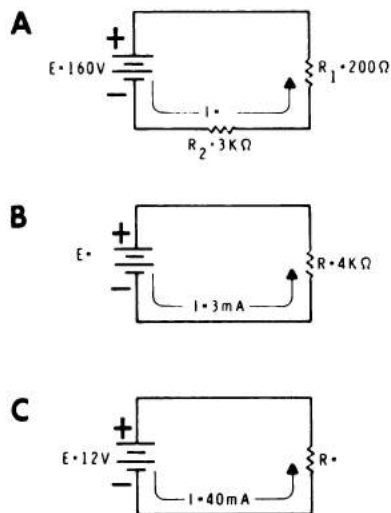
Figure 1-51
Use Ohm's law to find the unknown quantity.

Self-Review Answers

32. The three equation forms of Ohm's law are:

$$I = \frac{E}{R}$$

$$E = IR$$

$$R = \frac{E}{I}$$

33. To find current, you must first find the total resistance:

$$R_T = R_1 + R_2$$

$$R_T = 200\,\Omega + 3\,k\Omega$$

$$R_T = 200\,\Omega + 3000\,\Omega$$

$$R_T = 3200\,\Omega$$

$$R_T = 3.2\,k\Omega$$

Now use Ohm's law to find current:

$$I = \frac{E}{R_T}$$

$$I = \frac{160\,V}{3200\,\Omega}$$

$$I = 0.05A$$

$$I = 50\,mA$$

34. To find the applied voltage, use the Ohm's law equation:

$$E = IR$$

$$E = 3\,mA \times 4\,k\Omega$$

$$E = 0.003\,A \times 4000\,\Omega$$

$$E = 12\,V$$

35. Resistance is found by using the Ohm's law equation.

$$R = \frac{E}{I}$$

$$R = \frac{12 \text{ V}}{40 \text{ mA}}$$

$$R = \frac{12 \text{ V}}{0.04 \text{ A}}$$

$$R = 300 \ \Omega$$

POWER

In addition to the three basic electrical quantities (current, voltage, and resistance), a fourth quantity is also very important. This quantity is called power. **Power** is the amount of work done in a specific length of time.

The Watt

The unit of electrical power is the **watt**. One watt is equal to the work done by one volt moving a current of one ampere.

The abbreviation of watt is W. Thus, 4.2 watts is abbreviated 4.2 W. As with other electrical quantities, metric prefixes are attached to indicate smaller and larger units of power. Thus, one milliwatt equals 1/1000 watt, while one microwatt equals 1/1,000,000 watt. Also, one kilowatt equals 1000 watts while one megawatt equals 1,000,000 watts. In electronics, the letter P is used to represent power. Thus, in the shorthand of electronics, you could write the statement "the power is 5 watts" as an equation, "P = 5 W."

Power, Current, and Voltage

Power is directly proportional to both current and voltage. Thus, the formula for power is:

$$P = IE$$

where   P = power in watts

   I = current in amperes

   E = voltage in volts

As with Ohm's law, there are two other useful forms of this equation. The first expresses voltage in terms of current and power. It is found by rearranging the equation:

$$P = IE$$

Dividing both sides by I, we have:

$$\frac{P}{I} = \frac{IE}{I}$$

Canceling the two I's on the right side:

$$\frac{P}{I} = \frac{\cancel{I}E}{\cancel{I}}$$

Reversing the two sides of the equation:

$$E \quad = \frac{P}{I}$$

This states that voltage is equal to power divided by current. Another useful form of the equation is:

$$I \quad = \frac{P}{E}$$

See if you can derive this formula from the original power formula. To summarize, power, voltage, and current are related by the following formulas:

$$P \quad = IE$$

$$E = \frac{P}{I}$$

$$I = \frac{P}{E}$$

Now, let's work some sample problems using these formulas.

What is the power dissipated in a circuit in which an emf of 50 volts is applied and a current of 3.2 amperes flows?

$$P = IE$$

$$P = 3.2 \text{ A} \times 50 \text{ V}$$

$$P = 160 \text{ W}$$

How much current flows through a 75-watt light bulb that is connected across a 120 volt power line?

$$I = \frac{P}{E}$$

$$I = \frac{75 \text{ W}}{120 \text{ V}}$$

$$I = 0.625 \text{ A}$$

$$I = 625 \text{ mA}$$

What is the voltage drop across a light bulb that dissipates 60 watts when the current through the bulb is 0.5 amperes?

$$E = \frac{P}{I}$$

$$E = \frac{60 \text{W}}{0.5 \text{A}}$$

$$E = 120 \text{ V}$$

Power Dissipation in Resistors

In resistors and most other electronic components, power is dissipated in the form of heat. In some cases, this heat is a desired result, as in the resistance elements in toasters, heating pads, and electric stoves. But in most electronic devices the heat produced by resistors represents wasted power and since electrical power costs money, attempts are generally made to keep the power lost in resistors at a minimum.

Since resistors dissipate power, there must be some formula for power which involves resistance. Actually, there are two such formulas. One expresses power in terms of voltage and resistance. The other expresses power in terms of current and resistance. Let's see how these equations are derived.

First, consider how P can be expressed in terms of E and R. The basic formula for power is:

$$P = I \times E \quad \text{or} \quad P = E \times I$$

However, from Ohm's law you know that

$$I = \frac{E}{R}$$

Thus, you can substitute

$$\frac{E}{R}$$

for I in the basic power equation. When you do this, the equation becomes:

$$P = E \times \frac{E}{R}$$

or,

$$P = \frac{E^2}{R}$$

This formula is used when only the voltage and resistance are known and you wish to find the power. For example, how much power is dissipated by a 22 ohm resistor if the voltage drop across the resistor is 5 volts?

$$P = \frac{E^2}{R}$$

$$P = \frac{(5 \text{ V})^2}{22 \text{ } \Omega}$$

$$P = \frac{25}{22}$$

$$P = 1.136 \text{ W}$$

In some cases, only the current and resistance will be known. By combining the basic power formula with one of the Ohm's law formulas, you can derive an equation in which P is expressed in terms of I and R. Recall that the basic power formula is:

$$P = I \times E$$

From Ohm's law, we know that E = IR. Thus, you can substitute IR for E in the power formula. The equation becomes:

$$P = I \times IR$$

$$P = I^2R$$
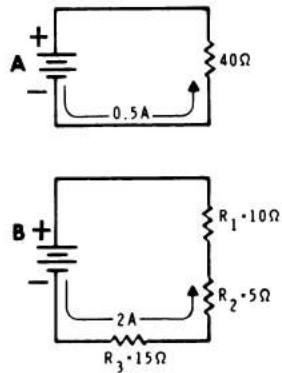
Figure 1-52
Finding power when R and I are known.

This is the formula you use when only current and resistance are known and you wish to find power. For example, how much power is dissipated by the circuit shown in Figure 1-52A?

$$P = I^2R$$

$$P = (0.5 \text{ A})^2 \times 40 \text{ } \Omega$$

$$P = 0.25 \times 40$$

$$P = 10 \text{ W}$$

How much power is dissipated by R_1 in Figure 1-52B?

$$P = I^2R_1$$

$$P = (2 \text{ A})^2 \times 10 \text{ } \Omega$$

$$P = 4 \times 10$$

$$P = 40 \text{ W}$$

The Wheel Diagram

There are also a number of other formulas you can derive from the basic power formula and Ohm's law.

Figure 1-53 is a wheel diagram that shows all the formulas in an easy-to-use form. The inner circle contains the four basic units: power, current, voltage, and resistance. The outer circle contains three formulas for each quantity. For example, the three formulas for power are:

$$P \quad = I \times E$$

$$P \quad = I^2 \times R$$

$$P \quad = \frac{E^2}{R}$$

Using this diagram, you can quickly find the proper formula for any problem in which the two quantities are known.
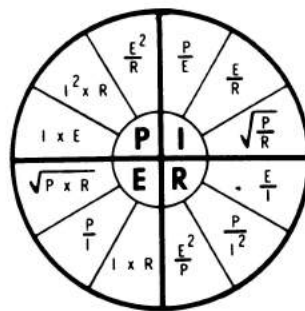
Figure 1-53
Wheel diagram for finding the proper equation for current,
voltage, resistance, and power.

Self-Review Questions

36. What is power? _____

    _____

37. The unit of electrical power is the _____ .

38. Write the basic power formula. _____.

39. What is the power dissipated when the applied voltage is 800 V and the current is 300 mA? P = _____

40. What is the voltage drop across a light bulb that dissipates 75 watts when the current through the bulb is 250 mA? E = _____

41. How much current flows through a 50-watt light bulb which is connected across a 250-volt power supply? I = _____

42. When an 8 Ω resistor has a voltage drop of 10 V, how much power is dissipated? P = _____

43. How much power is dissipated by a 100 Ω resistor when 200 mA flows through it? P = _____

Self-Review Answers

36. Power is the amount of work done in a specific length of time.

37. The unit of electrical power is the **watt**.

38. The basic power formula is:

    $P = IE$

39. $P = IE$

    $P = 300 \text{ mA} \times 800 \text{ V}$

    $P = 0.3 \text{ A} \times 800 \text{ V}$

    $P = 240 \text{ W}$

40. $E = \dfrac{P}{I}$

    $E = \dfrac{75 \text{ W}}{250 \text{ mA}}$

    $E = \dfrac{75 \text{ W}}{0.25 \text{ A}}$

    $E = 300 \text{ V}$

41. $I = \dfrac{P}{E}$

    $I = \dfrac{50 \text{ W}}{250 \text{ V}}$

    $I = 0.2 \text{ A}$

    $I = 200 \text{ mA}$

42. $P = \dfrac{E^2}{R}$

       $P = \dfrac{(10 \text{ V})^2}{8 \text{ } \Omega}$

       $P = \dfrac{100}{8}$

       $P = 12.5 \text{ W}$

43. $P = I^2 R$

       $P = (200 \text{ mA})^2 \times 100 \text{ } \Omega$

       $P = (0.2 \text{ A})^2 \times 100 \text{ } \Omega$

       $P = 0.04 \times 100$

       $P = 4 \text{ W}$

MAGNETISM
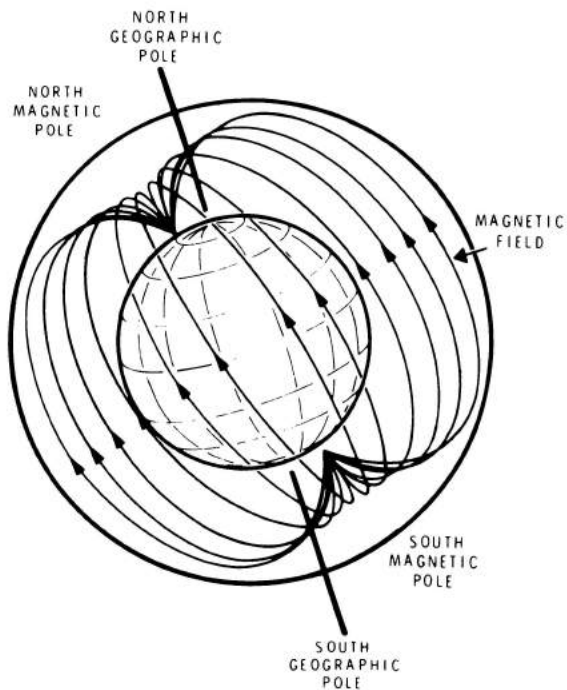
Figure 1-54
The Earth's magnetic field.

Most of us are familiar with the behavior of magnets. We know that they attract certain metallic materials. We also are aware that a compass works on the principles of magnetism. The compass needle aligns itself with the Earth's magnetic field. This magnetic field is shown in Figure 1-54.

The properties of magnetism are also used in many everyday items, such as loudspeakers, relays, and electric motors. Even a television set uses magnetism to help create the picture you see on the screen.

The Magnetic Field

Individual magnets also have a magnetic field, including north and south poles. This is shown in Figure 1-55A. The magnetic "lines of force" or **flux lines**, which make up the magnetic field, are invisible. However, if you place a magnet under a piece of paper and sprinkle iron filings over the paper, the iron filings will align themselves with these flux lines. You will then see evidence of the flux lines, as shown in Figure 1-55B.

A

B

IRON FILINGS
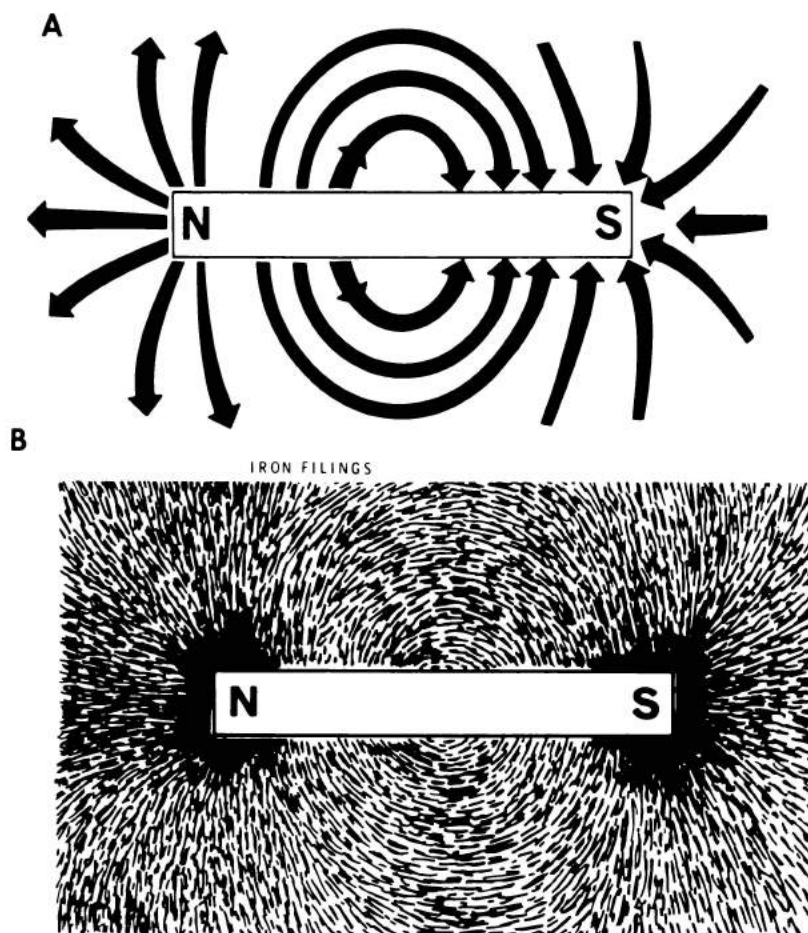
Figure 1-55
Flux lines surround the magnet.
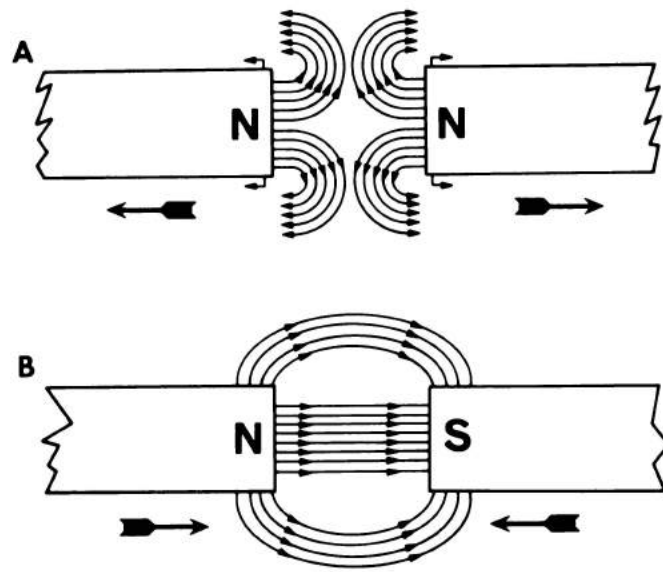
Figure 1-56
Like poles repel, unlike poles attract.

Laws of Magnetism

The attraction or repulsion between magnets is caused by the flux lines. Figure 1-56A shows that when two north poles are brought together, the flux lines push away from each other and cause the magnets to physically move apart. When the north and south poles of the magnets are brought together, as in Figure 1-56B, the flux lines join each other and physically pull the magnets together. The laws of magnetism are:

Like poles repel.

Unlike poles attract.

Electricity and Magnetism

Electricity and magnetism are closely related. When current flows through a wire, a magnetic field is developed around the wire. This is shown in Figure 1-57. The strength of the magnetic field is directly proportional to the current. When current increases, the magnetic field will also increase.
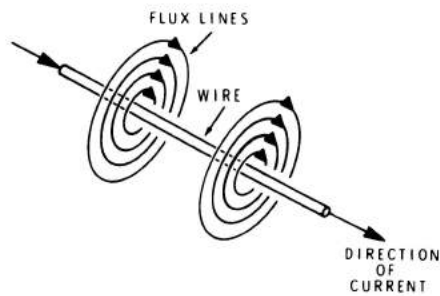
Figure 1-57
Current develops a magnetic field around a wire.

You can also increase the magnetic field by forming the wire into a loop. This is shown in Figure 1-58. Notice that this also creates north and south poles. In fact, this loop of wire has all the properties of a permanent magnet. The difference is that you can vary the intensity of the magnetic field by changing the amount of current through the loop. Another way to increase the intensity of the magnetic field is to add more turns of wire. This will form a **coil**.
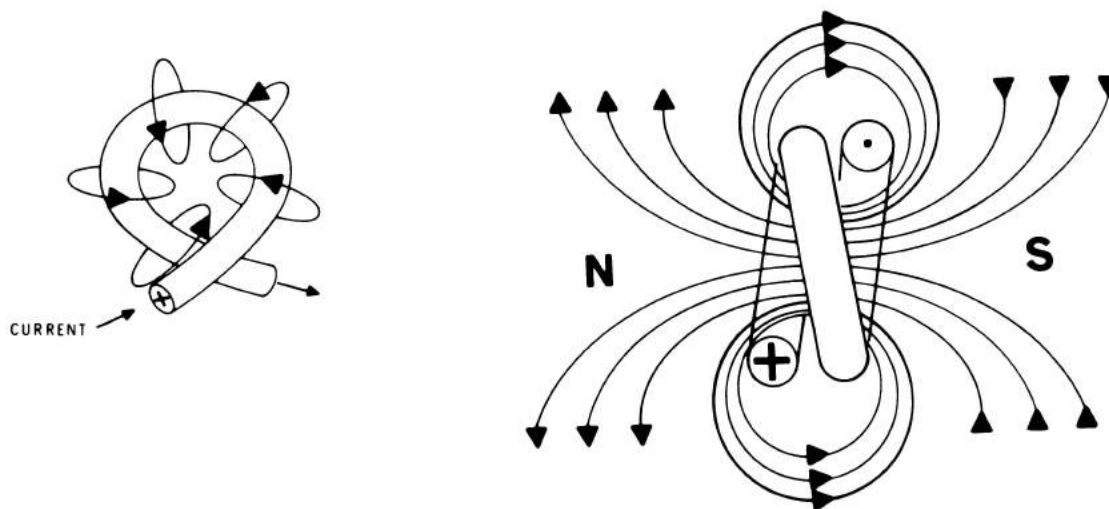
Figure 1-58
Flux lines around a loop of wire.
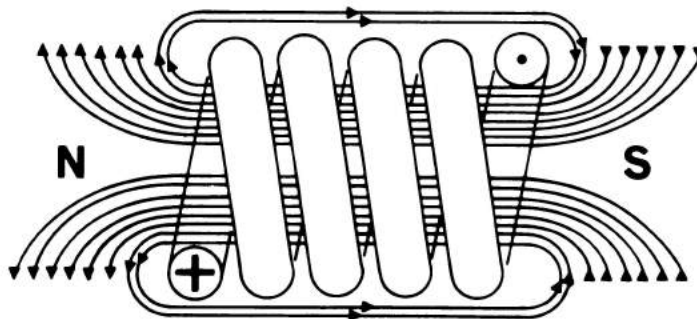
Figure 1-59
Magnetic field around a coil.

A coil of wire is shown in Figure 1-59. Note that there are many more flux lines present than with a single loop. The force which produces the flux lines in a coil is called **magnetomotive force**. The amount of magnetomotive force is directly proportional to the current and the number of turns in the coil. The unit of measure of magnetomotive force is the **ampere-turn**. This is the amount of force developed by one turn of wire when the current is one ampere.

You have seen two ways to increase the strength of the magnetic field around a coil. One way is to increase the current. Another is to increase the number of turns. However, a third method is the most dramatic improvement of all. If we insert an iron bar into the center of the coil, the strength of the magnetic field will greatly increase. The iron **core** can support many more flux lines than air. It is said to have a higher **permeability** than air. Permeability is the ability of a material to concentrate magnetic flux. Air has a permeability of one, while iron has a permeability of 7,000. The iron bar offers much less opposition to the flux lines than air. It, therefore, greatly increases the magnetic field.

Applications of Magnetism

The relay and the loudspeaker are two common applications of magnetism.
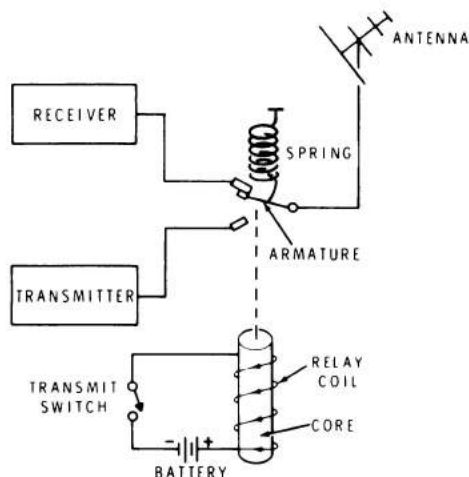
Figure 1-60
A relay can be used as a transmit-receive switch.

The relay is one of the simplest applications of magnetism, and it is also one of the more useful. Figure 1-60 shows how it operates. When the switch is closed, current flows from the battery through the relay coil. The current develops a magnetic field in the core which attracts the armature, pulling it down. This causes the lower contacts to close. The relay is being used here to change an antenna from the receiver to transmitter. When the switch is open, the antenna is connected to the receiver. When the switch is closed, the relay energizes, connecting the antenna to the transmitter. The schematic symbol for this relay is shown in Figure 1-61.
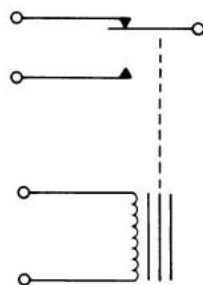
Figure 1-61
The schematic symbol for a relay.
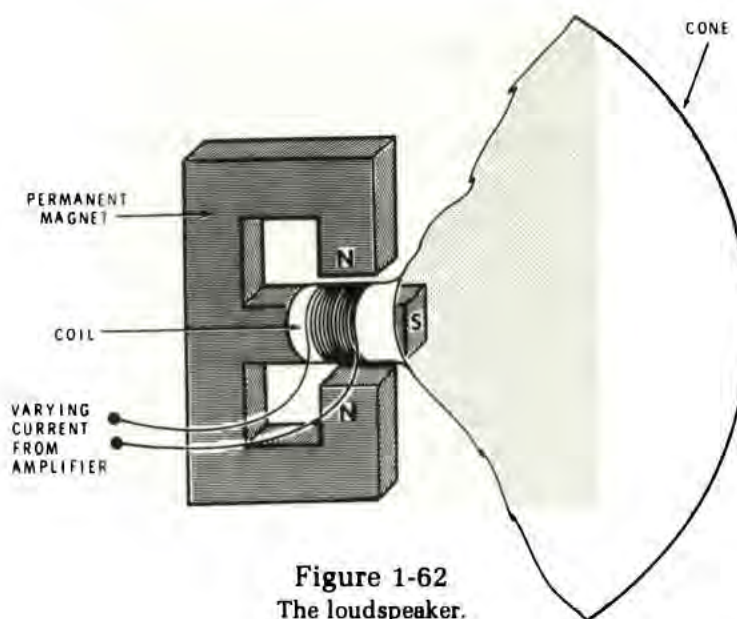
Figure 1-62
The loudspeaker.

Loudspeakers are used in all types of audio equipment. Most of them use a moving coil and a permanent magnet. The magnet establishes a strong, stationary magnetic field. The movable coil is placed around part of the magnet and in its magnetic field. A cut-away diagram of a loudspeaker is shown in Figure 1-62.

The coil is forced to move back and forth by the current it receives from the amplifier. This current, which varies at an audio rate, causes a correspondingly varying magnetic field around the coil. The coil is alternately attracted and repelled by the stationary field of the permanent magnet. Thus, the coil moves back and forth at the same rate as the varying current.

The moving coil is attached to a large cone or diaphragm. As the coil vibrates the cone also vibrates, setting the air around the cone in motion at the same rate. This reproduces the original sound. The schematic symbol for a loudspeaker is shown in Figure 1-63.
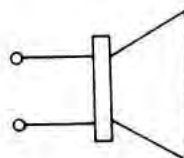
Figure 1-63
The schematic symbol for a loudspeaker.

Self-Review Questions

44. State the basic law of magnetism. _____

_____

_____

45. The force which produces the flux lines in a coil is called _____ .

46. What is the ampere-turn? _____

_____

47. The ability of a material to concentrate magnetic flux is called _____ .

Self-Review Answers

44. The basic law of magnetism is: like poles repel, unlike poles attract.

45. The force which produces the flux lines in a coil is called **magnetomotive force**.

46. The ampere-turn is the unit of magnetomotive force.

47. The ability of a material to concentrate magnetic flux is called **permeability**.

METERS

The most basic measuring instrument you will use is the meter. In this section, we will discuss the three basic meters: the **ammeter**, the **voltmeter**, and the **ohmmeter**. These are used to measure current, voltage, and resistance.
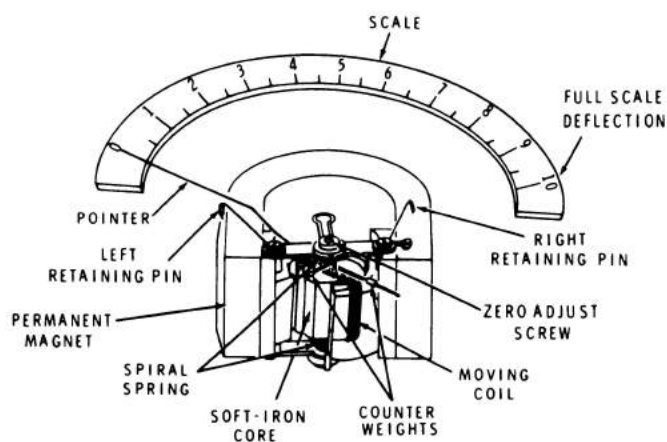
Figure 1-64
Permanent magnet, moving-coil meter movement.

The Meter Movement

Figure 1-64 shows the construction of the basic meter movement. Several important parts are listed. We will discuss these parts in detail, starting with the permanent magnet.
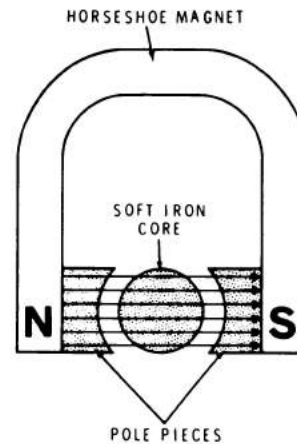
Figure 1-65
The permanent magnet.

Figure 1-65 shows the permanent magnet system. The horseshoe magnet produces a stationary magnetic field. To concentrate this magnetic field in the area of the moving coil, pole pieces are added to the magnet. These are made of soft iron, which concentrates the lines of flux in this area. Also, a stationary soft-iron core is placed between the pole pieces. Enough space is left between the pole pieces and the core so that the moving coil can rotate freely in this space. As you can see, the pole pieces and core restrict most of the flux to the area of the moving coil.
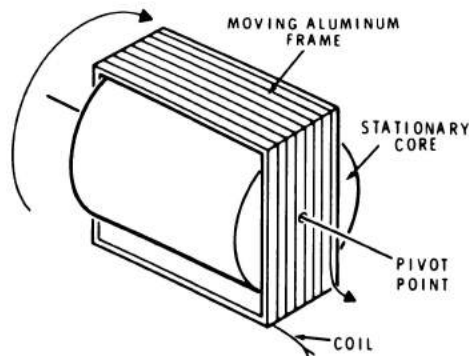
Figure 1-66
The moving coil.

Figure 1-66 shows how the moving coil fits around the soft iron core. The coil consists of many turns of extremely fine wire on an aluminum frame. The aluminum frame is very light so that little torque is needed to move it. The two ends of the coil connect to the leads of the ammeter, voltmeter, or ohmmeter.
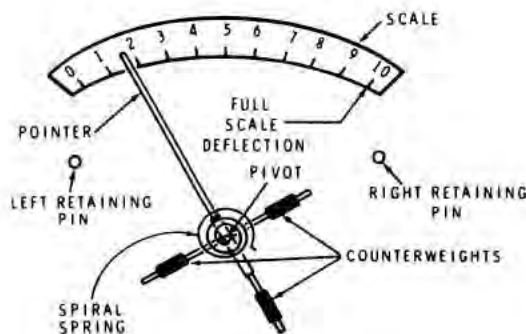
Figure 1-67
The pointer assembly.

Figure 1-67 shows the details of the pointer assembly. The pointer is attached to the moving coil so that it moves when the coil does. Counterweights are often attached to the pointer so that a perfect balance is achieved. This makes the pointer easier to move and helps the meter read the same in all positions. A well-balanced meter will read the same whether held vertically or horizontally. Retaining pins on either side of the movement limit the distance that the pointer and other rotating parts can move. Two spiral springs at opposite ends of the moving coil force the pointer back to the zero position when no current is flowing through the coil. In most movements, the spiral springs are also used to apply current to the moving coil. The two ends of the coil connect to the inner ends of the spiral spring. The outer end of the rear spring is fixed in place. However, the outer end of the front spring connects to a zero adjust screw. This allows you to set the pointer to exactly the zero point on the scale when no current is flowing through the coil.

The moving coil, pointer, and counter weight rotate around a pivot point. To hold the friction to an absolute minimum, jeweled bearings are used at this point just as they are in a fine watch.

Now that you have a good idea of how the meter is constructed, we will discuss how it operates. It is possible to create a motor effect by passing current through a conductor. The magnetic field set up around the conductor, due to the current flow, will push away from the magnetic field of the permanent magnet. This will cause the conductor to move if the magnetic field is strong enough. This is the principle upon which electric **motor** operation is based. The meter movement is actually a small motor.

When current flows through the moving coil, the coil is forced to move, due to the motor action. When the coil moves, the pointer will move up scale. The amount of torque produced by this tiny motor is proportional to the magnitude of the current which flows through the moving coil. The more current, the greater the torque will be and the further the pointer will be deflected.
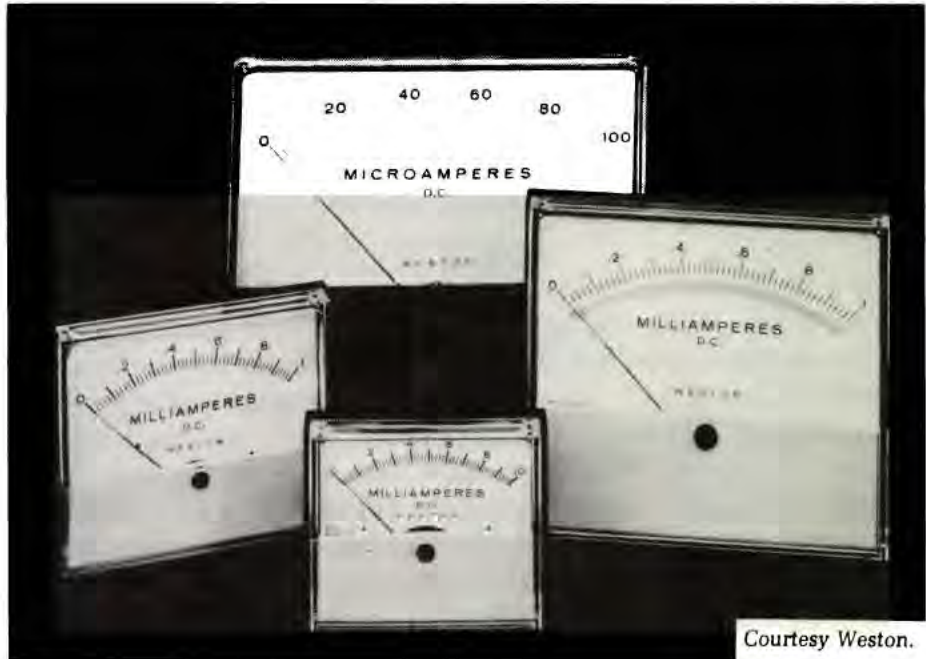
Courtesy Weston.

Meter movements come in different shapes, sizes, and sensitivities.

Meter movements are rated by the amount of current required to produce full-scale deflection. For example, a 50 microampere meter movement deflects full scale when only 50 microamperes of current flow through it. The 50 μA meter movement is one of the most commonly used types of meter movements. The 100 μA and 200 μA movements are also popular.

The d'Arsonval meter movement, which we have been discussing, is capable of measuring only direct current or DC. When alternating current or AC is applied to the meter movement, the pointer will attempt to change its direction of deflection each time the alternating current changes its direction. If the current changes direction more than a few times each second, the coil cannot follow the changes. Therefore, the d'Arsonval meter movement is a "DC only" meter.

Ammeter

The meter movement discussed in the previous section is basically a current meter. That is, it deflects when current flows through it. A current meter is also called an **ammeter**.

When using an ammeter to measure current, you must place the ammeter in the circuit in **series**. That is, you have to break the circuit at the point where the current is to be measured, and insert the ammeter. You must also observe polarity, the current must flow through the ammeter from the negative terminal to the positive terminal. One other precaution you must take is to be sure that the current to be measured does not exceed the rating of the ammeter. Failure to observe these precautions could result in damage to the meter.
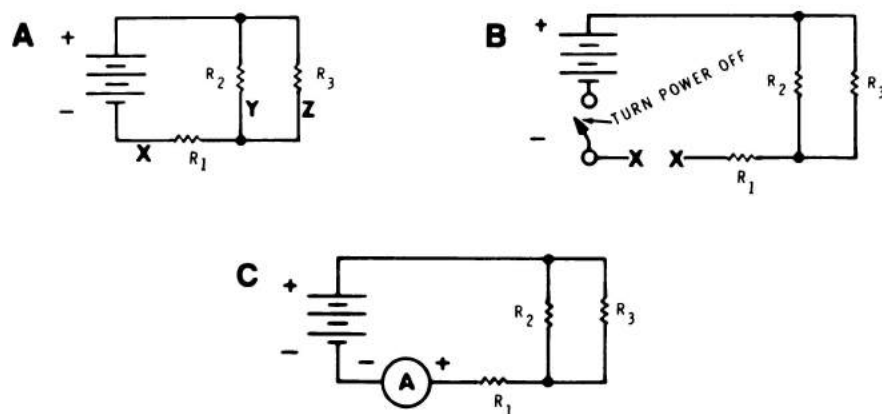
Figure 1-68
Using an ammeter to measure total current.

Figure 1-68A is a simple series-parallel circuit. To measure total current, the ammeter would be inserted at Point X. To measure current through R_2 or R_3, the ammeter would be inserted at Point Y or Z respectively. Suppose you want to measure the total current. You must first turn off the power to the circuit to eliminate any shock hazard. The circuit is then broken at point X. Next, observing the polarity of the ammeter, you would insert the ammeter in series. When power is applied to the circuit, the ammeter will indicate total current.

Each meter movement has a certain current rating. This is the current which will cause full-scale deflection. For example, an inexpensive meter movement may have a current rating of 1 milliampere. To obtain a usable reading, the current through the movement cannot be more than 1 mA. By itself, the movement has a single usable range of 0 to 1 mA.

Obviously, the meter would be much more useful if it could measure currents greater than 1 mA. Fortunately, there is an easy way to convert a sensitive meter movement to a less sensitive current meter. We do this by connecting a small value resistor in parallel with the meter movement. The resistor is called a **shunt**. Its purpose is to act as a low resistance path around the movement so that most of the current will flow through the shunt and only a small current will flow through the movement.
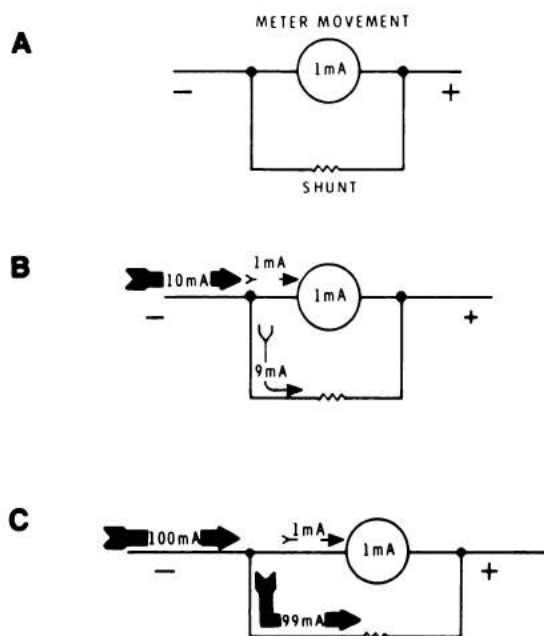
Figure 1-69
Increasing the current range of the ammeter.

Figure 1-69A shows a 1 mA meter movement connected across a low resistance shunt to form a higher range ammeter. The range will depend on how much current flows through the shunt. In Figure 1-69B, the

current applied to the ammeter is 10 mA. However, only 1 mA of this flows through the meter movement. The other 9 mA flows through the shunt. Thus, to convert the 1 mA movement to a 0-10 mA meter, the shunt must be chosen so that 9/10 of the applied current flows through the shunt. Once this is done, the full-scale position on the scale indicates 10 mA since this is the amount of current that must be applied before full-scale deflection is reached.

If the value of the shunt is made smaller, the meter can indicate even higher values of current. Figure 1-69C shows the requirements necessary to measure 100 mA. Here 99 mA or 99% of the applied current must go through the shunt. Thus, the resistance of the shunt must be much smaller than the resistance of the meter movement.

Voltmeter

The basic meter movement can be used to measure voltage as well as current. In fact, every meter movement has a certain voltage rating as well as current rating. This is the voltage which will cause full-scale deflection. Of course, the voltage rating is determined by the current rating and the meter resistance. For example, a 50 μA meter movement which has a resistance of 2,000 Ω, deflects full scale when connected across a voltage of:

E = I × R

E = 50 μA × 2,000 Ω

E = 0.1 volt

That is, the meter movement alone could be used to measure voltages up to 0.1 volt. Thus, the meter scale can be calibrated from 0-0.1 volt. However, if the meter movement is connected across a much higher voltage such as 10 volts, it may be damaged. Obviously, to be practical, we must extend the voltage range of the basic meter movement.

We have seen that a 50 μA, 2000 Ω meter movement can withstand a voltage of 0.1 volt without exceeding full scale.
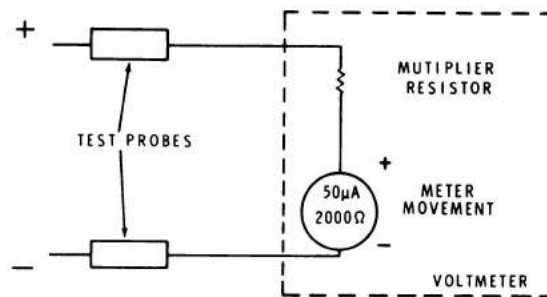
Figure 1-70
A voltmeter is formed by connecting multiplier
resistor in series with a microammeter.

To extend the range, we must insure that the voltage across the meter does not exceed 0.1 volt when the meter movement is connected across higher voltage. We do this by connecting a resistor in series with the meter movement as shown in Figure 1-70. The resistor is called a **multiplier** because it multiplies the range of the meter movement.

The purpose of the multiplier resistor is to limit the current that flows through the meter movement. For example, in the voltmeter shown in Figure 1-70, the current through the meter movement must be limited to 50 μA. Another way to look at it is that the multiplier must drop all the voltage applied to the voltmeter except the 0.1 volt allowed across the meter movement. For example, if the range is to be extended to 10 volts, then the multiplier must drop 10 −0.1 =9.9 volts. This will then mean that the current within the voltmeter will be 50 μA when 10 volts is applied across the test probes. Full-scale deflection will occur when 10 volts is applied.

That's how the voltmeter works, now we'll discuss how to use it. The test probes of the voltmeter shown in Figure 1-70 are marked with the symbols + and −. This means that we must also observe polarity when using the voltmeter, just as we did with the ammeter. The important difference between the ammeter and the voltmeter is in the method of connection.
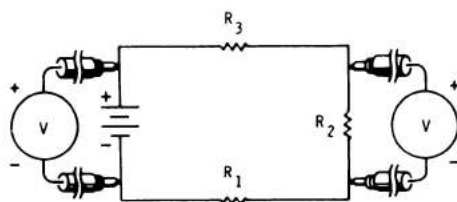
Figure 1-71
Using a voltmeter.

The voltmeter is connected **across** or in **parallel** with the voltage to be measured. This greatly simplifies measurement. We do not need to break any connections in the circuit or disturb it in any way. We need merely to exercise care when connecting the meter across the source of voltage to be measured.

Figure 1-71 shows a simple series circuit. We can measure the voltage of the battery by connecting the voltmeter across the battery, being careful to observe the polarity of the battery. The voltage drop across any resistor can also be measured by placing the voltmeter in parallel with or across the resistor. Figure 1-71 shows the meter connected to measure the voltage drop across R_2. Note that the polarity must be observed when connecting the voltmeter.

Usually. you will be using a multi-range voltmeter. This is a voltmeter which has several values of multiplier resistors and a switch with which to change the multiplier resistors and thus the full-scale deflection of the meter. When measuring an unknown voltage, it is always best to set the voltmeter on its highest range. This will avoid damaging the voltmeter. Then you can turn the range down until you get a suitable reading on the meter. Be careful not to switch the range down too far or you may damage the meter movement.

Ohmmeter

The basic meter movement can also be used to measure resistance. The resulting circuit is called an ohmmeter. In its most basic form, the ohmmeter is nothing more than a meter movement, a battery, and a series resistance.
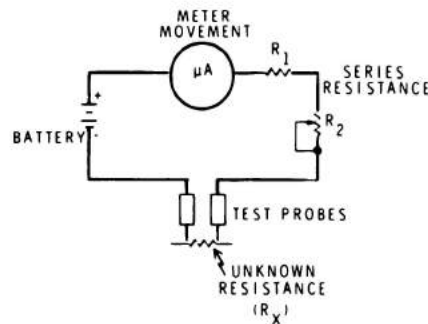
Figure 1-72
The basic ohmmeter.

Figure 1-72 shows the basic circuit of the ohmmeter. An ohmmeter forces a current to flow through an unknown resistance, then measures the current. For a given voltage, the current will be determined by the unknown resistance. That is, the amount of current measured by the meter is an indication of the unknown resistance. Thus, the scale of the meter movement can be marked off in ohms.

The purpose of the battery is to force current through the unknown resistance. The meter movement is used to measure the resulting current. The test probes have long leads and they simplify the job of connecting the ohmmeter to the unknown resistor (R_x). Fixed resistor R_1 limits the current through the meter to a safe level. Variable resistor R_2 is called the ZERO OHMS adjustment. Its purpose is to compensate for battery aging.
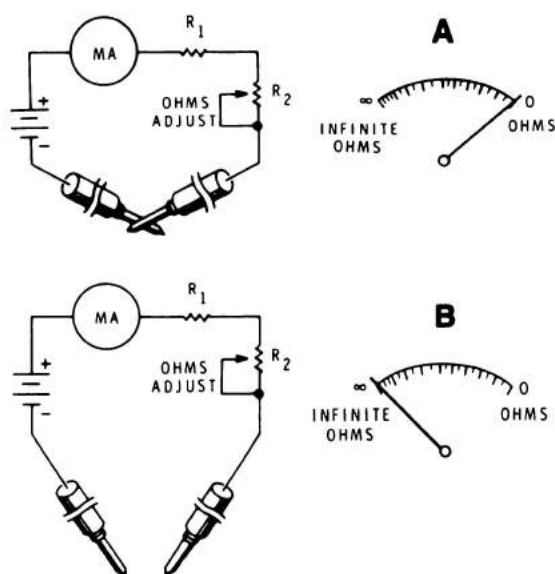
Figure 1-73
You must "zero" the ohmmeter before using it.

When using an ohmmeter, all voltage sources must be disconnected from the circuit under test. **Never use an ohmmeter on a circuit which has voltage applied to it**. When using the ohmmeter, it is first necessary to calibrate or "zero" it. This is shown in Figure 1-73A. When you short the leads of the ohmmeter together, the resistance value being measured is zero ohms. The meter will deflect to full scale. This is the zero indication of the meter. At times, you will have to adjust the OHMS ADJUST potentiometer (R_2) until the meter reads full scale. This is known as calibrating the meter. When the probes are apart and an open circuit exists, there will be no deflection of the meter, as no current exists in the circuit. This point on the scale is calibrated as infinite ohms. This is shown in Figure 1-73B. When measuring other values of resistance, the meter will deflect to points in between zero and infinite. The meter scale is calibrated to indicate the value of resistance being measured. The ohmmeter range may be changed by increasing the battery voltage, and the value of the series current limiting resistor R_1.

Multimeters

The multimeter or volt-ohm-milliammeter (VOM) is nothing more than the three basic meters — voltmeter, ammeter, and ohmmeter — combined into a single case. Cost and convenience are the primary reasons for this. Using a single movement and a single case costs less than three separate meters. It is also much simpler and easier to carry a single instrument than three separate ones.

A popular multimeter is shown in Figure 1-74. This particular instrument has a single switch, which is a combination range and function switch. This instrument is capable of four functions — DC voltmeter, AC voltmeter (which is discussed in the next unit), DC ammeter, and ohmmeter. The function switch selects the appropriate circuitry inside the meter and also the desired range by changing the multiplier, shunt, etc. All the functions work just like the meters previously described.

The common (−) and positive (+) jacks are normally used for the meter probes; however, there are a number of other jacks present which can be used to extend the range of the instrument beyond those normally selected by the range switch.

1 KV DC SOCKET - This socket provides the connection for dc voltages between 500 Vdc and 1000 Vdc with the RANGE switch in the 1 KV position.

1 KV AC SOCKET - This socket provides the connection for ac voltages between 500 Vac and 1000 Vac with the RANGE switch in the 1 KV position.

+/- SWITCH - This two-position switch is used to reverse the meter connections to the circuit being tested, thereby eliminating the necessity of changing the test lead connections when measuring dc.

RANGE SWITCH - This multiple-position switch may be turned in either direction to select the required range for a voltage, current, or resistance measurement.

Ω CONTROL - This control is used to adjust the pointer for a zero reading on the OHMS scale.

5 KV AC SOCKET - This socket provides the connection for ac voltages between 1000 Vac and 5000 Vac with the RANGE switch in the 5 KV position.

5 KV DC SOCKET - This socket provides the connection for dc voltages between 1000 Vdc and 5000 Vdc with the RANGE switch in the 5 KV position.

OUTPUT SOCKET - This socket provides the connection for pure ac measurements.

10 ADC SOCKET - This socket provides the connection for dc current measurements between 0.5 amperes (500 mA) and 10 amperes with the RANGE switch in the 10 A position.

- COM SOCKET - This socket provides the common or ground connection for the various inputs to the Meter.

+ SOCKET - This socket provides the input connection for all the functions and ranges, except the 1 kV, 5 kV, and 10 A ranges.

Figure 1-74

It is not always easy to tell an electrical meter (A) from the electronic type (B). The digital voltmeter (C) is easily recognized by its numerical readout.

All of the meters discussed up to this point are classified as electrical meters. They consist of a meter movement, precision resistors, and a battery for the ohmmeter. However, there is another family of meters called electronic meters. These contain electronic circuits which can amplify small voltages and currents. These meters use vacuum tubes, transistors or integrated circuits to amplify these voltages and currents. They may even use a digital (numerical) readout rather than a meter movement. This is why electronic meters are much more accurate than electrical meters.
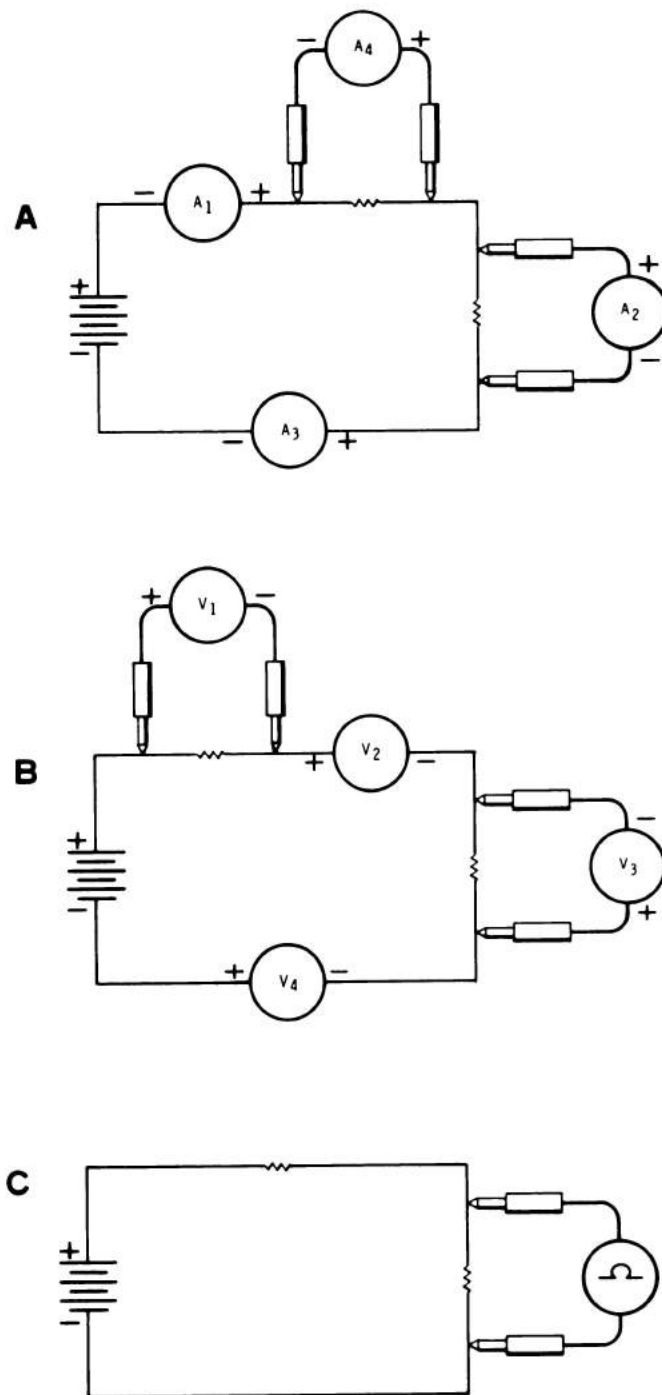
Figure 1-75
Circuits for self-review questions.

Self-Review Questions

48. An ammeter is used to measure _____ .

49. Which ammeter is connected properly in the circuit of Figure 1-75A? _____

50. A voltmeter is used to measure _____ .

51. Which voltmeter is connected properly in the circuit of Figure 1-75B? _____

52. An ohmmeter is used to measure _____ .

53. What is wrong with the way the ohmmeter is being used in Figure 1-75C? _____
     _____
     _____

54. What is a multimeter? _____
     _____
     _____

Self-Review Answers

48. An ammeter is used to measure **current**.

49. Ammeter A_3 is connected correctly.

50. A voltmeter is used to measure **voltage**.

51. Voltmeter V_1 is connected correctly.

52. An ohmmeter is used to measure **resistance**.

53. An ohmmeter must never be connected to a circuit which has voltage applied to it.

54. A multimeter is the three basic meters — voltmeter, ammeter, and ohmmeter — combined into a single case.

APPENDIX A

Scientific Notation

In electronics, it is common to deal with both very large and very small numbers. An example of a very large number is the speed at which electricity travels. It travels at the speed of light, which is approximately 1,000,000,000 feet per second or about 300,000,000 meters per second. As for very small numbers, consider the size and weight of an electron. It is believed that the electron has a diameter of approximately 0.000 000 000 0022 inch and a weight of about 0.000 000 000 000 000 000 000 000 0009 gram. Sometimes, we perform arithmetic with numbers such as these. To simplify such arithmetic, a shorthand method has been developed to express numbers. This shorthand method is called scientific notation.

The following programmed instruction sequence will serve as an introduction to scientific notation. As you read each numbered frame, cover up the frame below it until you have answered the question. The correct answer will be at the beginning of the next frame. Be sure to write in each answer; the act of writing is part of the memory training process.

1. As mentioned above, scientific notation is a shorthand method of expressing numbers. While any number can be expressed in scientific notation, this technique is particularly helpful in expressing very large and very ＿＿＿＿＿＿＿ numbers.

2. (small) Scientific notation is based on a concept called powers of ten. Thus, in order to understand scientific notation we should first learn what is meant by powers of ＿＿＿＿＿＿＿.

3. (ten) In mathematics, a number is raised to a power by multiplying the number times itself one or more times. Thus, we raise 5 to the second power by multiplying 5 times itself. That is, 5 to the second power is $5 \times 5 =$＿＿＿＿＿＿＿.

4. (25) Also, 5 to the third power is the same as saying $5 \times 5 \times 5 =$＿＿＿＿＿＿＿.

5. (125) Thus, 5 can be raised to any power simply by multiplying it times itself the required number of times. For example, 5 × 5 × 5 × 5 = 625. Consequently, 5 raised to the _____ power is equal to 625.

6. (fourth) The above examples use powers of five. However, any number can be raised to a power by the technique of multiplying it times itself the required number of times. Thus, the powers of two would look like this:

2 to the second power equals 2 × 2 = 4

2 to the third power equals 2 × 2 × 2 = 8

2 to the fourth power equals 2 × 2 × 2 × 2 = 16

2 to the fifth power equals 2 × 2 × 2 × 2 × 2 = 32

2 to the sixth power equals 2 × 2 × 2 × 2 × 2 × 2 = _____

7. (64) In mathematics, the number which is raised to a power is called the *base*. If 5 is raised to the third power, 5 is considered the _____.

8. (base) The power to which the number is raised is called the *exponent*. If 5 is raised to the third power, then the exponent is 3. In the same way, if 2 is raised to the sixth power, then 2 is the base while 6 is the _____.

9. (exponent) There is a shorthand method for writing "2 raised to the sixth power." It is:

$$2^6$$

Notice that the exponent is written as a small number at the top right of the base. Remember *this* number is the base while *this* number is the exponent.

$$2^6$$

Therefore in the example 3^4, 3 is the _____ while 4 is the _____.

10. (base, exponent) The number 3^4 is read "3 raised to the fourth power." It is equal to:

$$3 \times 3 \times 3 \times 3 = 81.$$

The number 4^6 is read _____.

11. (4 raised to the sixth power) Scientific notation uses powers of ten. Several powers of ten are listed below:

$$10^2 = 10 \times 10 = 100$$
$$10^3 = 10 \times 10 \times 10 = 1000$$
$$10^4 = 10 \times 10 \times 10 \times 10 = 10,000$$
$$10^5 = 10 \times 10 \times 10 \times 10 \times 10 = 100,000$$
$$10^6 = 10 \times 10 \times 10 \times 10 \times 10 \times 10 = \text{_____}.$$

12. (1,000,000) Multiplication by 10 is extremely easy since all we have to do is add one zero for each multiplication. Another way to look at it is that multiplication by ten is the same as moving the decimal point one place to the right. Thus, we can find the equivalent of 10^2 by multiplying $10 \times 10 = 100$; or, simply by adding a 0 after 10 to form 100; or by moving the decimal point one place to the right to form $10.0. = 100$. In any event, 10^2 is equal to

_____.

13. (100) There is a simple procedure for converting a number expressed as a power of ten to its equivalent number. We simply write down a 1 and after it write the number of zeros indicated by the exponent. For example, 10^6 is equal to 1 with 6 zeros after it. In the same way 10^{11} is equal to 1 with _____ zeros after it.

14. (11) This illustrates one of the advantages of power of ten. It is easier to write and remember 10^{21} than its equivalent number: 1,000,000,000,000,000,000,000. Try it yourself and see if it isn't easier to write 10^{35} than to write its equivalent number of:

_____.

15. (100,000,000,000,000,000,000,000,000,000,000,000). In the above examples, we converted a number expressed in powers of ten to its equivalent number. Now let's see how we convert in the opposite direction. Remember the number must be expressed using 10 as the base with the appropriate exponent. The exponent is determined simply by counting the zeros which fall on the right side of the 1. Thus, 1,000,000 becomes 10^6 because there are 6 zeros in the number. In the same way, 10,000,000,000 is expressed as

_____.

16. (10^{10}) To be sure you have the right idea, study each of the groups below. Which group contains an error? _____.

Group A	Group B	Group C
$10^6 = 1,000,000$	$1000 = 10^3$	$10^7 \ = 10,000,000$
$10^2 = 100$	$10,000 = 10^4$	$10^9 \ = 1,000,000,000$
$10^9 = 1,000,000,000$	$100 = 10^2$	$10^{11} = 10,000,000,000$

17. (Group C) There are two special cases of powers of ten which require some additional explanation. The first is 10^1. Here the exponent of 10 is 1. If we follow the procedure developed in Frame 13 we find that $10^1 = 10$. That is, we put down a 1 and add the number of zeros indicated by the exponent. Thus $10^1 =$

_____.

18. (10) The other special case is 10^0. Here the exponent is 0. Once again we follow the procedure outlined in Frame 13. Here again we write down a 1 and add the number of zeros indicated by the exponent. However, since the exponent is 0, we add no zeros. Thus, the equivalent number of 10^0 is 1. That is $10^0 =$ _____.

19. (1) Any base number with an exponent of 1 is equal to the base number. Any base number with an exponent of 0 is equal to 1. Thus, $X^1 =$ _____ and $X^0 =$ _____.

20. (X, 1) In the examples given above, the exponents have been positive numbers. For simplicity the plus sign has been omitted. Therefore, 10^2 is the same as 10^{+2}. Also, 10^6 is the same as _____.

21. (10^{+6}) Positive exponents represent numbers larger than 1. Thus, numbers such as 10^2, 10^7, and 10^{15} are greater than 1 and require _____ exponents.

22. (positive) Numbers smaller than 1 are indicated by negative exponents. Thus, numbers like 0.01, 0.0001, and 0.00001 are expressed as negative powers of ten because these numbers are less than _____.

23. (1) Some of the negative powers of ten are listed below:
$10^{-1} = 0.1$
$10^{-2} = 0.01$
$10^{-3} = 0.001$
$10^{-4} = 0.0001$
$10^{-5} = $ _____.

24. (0.00001) A brief study of this list will show that this is simply a continuation of the list shown earlier in frame 11. If the two lists are combined in a descending order, the result will look like this:
$10^6 = 1,000,000.$
$10^5 = \quad 100,000.$
$10^4 = \quad\quad 10,000.$
$10^3 = \quad\quad\quad 1,000.$
$10^2 = \quad\quad\quad\quad 100.$
$10^1 = \quad\quad\quad\quad\quad 10.$
$10^0 = \quad\quad\quad\quad\quad\quad 1.$
$10^{-1} = \quad\quad\quad\quad\quad 0.1$
$10^{-2} = \quad\quad\quad\quad\quad 0.01$
$10^{-3} = \quad\quad\quad\quad\quad 0.001$
$10^{-4} = \quad\quad\quad\quad\quad 0.0001$
$10^{-5} = $

25. (0.00001) We can think of the negative exponent as an indication of how far the decimal point should be moved to the left to obtain the equivalent number. Thus, the procedure for converting a negative power of ten to its equivalent number can be developed. The procedure is to write down the number 1 and move the decimal point to the left the number of places indicated by the negative exponent. For example, 10^{-4} becomes:

0.0001. or 0.0001

Notice that the -4 exponent indicates that the decimal point should be moved _____ places to the _____.

26. (4, left) Up to now we have used powers of ten to express only those numbers which are exact multiples of ten such as 100, 1000, 10,000, etc. Obviously, if these were the only numbers which could be expressed as powers of ten, this method of writing numbers would be of little use. Actually, any _____ can be expressed in powers-of-ten notation.

27. (number) The technique by which this is done can be shown by an example. If 1,000,000 can be represented by 10^6, then 2,000,000 can be represented by 2×10^6. That is, we express the quantity as a number multiplied by the appropriate power of ten. As another example, $2,500,000 = 2.5 \times 10^6$. Also, $3,000,000 =$

_____.

28. (3×10^6) In the same way, we can write 5,000 as 5×10^3. Some other examples are:

$$200 = 2 \times 10^2$$
$$1500 = 15 \times 10^2$$
$$22,000 = 22 \times 10^3$$
$$120,000 = 12 \times 10^4$$
$$1,700,000 = 17 \times 10^5$$
$$9,000,000 = \underline{\qquad}$$

29. (9×10^6) By the same token, we can convert in the opposite direction. Thus, 2×10^5 becomes $2 \times 100,000$ or $200,000$. Also, $2.2 \times 10^3 = 2.2 \times 1000 = 2,200$. And, $66 \times 10^4 = $ _____.

30. $(660,000)$ You may have noticed that when we use powers of ten there are several different ways to write a number. For example, $25,000$ can be written as 25×10^3 because 25×1000 equals $25,000$. However; it can also be written as 2.5×10^4 because $2.5 \times 10,000$ equals $25,000$. It can even be written as 250×10^2 since $250 \times 100 = 25,000$. In the same way, 4.7×10^4, 47×10^3, and 470×10^2 are three different ways of writing the number _____.

31. $(47,000)$ Numbers smaller than one are expressed as negative powers of ten in much the same way. Thus, $.0039$ can be expressed as 3.9×10^{-3}, 39×10^{-4}, or $.39 \times 10^{-2}$. Also, 6.8×10^{-5}, 68×10^{-6}, and $.68 \times 10^{-4}$ are three different ways of expressing the number _____.

32. $(.000068)$ As you can see there are several different ways in which a number can be written as a power of ten. Scientific notation is a way of using powers of ten so that all numbers can be expressed in a uniform way. To see exactly what scientific notation is, consider the following examples of numbers written in scientific notation:

$$6.25 \times 10^{18}$$
$$3.7 \ \times 10^{6}$$
$$4.0 \ \times 10^{2}$$
$$6.8 \ \times 10^{-4}$$
$$3.9 \ \times 10^{-6}$$
$$2.2 \ \times 10^{-12}$$

Notice that the numbers range from a very large number to an extremely small number. And yet, all these numbers are written in a uniform way. This method of writing numbers is called scientific _____.

33. (notation) The rules for writing a number in scientific notation are quite simple. First, the decimal point is always placed after the first digit on the left which is not a zero. Therefore, the final number will appear in this form: 6.25, 7.3, 9.65, 8.31, 2.0 and so forth. It must never appear in a form such as: .625, 73, 96.5, .831 or 20. Thus, there is always one and only one digit on the _____ side of the decimal point.

34. (left) The second rule involves the sign of the exponent. If the original number is greater than 1, the exponent must be positive. If the number is less than 1, the exponent must be negative. Thus, 67,000 requires a positive exponent but 0.00327 requires a _____ exponent.

35. (negative) Finally, the magnitude of the exponent is determined by the number of places that the decimal point is moved. For example, 39,000.0 is expressed as 3.9×10^4 because the decimal point must be moved 4 places in order to have only one digit to the left of it. Using this rule, 6,700,000,000 is expressed as $6.7 \times$ _____.

36. (10^9) The number 0.00327 is expressed as 3.27×10^{-3}. Here the decimal point is moved 3 places in order to have one digit which is not zero to the left of the decimal. Likewise 0.00027 is expressed as $2.7 \times$ _____.

37. (10^{-4}) To be sure you have the idea look at the groups of numbers below. Which of the following groups contains a number that is not expressed properly in scientific notation? _____

Group A	Group B	Group C
6.25×10^{18}	1.11×10^{11}	6.9×10^{10}
3.75×10^{-9}	-3.1×10^{2}	3.4×10^{7}
4.20×10^{1}	-3.1×10^{-2}	39.5×10^{2}
7.93×10^{0}	2.00×10^{2}	6.0×10^{4}

38. (Group C) The number 39.5×10^2 is not written in scientific notation because there are two digits on the left side of the decimal point. The minus signs in Group B may have confused you. Although, it has not been mentioned, negative numbers can also be expressed in scientific notation. Thus, a number like $-6,200,000$ becomes -6.2×10^6. All the rules previously stated hold true except that now a _____ sign is placed before the number.

39. (minus) Small negative numbers are handled in the same way. Thus -0.0092 becomes -9.2×10^{-3}. The minus sign before the number indicates that this is a negative number. The minus sign before the exponent indicates that this number is less than _____.

40. (1) Listed below are numbers which are converted to scientific notation. Which one of these groups contains an error? _____

Group A	Group B	Group C
$2,200 = 2.2 \times 10^3$	$119,000 = 1.19 \times 10^5$	$119 = 1.19 \times 10^2$
$32,000 = 3.2 \times 10^4$	$1,633,000 = 1.633 \times 10^6$	$93 = 9.3 \times 10^1$
$963,000 = 9.63 \times 10^5$	$937,000 = 9.37 \times 10^4$	$7.7 = 7.7 \times 10^0$
$660 = 6.6 \times 10^2$	$6,800 = 6.8 \times 10^3$	$131.2 = 1.312 \times 10^2$

41. (Group B) 937,000 converts to 9.37×10^5 and not to 9.37×10^4. Which of the groups below contains an error? _____.

Group A	Group B	Group C
$0.00037 = 3.7 \times 10^{-4}$	$0.44 = 4.4 \times 10^{-1}$	$.37 = 3.7 \times 10^{-1}$
$0.312 = 3.12 \times 10^{-1}$	$0.0002 = 2.0 \times 10^{-4}$	$.0098 = 9.8 \times 10^{-3}$
$0.068 = 6.8 \times 10^{-2}$	$0.0798 = 7.98 \times 10^{-2}$	$.00001 = 1.0 \times 10^{-5}$
$0.0092 = 9.2 \times 10^3$	$0.644 = 6.44 \times 10^{-1}$	$0.0075 = 7.5 \times 10^{-3}$

42. (Group A) The final number in group A requires a negative exponent. Which of the groups below contains an error?_____.

Group A	Group B	Group C
$3{,}700{,}000 = 3.7 \times 10^6$	$9440 = 9.44 \times 10^3$	$20 = 2.0 \times 10^1$
$-5{,}500 = -5.5 \times 10^3$	$-110 = -1.1 \times 10^2$	$0.02 = 2.0 \times 10^{-2}$
$0.058 = 5.8 \times 10^{-2}$	$0.0062 = 6.2 \times 10^{-4}$	$-200{,}000 = -2.0 \times 10^5$
$-0.0034 = -3.4 \times 10^{-3}$	$-0.0123 = -1.23 \times 10^{-2}$	$-0.000200 = -2.0 \times 10^{-4}$

43. (Group B) 0.0062 is equal to 6.2×10^{-3}. Match the following:

1.	16	a.	1.6×10^{-3}
2.	.0016	b.	1.6×10^4
3.	160,000	c.	1.6×10^0
4.	1.6	d.	1.6×10^1
5.	.016	e.	1.6×10^{-2}
6.	16,000	f.	1.6×10^5

44. (1-d, 2-a, 3-f, 4-c, 5-e, 6-b). Another concept that goes hand in hand with powers of ten and scientific notation is metric prefixes. These are prefixes such as *mega* and *kilo* which when placed before a word change the meaning of the word. For example, the prefix *kilo* means *thousand*. When kilo and meter are combined the word kilometer is formed. This word means 1000 meters. In the same way, the word kilogram means _____ grams.

45. (1,000) Since kilo means 1,000 we can think of it as multiplying any quantity times 1000 or 10^3. Thus, kilo means 10^3. Another popular metric prefix is *mega*. *Mega* means *million*. Thus a megaton is one million tons or 10^6 tons. In the same way one million volts is referred to as a _____ volt.

46. (mega) One thousand watts can be called a kilowatt. Also one million watts can be called a _____.

47. (megawatt) A kilowatt is equal to 10^3 watts while a megawatt is equal to _____ watts.

48. (10^6) Often it is convenient to convert from one prefix to another. For example, since a megaton is 10^6 tons and a kiloton is 10^3 tons, a megaton equals 1000 kilotons. And, since a megaton is one thousand times greater than a kiloton, the kiloton is equal to .001 megaton. Now, consider the quantity 100,000 tons. This is equal to 100 kilotons or _____ megatons.

49. (0.1) Kilo is often abbreviated k. Thus, 100 kilowatts may be expressed as 100 k watt. Mega is abbreviated M. Therefore 10 megawatts may be expressed as _____ watts.

50. (10M) The quantity 5 k volts is 5 kilovolts or 5000 volts. Also, 5 M volts is 5 megavolts or _____ volts.

51. (5,000,000) There are also prefixes which have values less than one. The most used are:

milli   which means *thousandth* (.001) or 10^{-3}, and
micro   which means *millionth* (.000 001) or 10^{-6}.

One thousandth of an ampere is called a milliampere. Also, one thousandth of a volt is called a _____.

52. (millivolt) If a second is divided into one million equal parts each part is called a microsecond. Also, the millionth part of a volt is called a _____.

53. (microvolt) One volt is equal to 1000 millivolts or 1,000,000 microvolts. Or, 1 volt equals 10^3 millivolts and 10^6 microvolts. Expressed another way, 1 millivolt equals .001 volt while 1 microvolt equals .000001 volt. Thus, 1 millivolt equals 10^{-3} volts while 1 microvolt equals _____ volt.

54. (10^{-6}) Powers of ten allow us to express a quantity using whichever metric prefix we prefer. For example, we can express 50 millivolts as 50×10^{-3} volts simply by replacing the prefix milli with its equivalent power of ten. In the same way 50 microvolts is equal to $50 \times$ _____ volts.

55. (10^{-6}) When writing abbreviation for the prefix milli the letter small *m* is used. A small *m* is used to distinguish it from mega which used a capital M. Obviously, the abbreviation for micro cannot also be *m*. To represent micro the Greek letter μ (pronounced mu) is used. Thus, 10 millivolts is abbreviated 10 m volts while 10 microvolts is abbreviated 10 μ volts. Remember, m means 10^{-3} while μ means _____.

56. (10^{-6}) Match the following:

1. M watt   a. 10^{-3} watts

2. k watt   b. 10^{-6} watts

3. m watt   c. 500×10^{-3} watts

4. μ watt   d. 10^6 watts

5. .5 watt   e. .5k watts

6. 500 watts   f. 10^3 watts

7. 500,000 watts   g. .5 M watts

8. .00005 watts   h. .05 k watts

9. 50 watts   i. 5 m watts

10. .005 watts   j. 50 μ watts

57. (1-d, 2-f, 3-a, 4-b, 5-c, 6-e, 7-g, 8-j, 9-h, 10-i)

Unit 2

ALTERNATING CURRENT

CONTENTS

INTRODUCTION

Alternating current (commonly identified as AC or ac) plays an important role in electronics. In fact, it is used more extensively than direct current (DC) because it has a wider range of practical applications. Alternating current is used so extensively that it is almost impossible to avoid using it, or at least being affected by it. For example, each time you plug in your electric shaver, toaster, or drill, you are using alternating current. When you turn on your radio or television set, the music or voice information that you hear is produced by alternating current. Even when you drive down a city street, you are still under its influence since the traffic signals that you obey and the street lights which illuminate your way all rely on alternating current for operation.

This unit will teach you what alternating current is and how it behaves in simple circuits.

The "Unit Objectives" listed next state exactly what you are expected to learn from this unit. Study this list now and refer to it often as you study the text.

UNIT OBJECTIVES

When you complete this unit, you should be able to:

1. Define AC and DC.

2. Identify a sine wave.

3. Determine the peak value, peak-to-peak value, and rms or effective value of a sine wave.

4. Find peak value when given the rms value.

5. Define hertz, kilohertz, and megahertz.

6. Find wavelength when given frequency.

7. Determine the phase angle between two sine waves.

8. State how a DC moving-coil meter can be used to measure alternating current.

9. Tell what an oscilloscope does.

10. List the measurements that an oscilloscope can make.

11. Measure amplitude, period, and frequency using an oscilloscope.

12. Define inductance, henry, and mutual inductance.

13. Solve problems involving the turns ratio of a transformer with regard to voltage, current, and impedance ratios.

14. State why impedance matching is important.

15. Name two sources of loss in a transformer.

16. Define inductive reactance and impedance.

17. Write the formula for inductive reactance.

18. Find inductive reactance when given frequency and inductance.

19. Find the total inductance of series and parallel connected inductors.

20. State the phase angle between voltage and current in an inductor.

21. Find the impedance of a series RL circuit.

22. Define capacitance, farad, and capacitive reactance.

23. State the three factors which determine capacitance.

24. Find the total capacitance of series and parallel connected capacitors.

25. Write the formula for capacitive reactance.

26. Find capacitive reactance when given frequency and capacitance.

27. State the phase angle between voltage and current in a capacitor.

28. Find the impedance of a series RC circuit, a series RLC circuit, and a parallel RLC circuit.

29. Define resonance.

30. Write the formula for resonance.

31. State the characteristics of series and parallel resonant circuits.

32. Define circuit Q and state its effect on the bandwidth of a resonant circuit.

33. List four types of filters and state the purpose of each one.

WHAT IS AC?

Up to this point, we have been primarily concerned with direct current or DC. In this unit, you will learn about alternating current or AC.

Alternating current (AC) is fundamentally different from direct current (DC), since DC flows in only one direction, and AC periodically changes its direction of flow. The AC supplied by your local power company changes direction sixty times per second.

While we can use a battery to supply DC, AC must be supplied by a generator. In this section, you will learn how AC is generated and what an AC waveform looks like.

Generating an Alternating Voltage

The most common method of generating an alternating voltage is by moving a conductor through a magnetic field. Figure 2-1 illustrates the principles involved. When the conductor is moved across the magnetic field, the free electrons in the conductor are forced to move. If the conductor is moved up, the free electrons are forced to the right end of the conductor. This causes an excess of electrons, or negative potential, at the right end of the conductor and a deficiency of electrons, or positive potential, at the left end. The result is a potential difference across the conductor.
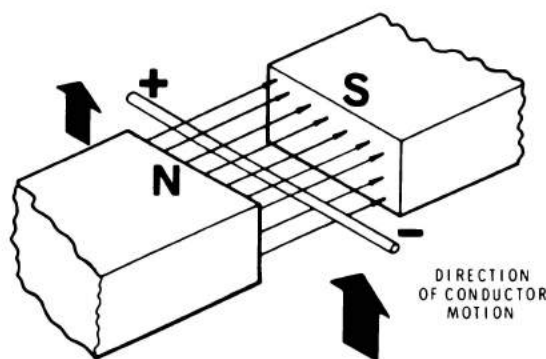
Figure 2-1
Electromagnetic induction.
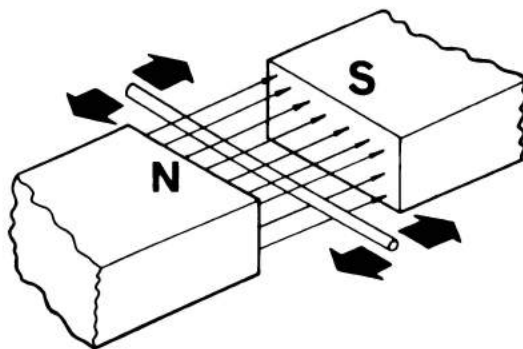
Figure 2-2
No voltage is induced when the conduc-
tor moves parallel to the flux lines.

The key to producing the potential difference is the relative motion
between the conductor and the magnetic field. If the motion stops, the
electrons return to their normal positions, and the potential difference
disappears. Furthermore, the motion must cause the conductor to **cut
across** the magnetic field. Figure 2-2 shows what happens when the
conductor motion is parallel to the magnetic flux lines. No voltage is
"induced" into the conductor because no flux lines are cut.

Another important point is that when the conductor motion is down
through the magnetic field, the polarity of the voltage is opposite to that
shown in Figure 2-1. By moving the conductor up and down in the
magnetic field, the induced voltage will alternately reverse polarity.

If we move the conductor at a faster rate or if the magnetic field is
stronger, a higher voltage will be induced in the conductor. The mag-
nitude of voltage induced in the conductor is directly related to the
number of flux lines which are cut by the conductor.

A simple AC generator is shown in Figure 2-3. Notice that we have
formed the conductor into a loop known as an armature. When the loop or
armature is rotated, one half moves up through the field near the south
pole, while the other half moves down through the field near the north
pole. You can see that the voltage induced into one half aids the other
half, just as if two batteries were connected in series.
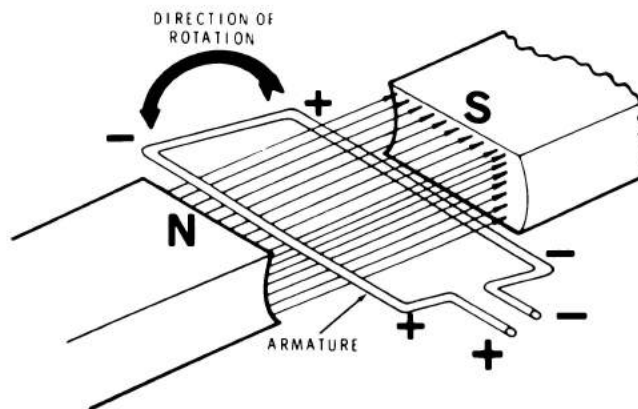
Figure 2-3
The AC generator.

Figure 2-4A shows the output voltage from this AC generator. To see how this voltage is produced, we must follow the armature through one complete revolution. In each case, we will consider the voltage at Point A with respect to Point B. Figure 2-4B shows the armature at 90° increments.
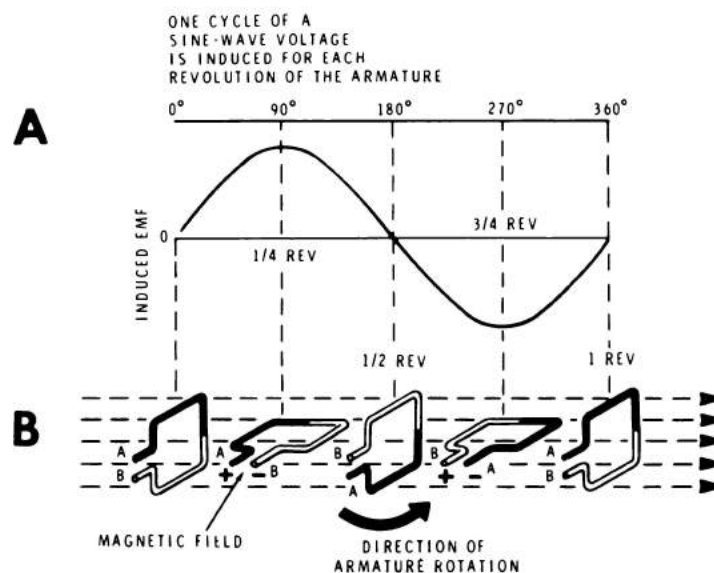
Figure 2-4
Generating a sine-wave.

At 0°, the sides of the armature are moving parallel to the lines of flux, thus, there is no induced voltage at this time. However, as the armature rotates, the loop begins cutting the lines of flux and a voltage is induced. Point A becomes positive with respect to Point B. The voltage begins to rise and it reaches its maximum value at 90° of rotation. The voltage is maximum at this point because the armature is cutting the flux at a right angle. Thus, it cuts the maximum number of lines at this point. Once past 90°, the voltage begins to decrease because fewer lines per second are being cut.

At 180°, the induced voltage is again zero because the armature is moving parallel to the lines. As the loop passes 180° and starts cutting the lines again, the voltage is once more induced. However, this time Point A becomes negative with respect to Point B; because side A is now moving up, just the opposite to its earlier movement.

The maximum negative voltage is produced at 270° when the armature is once again cutting the lines at right angles. As the armature heads back toward its starting point, the voltage begins to decrease back toward 0. At 360°, the armature is back where it started and the induced voltage is again zero.

The voltage shown in Figure 2-4A is called a **sine wave**. One cycle of the sine wave is produced for each revolution of the armature. If a load is connected across the armature between Points A and B, current will flow through the load. For the first half cycle, current will flow from Point B through the load to Point A. However, during the next half cycle, current will flow from Point A, through the load in the opposite direction to Point B. Thus, during each cycle, the current reverses direction — flowing in one direction half of the time and flowing in the opposite direction the other half of the time. This is called alternating current and is abbreviated AC.

The voltages supplied to our homes, offices, and factories are AC voltages. The armature of the generator at the plant which provides our power is rotating 60 times each second. Thus, the voltage supplied by these power stations goes through 60 cycles, like the one shown in Figure 2-4A, each second. Most of the appliances in your home require 115 volts at 60 cycles per second.

The AC generator is often called an alternator because it produces alternating current. The simple machine shown here would not produce useful power because the armature consists of only one turn of wire. In a practical alternator, hundreds of turns are wound into an armature which can produce considerable power.

Self-Review Questions

1. Define DC. _____
    _____

2. Define AC. _____
    _____

3. The waveform shown in Figure 2-5 is a _____ .
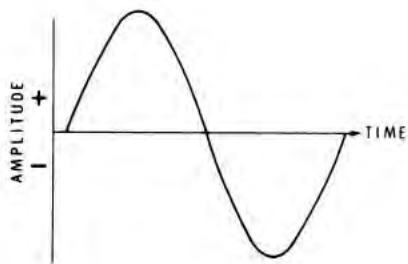
Figure 2-5
What is this waveform?

Self-Review Answers

1. DC stands for direct current. Direct current flows in only one direction.

2. AC stands for alternating current. Alternating current periodically changes its direction of flow.

3. The waveform shown in Figure 2-5 is a **sine wave**.

WAVEFORM MEASUREMENTS

An alternating sine wave of voltage or current is constantly changing in value. In order to compare one sine wave to another, it is necessary to define specific values.

Voltage and Current Values

There are several different methods of specifying the amplitude of a sinewave. Figure 2-6 shows the three most common.
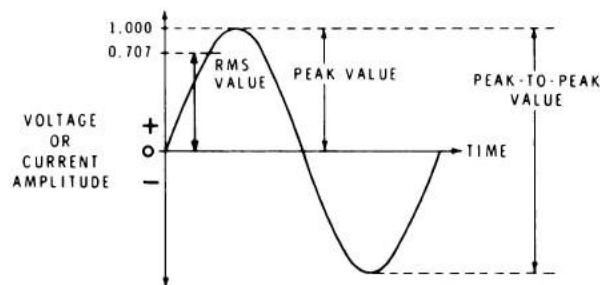
Figure 2-6
AC values for a sine wave of voltage or current.

One measure of a sine wave's voltage or current amplitude is the **peak** value. Figure 2-6 shows that the peak value of a sine wave is measured from the zero line to the maximum positive (or negative) point. As an example, the sine wave of Figure 2-7 has a positive peak of 10 volts. Its negative peak is also 10 volts.
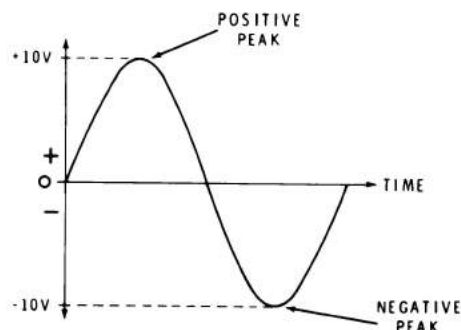
Figure 2-7
What is the peak-to-peak value of this sine wave?

Another measure of a sine wave's amplitude is the **peak-to-peak** value. It is measured from the positive to the negative peak. Therefore, the waveform of Figure 2-7 has a peak-to-peak value of 20 volts. Since a sine wave is symmetrical about the zero line, the peak-to-peak value is equal to twice the peak value. The following formulas show the relationship between the peak and peak-to-peak values of a sine wave.

$$\text{peak value} = \frac{\text{peak-to-peak value}}{2}$$

peak-to-peak value = peak value × 2

The most commonly used measure of a sine wave is the **rms** value. The rms value of a sine wave is determined by using the formula below:

rms value = peak value × 0.707

The factor 0.707 is determined mathematically by the root-mean-square or rms method. The rms value is also referred to as the **effective** value because it corresponds to the same amount of direct current or voltage in heating power. As an example, 120 volts rms gives exactly the same amount of heating power as 120 volts DC.

It is often necessary to convert rms values to peak or peak-to-peak values. When converting rms values to peak values, use the formula:

peak value = rms value × 1.414

As an example, the 120 volts AC supplied by your local power company is an rms value. The peak value of this voltage is:

peak value = 120 volts × 1.414

peak value = 169.68 volts or approximately 170 volts

The peak-to-peak value would then be:

peak-to-peak value = peak value × 2

peak-to-peak value = 170 volts × 2

peak-to-peak value = 340 volts

These formulas apply to both current and voltage values.
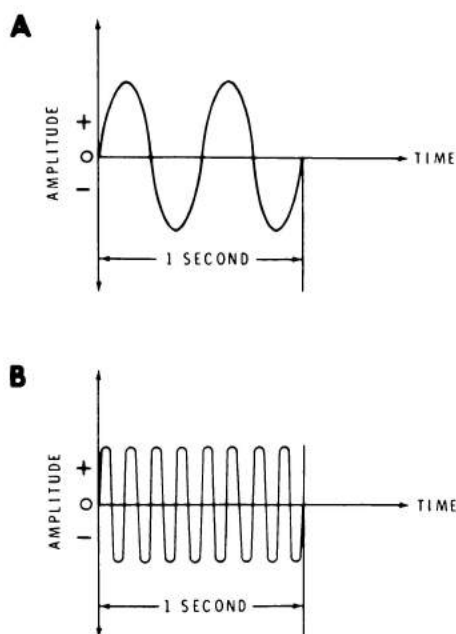
Figure 2-8
What is the frequency of these sine waves?

Frequency

The **frequency** of a sine wave is determined by the number of cycles per second. One cycle is one full rotation of a generator's armature, or a full 360° of a sine wave.

The sine wave shown in Figure 2-8A completes two cycles in one second. Its frequency is 2 hertz. **Hertz** is the unit of frequency and stands for cycles per second. The abbreviation for hertz is Hz. What is the frequency of the sine wave shown in Figure 2-8B? Eight cycles are completed in one second. Therefore, the frequency is 8 Hz.

To simplify writing frequencies, you should use metric prefixes:

1,000 Hz = 1 kilohertz or 1 kHz (1×10^3)

1,000,000 Hz = 1 megahertz or 1 MHz (1×10^6)

1,000,000,000 Hz = 1 gigahertz or 1 GHz (1×10^9)

For example, TV channel 2 has a frequency of 55.25 MHz. This can also be expressed as 55,250 kilohertz or 55,250,000 hertz.

Wavelength

The wavelength is the distance a wave travels in one complete cycle. When a radio wave leaves an antenna, it travels at the speed of light. The distance a wave travels is dependent on its velocity and frequency:

$$\lambda \quad = \quad \frac{\text{velocity}}{\text{frequency}}$$

Where λ (Lambda) is the symbol for wavelength.

For radio waves, the velocity is the speed of light which is 300,000,000 meters per second or 3×10^8 m/sec. Therefore:

$$\lambda \text{ (meters)} \quad = \quad \frac{3 \times 10^8 \text{ m/sec}}{\text{frequency (Hz)}}$$

Let's work some example wavelength problems. What is the wavelength for a frequency of 3.75 MHz?

$$\lambda \quad = \quad \frac{3 \times 10^8}{\text{frequency (Hz)}}$$

$$\lambda = \frac{3 \times 10^8}{3.75 \times 10^6 \text{ Hz}}$$

When dividing powers of ten, subtract the denominator's exponent from the numerator's exponent, then divide the base numbers. In this example:

$$\lambda = \frac{3 \times 10^8}{3.75 \times 10^6} = \frac{3}{3.75} \times 10^2$$

$$\lambda = 0.8 \times 10^2$$

$$\lambda = 80 \text{ meters}$$

What is the wavelength for a frequency of 28.5 MHz?

$$\lambda = \frac{3 \times 10^8}{28.5 \times 10^6} = \frac{3}{28.5} \times 10^2$$

$$\lambda = 0.1053 \times 10^2$$

$$\lambda = 10.53 \text{ meters}$$

Phase Angle

The three ways AC sine waves can differ are frequency, amplitude, and phase. Phase is the number of electrical degrees that one wave leads or lags another wave. Figure 2-9 shows four sine waves. Their amplitude and frequency are the same. The only difference is the phase variation between them. Using the waveform in A of Figure 2-9 as our reference, B is exactly in phase. The waveform in C crosses the zero line 90° later than our reference waveform in A, so the waveform in C is said to lag the waveform in A by 90°. The waveform in D crosses the zero line 180° after our reference, so it lags the waveform in A by 180°.

What is the phase difference between waveforms C and D? D lags C by 90° because waveform D crosses the zero line 90° after C.
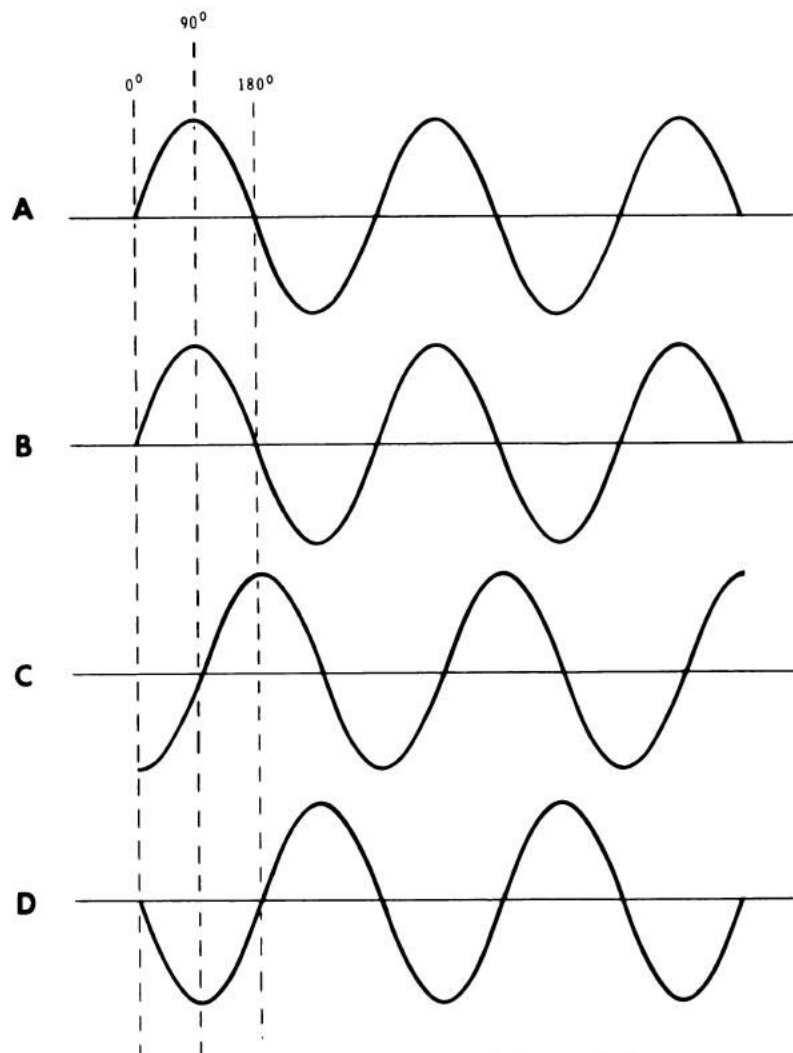
Figure 2-9
Phase relationships between AC sine waves.

Self-Review Questions
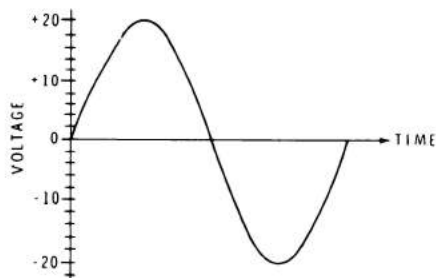
Figure 2-10
What is the peak voltage of this sine wave?

4. What is the peak voltage of the sine wave shown in Figure 2-10?

5. What is the rms voltage of the sine wave shown in Figure 2-10?

6. What is the peak-to-peak voltage of the sine wave shown in Figure 2-10? _____

7. A 240-volt rms sine wave has a peak voltage of _____.

8. 1 cycle per second is equal to _____ hertz. 1,000 Hz is equal to _____ kHz. 1,000 kHz is equal to _____ MHz.

9. What is the wavelength of a frequency of 21.42 MHz? _____.

10. What is the wavelength of a frequency of 1,875 kHz? _____.

11. In Figure 2-11, the sine wave in A (leads, lags) the sine wave in B by (90, 180) degrees.
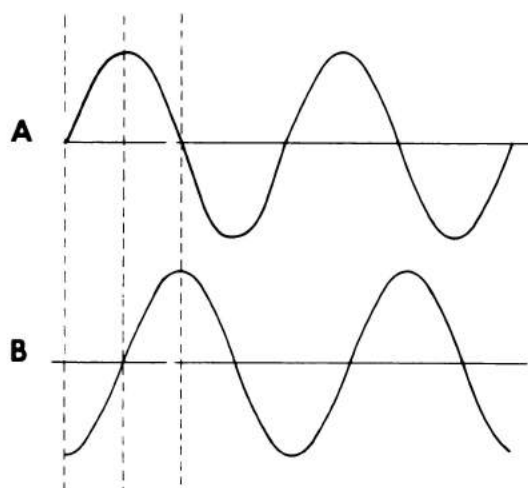
Figure 2-11
What is the phase difference?

Self-Review Answers

4. The peak voltage of the sine wave shown in Figure 2-10 is 20 volts.

5. The rms voltage of the sine wave shown in Figure 2-10 is:

    rms voltage = peak voltage × 0.707

    rms voltage = 20 volts × 0.707

    rms voltage = 14.14 volts

6. The peak-to-peak voltage of the sine wave shown in Figure 2-10 is 40 volts.

7. 240 volts rms has a peak voltage of:

    peak voltage = rms voltage × 1.414

    peak voltage = 240 × 1.414

    peak voltage = 340 volts.

8. 1 cycle per second is equal to 1 hertz. 1,000 Hz is equal to 1 kHz. 1,000 kHz is equal to 1 MHz.

9. The wavelength of a frequency of 21.42 MHz is:

$$\lambda = \frac{3 \times 10^8}{\text{frequency (Hz)}}$$

$$\lambda = \frac{3 \times 10^8}{21.42 \times 10^6 \text{ Hz}}$$

$$\lambda = 14 \text{ meters}$$

10. The wavelength of a frequency of 1875 kHz is:

$$\lambda = \frac{3 \times 10^8}{1.875 \times 10^6 \text{ Hz}}$$

$$\lambda = 160 \text{ meters}$$

11. In Figure 2-11, the sine wave in **A leads** the sine wave in **B** by **90 degrees.**

AC METERS

The moving-coil or d'Arsonval meter movement may be used to measure alternating current if the AC is converted to DC before it is applied to the meter movement. This is usually accomplished by a group of rectifier diodes. These rectifiers, which are connected between the input AC and the meter movement, allow current to flow in only one direction through the meter.

Rectifiers

One of the early rectifier diodes that was once commonly used in conjunction with meter movements is illustrated in Figure 2-12A. It consists of a copper disk which has a layer of copper oxide on one side. A lead disk is placed against the copper oxide layer and the entire unit is compressed between two metal pressure plates which are held in place by a bolt. This type of rectifier diode is commonly referred to as a **copper oxide rectifier**.
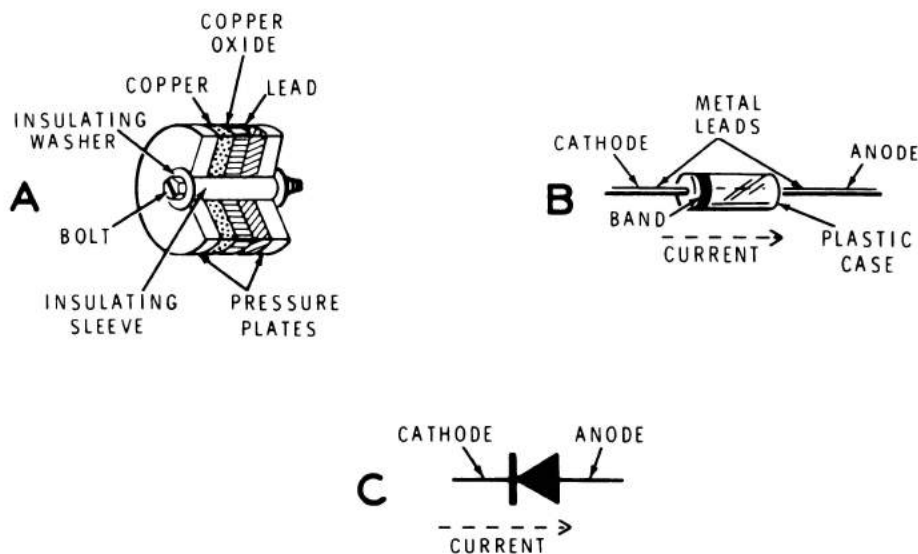
Figure 2-12
A typical copper oxide rectifier (A),
junction diode rectifier (B),
and rectifier symbol (C).

The copper oxide rectifier will allow current to flow readily from the copper to the copper oxide, but will allow only a small current to flow in the reverse direction. Therefore, it is essentially a unidirectional, or one-way, conductor of electrical current. An ideal rectifier would act as a short in one direction and an open in the other direction. However, this ideal situation cannot be achieved in practice and most practical rectifiers have substantially less than ideal characteristics.

The copper oxide rectifer has now been replaced in most applications by a more efficient device known as a **junction diode rectifier** or simply a junction diode. The junction diode is made from a semiconductor material (a material which is not a good conductor or a good insulator) such as germanium or silicon. The semiconductor material is processed in a unique manner so that it takes on unidirectional characteristics. The semiconductor device is then mounted in a glass, metal, or plastic package which protects it from environmental hazards. A typical junction diode is shown in Figure 2-12B. The device shown has two metal leads which extend from opposite ends of a plastic case. The leads are internally connected to the opposite ends of the junction diode rectifier. Current can readily flow into one lead and out of the other lead in the direction shown but only a small leakage current can flow in the opposite direction.

One end of the rectifier shown in Figure 2-12B is marked with a band. This end of the diode is generally referred to as the **cathode** and the opposite end is called the **anode**. Current always flows from the cathode end to the anode end as shown.

The schematic symbol that is commonly used to represent a rectifier (of any type) is shown in Figure 2-12C. Notice that the symbol consists of a bar and an arrow and that current (electron) flow is always opposite to the direction that the arrow is pointing as shown. In other words, current can flow from the bar to the arrow but not in the opposite direction.

The Complete AC Meter

The schematic diagram of a rectifier-type, moving-coil meter is shown in Figure 2-13A. Notice that four rectifiers are used in conjunction with one meter movement. The four rectifiers are identified as D_1, D_2, D_3, and D_4 and are arranged in what is called a **bridge rectifier** configuration. The two input terminals of the circuit are identified as A and B.

Assume that an AC generator is connected to input terminals A and B and that this generator is supplying a continuous alternating current which varies in a sinusoidal manner. Also assume that during each positive alternation of the AC sine wave, terminal A is positive with respect to terminal B. During each positive alternation the circuit current would therefore be forced to flow along the path indicated by the long dashed lines. It would flow from terminal B, through D_3, through the meter movement, through D_2, and back to terminal A. During each negative alternation, when B is positive with respect to A, current must flow along the path indicated by the short dashed lines. In other words, it would flow from terminal A, through D_1, through the meter movement, through D_4, and back to terminal B. Notice that even though the input current changes direction, the current through the meter movement is always in the same direction. The four rectifiers therefore convert the input AC into DC (actually a pulsating DC) as shown in Figure 2-13B.
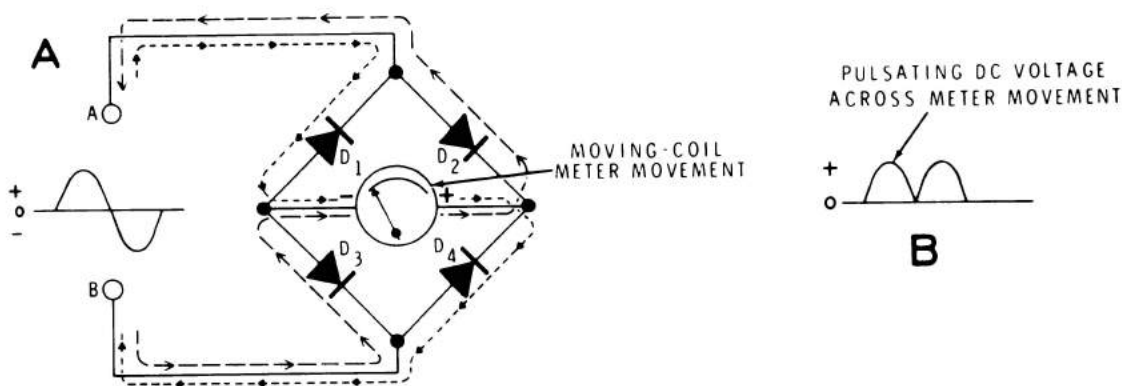
Figure 2-13
A rectifier-type, moving-coil meter.

The process of converting AC to DC is referred to as **rectification** and this is why the diodes used in this application are referred to as simply rectifiers. Furthermore, the rectifiers in this circuit convert both halves of the AC sine wave into a pulsating direct current and are therefore said to be providing **full-wave** rectification. Less complicated circuits, which use only one or two rectifiers, may also be used to provide **half-wave** rectification. In this last process, only one-half of the AC sine wave is allowed to flow through the meter movement.

The current through the meter movement flows in pulses, since each alternation rises from zero to a peak value and drops back to zero again. Unless the frequency of AC input is extremely low, the meter movement will not be able to follow the variations in the pulsating current. Instead, the meter's pointer responds to the **average value** of the AC sine wave, which is 0.636 times the peak value. However, the scale on the meter is usually calibrated in effective or rms values. In other words, the numbers on the meter scale represent effective values that are equal to 0.707 times the peak value. The effective value of a sine wave is much more important than the average value, since effective values are used in most AC calculations involving current and voltage.

Using the AC Meter

The rectifier-type, moving-coil meter is useful for measuring alternating currents over a specific frequency range. This type of meter is usually quite accurate over a frequency range that extends from approximately 10 Hz to as high as 10,000 to 15,000 Hz. Below a lower limit of approximately 10 Hz, the meter's pointer tends to fluctuate in accordance with the changes in input current, thus making it difficult to read the meter. Above the upper frequency limits just mentioned, the meter readings are usually too inaccurate to be usable. In fact, the accuracy of the meter becomes progressively worse as frequency increases beyond just a few hundred hertz. This is due to internal variations in the meter's AC resistance as the frequency increases.
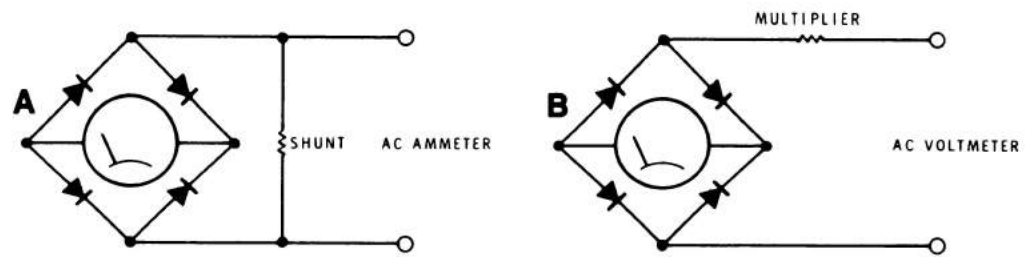
Figure 2-14
AC ammeter and AC voltmeter.

Just as a DC meter movement can be used as an ammeter or voltmeter, so too can the AC meter movement. Figure 2-14A shows that by adding a shunt to the rectifier-type, moving-coil meter, an AC ammeter is formed. Here, as with the DC ammeter, the sensitivity can be varied by changing the shunt resistor. An AC voltmeter is shown in Figure 2-14B. Notice that it uses a multiplier to increase its voltage range, just as the DC voltmeter does. When you use AC meters, follow the same rule as for DC meters. The only exception is that you do not have to observe polarity.

Self-Review Questions

12. State how a DC moving-coil meter can be used to measure alternating current. _____

_____

_____

13. An AC meter is normally calibrated to indicate _____ values of alternating current.

14. An AC rectifier type meter _____ frequency sensitive.

is/is not

Self-Review Answers

12. A DC moving-coil meter can be used to measure alternating current by using a rectifier to convert the AC to DC.

13. An AC meter is normally calibrated to indicate **rms or effective** values of alternating current.

14. An AC rectifier type meter **is** frequency sensitive.

OSCILLOSCOPES

The AC meter previously described provides a reasonably accurate measurement of current and voltage, but it does not allow you to see what the AC quantities actually look like. It is usually calibrated to indicate the effective or rms value of a sine wave and, if it is used to measure non-sinusoidal waveforms, the scale does not provide a true rms reading. It is also limited to relatively low frequencies.

To troubleshoot or analyze almost any electronic equipment, it is often necessary to know exactly what an AC waveform looks like. In many cases, it is necessary to know its peak and peak-to-peak values, its instantaneous values, and also its frequency and period. These measurements, as well as others, can be performed by using a device known as an **oscilloscope.**

The oscilloscope may be used to analyze any type of AC waveform and measure its most important electrical characteristics. The oscilloscope, or **scope** as it is commonly called, is one of the most important test instruments for measuring AC quantities, but it may also be used to measure DC quantities.

Oscilloscope Operation

Oscilloscopes do not contain moving parts like the meters previously discussed. They are electronic test instruments which use various types of electronic circuits. The early oscilloscopes used vacuum tubes as the principle controlling elements. However, most modern scopes now use semiconductor components such as transistors and IC's.

An oscilloscope is capable of measuring an AC or DC voltage and displaying it in a graphical manner. The AC or DC voltage appears as a picture on a screen which is similar to the type of screen used in a television set. The oscilloscope contains a number of controls that are used to adjust the size and the number of complete AC waveforms that are displayed. Most oscilloscopes are calibrated so that the waveform presented on the screen can be analyzed and its most important characteristics determined.
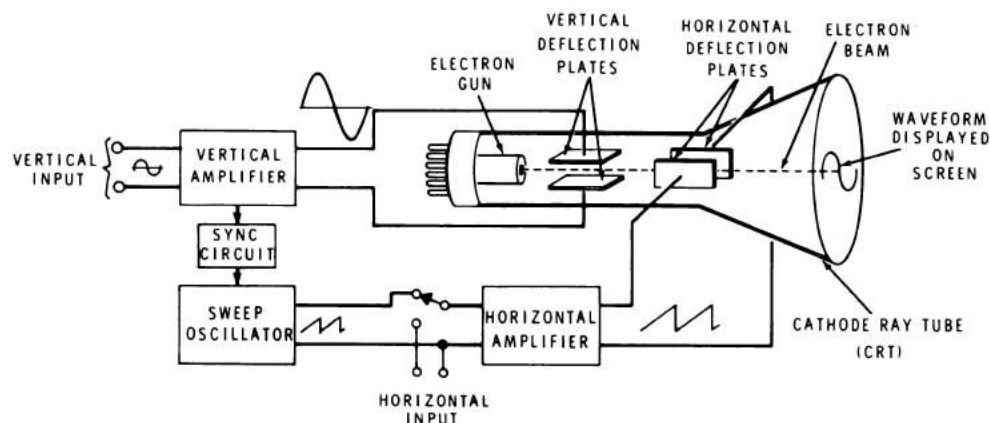
Figure 2-15
A basic oscilloscope.

A simplified block diagram of an oscilloscope is shown in Figure 2-15. It has two input terminals that are used to measure AC or DC voltages. These terminals must be connected in parallel with the voltage source to be measured, and they are referred to as the **vertical input terminals**. The voltage at these terminals is applied to an amplifier which increases the amplitude of the voltage before it is applied to the **cathode ray tube** or CRT. The CRT graphically displays the AC waveform being measured.

As shown in Figure 2-15, the CRT contains an electron gun and two sets of deflection plates. These components are mounted inside a large glass tube that fans out at one end to form a screen that closely resembles the screen on a television picture tube. The air is pumped out of the tube and the end is sealed, just as is done with an ordinary vacuum tube. The electron gun produces a stream of electrons that are focused into a narrow beam and aimed at the CRT screen. When the beam strikes the screen, it illuminates a phosphorus coating on the screen so that a spot of light is produced. This electron beam must also flow between the two sets of deflection plates.

The AC voltage from the vertical amplifier is applied across the **vertical deflection plates**. This alternating voltage causes the plates to become positively and negatively charged. Since it is an alternating voltage, the polarity of the plates will be continually changing. The electrons in the beam are negatively charged and, therefore, deflect toward the positive plate and away from the negative plate, thus causing the electron beam to bend. Since the charges on the vertical plates continually change polar-

ity, the electron beam is deflected up and down, causing a vertical trace to appear on the CRT screen as shown in Figure 2-16A. The height of this trace will depend on the amplitude of the AC voltage being measured and the amount of amplification provided by the amplifier circuit.
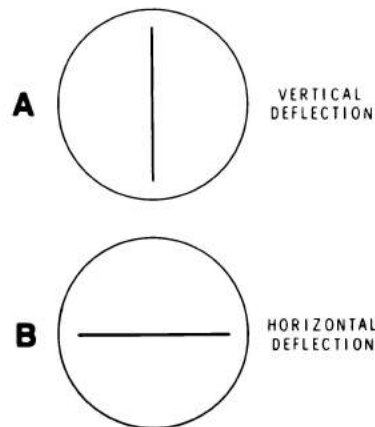
Figure 2-16
Horizontal and vertical oscilloscope traces.

If the electron beam was simply moved up and down, only a vertical line or trace would appear as shown. Such a display indicates the peak-to-peak amplitude of a waveform but would not indicate the exact shape of the waveform. In order to show how the waveform varies, it is necessary to move the electron beam horizontally across the screen. This is accomplished by the **sweep oscillator**. It generates an AC sawtooth waveform which is amplified by a **horizontal amplifier** and then applied to the **horizontal deflection plates**. The sawtooth voltage increases at a linear rate from a negative peak value to a positive voltage value and then almost instantly changes back to a negative value again. The positive and negative charges on the horizontal deflection plates vary in the same manner, thus causing the electron beam to move from left to right across the screen at a linear rate and then immediately jump back to the left side and start over again.

If only the sawtooth waveform were applied to the horizontal plates, only a horizontal trace would appear on the screen as shown in Figure 2-16B. Such a trace should be thought of as a horizontal time base upon which the vertical signal can be made to ride. The beam simply moves from left to right in a specific period of time and then repeats this action again and again.

When the vertical AC voltage and the horizontal sawtooth voltage are both applied to the CRT, an AC waveform will be displayed. As the beam moves from left to right at a linear rate with respect to time, the vertical AC voltage causes the beam to move up and down in accordance with the variations in AC voltage. If the time required for the beam to move across the screen from left to right is equal to the time required to generate one cycle of the AC input voltage, one cycle of the AC waveform will appear on the screen.
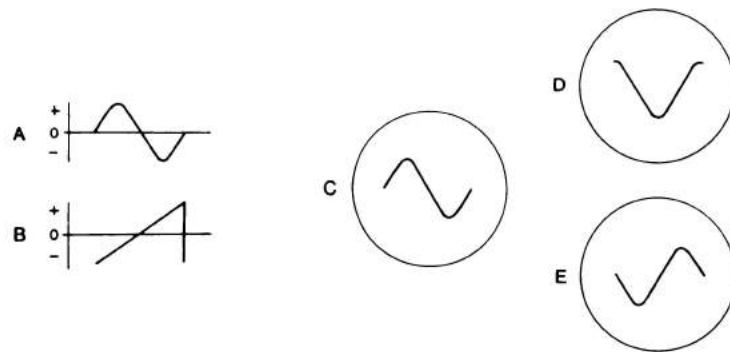
Figure 2-17
The input AC signal (A) and the sawtooth waveform (B) must have the same frequency for one cycle to be displayed (C).

The relationship between the input AC sine wave, the sawtooth wave, and the displayed waveform is shown in Figure 2-17. Notice that one complete cycle of the sine wave (Figure 2-17A) must occur in the time required to generate one complete cycle of the sawtooth wave (Figure 2-17B) in order to display one complete AC sine wave. The resulting display is shown in Figure 2-17C. In other words, the frequency of the input AC waveform must be equal to the frequency of the sawtooth waveform. Furthermore, the sine wave and sawtooth wave must begin

their cycles at the same time in order to display a sine wave that is properly oriented. If the two waveforms are not properly synchronized, the displayed waveform might appear as shown in Figure 2-17D, or Figure 2-17E. Although these are complete AC cycles, they are not properly oriented.

To insure that the input AC waveform and the sawtooth waveform are correctly synchronized, a synchronization or **sync** circuit is included in the oscilloscope. This circuit samples the incoming AC signal and produces a control signal which is applied to the sawtooth oscillator. This insures that the sawtooth begins its cycle at the proper time.

The input AC waveform and sawtooth waveform do not simply occur just once. These waveforms must occur repeatedly in order to produce a picture on the screen. That is, the electron beam follows the pattern of the waveform again and again, which results in a constant picture on the screen. The phosphor on the screen produces light for only an instant after the electron beam strikes it and moves on. Therefore, this constant repetition is necessary to produce a pattern that is constantly illuminated.

Using the Oscilloscope

The oscilloscope can be used to observe various types of AC waveforms and it can be used to measure important AC values. We will now briefly consider some of the ways in which this important test instrument can be used.

MEASURING VOLTAGE

Since the oscilloscope displays an entire AC waveform, it can be used to determine instantaneous values as well as peak-to-peak values.
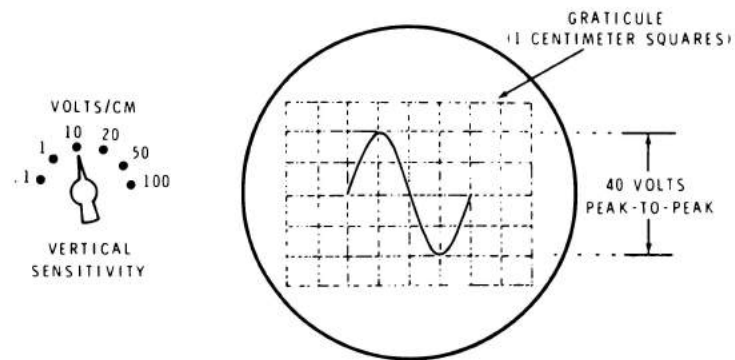
Figure 2-18
Measuring voltage with an oscilloscope.

A typical oscilloscope display is shown in Figure 2-18. Notice that the screen of the oscilloscope is marked with vertical and horizontal lines that form squares. This grid pattern is commonly referred to as a **graticule**. The squares are usually 1 centimeter high and 1 centimeter wide and are used in much the same way as a sheet of **graph paper**.

When observing an AC waveform as shown in Figure 2-18, the amplification of the vertical amplifier can be controlled so that the vertical height of the waveform can be adjusted. Furthermore, the vertical amplifier controls are usually calibrated so that a given input voltage will produce a specific amount of vertical deflection on the screen. The vertical amplification is usually adjusted by a **vertical sensitivity** control or **vertical attenuator** control. For example, suppose the vertical sensitivity control was set to the 10-volts-per-centimeter (cm) position as shown in Figure 2-18. This means that each centimeter of vertical height or deflection would represent 10 volts at the vertical input terminals. The waveform being observed in Figure 2-18 is 4 cm or 4 squares high and,

therefore, has a peak-to-peak amplitude of 4 times 10-volts or 40 volts. The peak value of the waveform is therefore equal to one-half of 40 volts, or 20 volts.

In a similar manner, the value at any point on the waveform can be determined by simply referring to the squares on the graticule. Also, the vertical sensitivity control can be set to other positions to either increase or decrease the sensitivity of the oscilloscope. When this is done, the squares represent other values of voltage which are either lower or higher than 10 volts.

Some of the less expensive oscilloscopes are not calibrated as shown in Figure 2-18. When using an oscilloscope of this type, it is necessary to apply a known DC or AC voltage to the vertical terminals and adjust the vertical amplifier for a specific amount of deflection. In other words, you calibrate the oscilloscope with a known voltage. Then you may apply the unknown voltage and determine its value.

MEASURING THE PERIOD

The oscilloscope may also be used to measure the period of an AC waveform. The **period** is the amount of time it takes to complete one cycle of the AC signal. It can be readily determined by observing the horizontal width of the waveform displayed on the screen. The oscilloscope's sawtooth oscillator can usually be adjusted so that the electron beam will move from left to right across the screen at a specific speed. The time required for the beam to move horizontally across the screen is usually referred to as the **sweep time**. The sweep time can be adjusted by a control that is mounted on the oscilloscope. This control sets the amount of time (in seconds, milliseconds, or microseconds) required for the trace to move horizontally a distance of 1 centimeter.
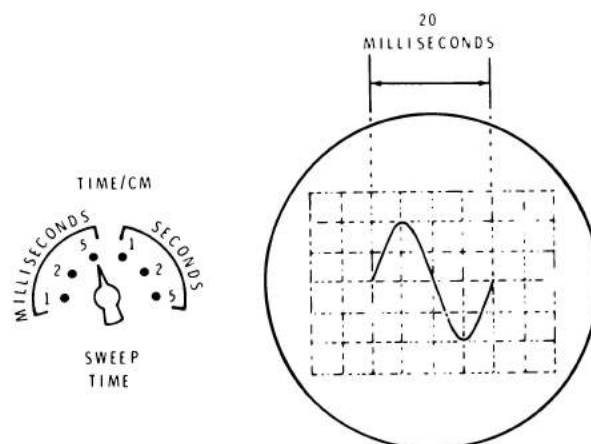
Figure 2-19
Measuring the period of an AC waveform.

Assume that the oscilloscope's sweep time control is set to the 5-milliseconds-per-centimeter position as shown in Figure 2-19. This means that each centimeter of horizontal deflection represents a time interval of 5 milliseconds. The waveform being displayed in Figure 2-19 is 4 centimeters or 4 squares wide. In other words, one complete cycle occupies 4 centimeters of the trace. Therefore, the period of the waveform is equal to 4 times 5 mS, or 20 mS.

The sweep time control can be set to various positions so that the oscilloscope can be used to measure waveforms that have very long and very short periods. In many cases, it is desirable to display only one cycle of the waveform as shown. However, the oscilloscope is capable of displaying any number of complete AC cycles on its screen. This is done by adjusting the sweep time control so that the time required to produce one complete sweep of the trace across the screen is equal to the time required to produce any given number of input AC cycles. For example, if 4 cycles of input AC occur during the time required to generate one cycle of the sawtooth sweep waveform, then 4 cycles of the input waveform will be displayed.

An oscilloscope may therefore display a number of input AC cycles. When a number of cycles are displayed on the screen, it is important to remember that it is only necessary to determine the time for one cycle in order to determine the period of the waveform.

MEASURING FREQUENCY

The frequency of an AC waveform can be determined by first measuring its period and then calculating the frequency. The frequency of an AC waveform is equal to the reciprocal of the period and is expressed mathematically as:

$$F = \frac{1}{T}$$

Where F = frequency in Hertz

T = period in seconds

For example, the waveform in Figure 2-19 has a period of 20 milliseconds. This waveform therefore has a frequency of:

$$F = \frac{1}{20 \text{ mS}}$$

$$F = \frac{1}{20 \times 10^{-3} \text{ sec.}}$$

$$F = 50 \text{ Hz}$$

The frequency is therefore 50 Hz.

This is an excellent oscilloscope that the Hobbyist can use as a test instrument.

Self-Review Questions

15. What is an oscilloscope? _____
_____
_____

16. What device does an oscilloscope use to display a waveform graphically? _____
_____

17. What path does the input AC signal follow in an oscilloscope?
_____
_____

18. What is the purpose of the sweep oscillator in an oscilloscope?
_____

19. What measurements can be made with an oscilloscope? _____
_____
_____
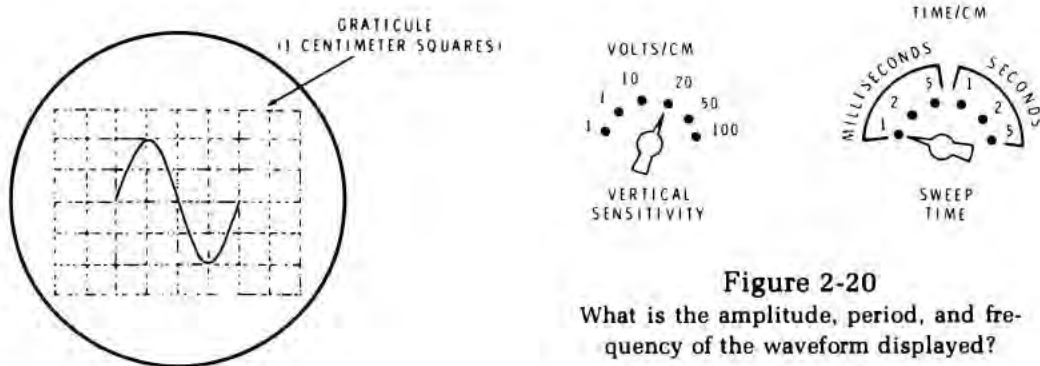
Figure 2-20
What is the amplitude, period, and frequency of the waveform displayed?

20. Refer to Figure 2-20. With the display and controls as shown, what is the waveform's peak-to-peak amplitude? _____
_____

What is the waveform's period? _____

What is the waveform's frequency? _____
_____

Self-Review Answers

15. An oscilloscope is an electronic instrument that displays complex waveforms graphically for analysis and measurement.

16. An oscilloscope uses a **cathode ray tube** or CRT to display waveforms graphically.

17. The input AC signal is applied to the **vertical input terminals**, is amplified in the **vertical amplifier**, and then applied to the **vertical deflection plates** of the CRT.

18. The sweep oscillator generates a sawtooth waveform which causes the CRT's beam to deflect horizontally, thus causing the entire AC waveform to be displayed.

19. An oscilloscope can be used to measure voltage, period, and frequency.

20. The vertical deflection of the waveform is 4 cm. Since the vertical sensitivity control is set to 20 volts/cm, the waveform's peak-to-peak amplitude is 4 × 20 volts = **80 volts.**

    The horizontal width of one cycle is 4 cm. Since the sweep time control is set to 1 millisecond/cm, the waveform period is 4 × 1 mS = **4 mS**.

    Frequency is equal to the reciprocal of the waveform's period. Therefore, the frequency is:

    $$F = \frac{1}{T} = \frac{1}{4 \text{ mS}}$$

    $$F = \frac{1}{4 \times 10^{-3} \text{ S}} = \textbf{250 Hz}$$

INDUCTANCE AND TRANSFORMERS

Inductors and transformers have a wide variety of applications. **Inductors** are used in almost every radio circuit we will discuss in this course. **Transformers** are used to step down high voltage AC to the 110 VAC you use in your home. They also have a number of applications in audio, radio, and power supply circuits.

In order to understand the operation of these two components, you must study inductance. Earlier, you learned two rules that are important to the study of inductance. First, you saw that current flowing in a conductor creates a magnetic field. Second, you saw that a moving magnetic field induces a voltage into a conductor. These two rules are the basis of inductance.

Inductance

Self-Inductance When a voltage is applied to a conductor, current flows and a magnetic field is generated by this current flow. The strength of the magnetic field builds up as the current flow rises from zero to its maximum value. As the magnetic lines of force expand outward from the center of the conductor, they induce a voltage into the conductor. The outward expansion of the magnetic field creates the relative motion between the magnetic field and conductor which induces this voltage. The polarity of the induced voltage opposes the polarity of the original voltage.

If the voltage applied to the conductor is removed, current flow stops. With no movement of electrons in the conductor, the magnetic field collapses. The collapsing lines of force cause relative motion between the conductor and the magnetic field, inducing a voltage in the conductor. The polarity of the induced voltage is such that it attempts to keep current flowing in the same direction. In other words, the induced voltage opposes the decrease of the applied voltage.

The application or removal of the voltage source causes a self-induced voltage, which takes place for any current changes that occur. Increasing or decreasing the current in a circuit causes the magnetic lines of force to expand or collapse. The magnetic lines of force induce a voltage that opposes the applied voltage. The induced voltage is referred to as **counter EMF** since it always opposes the applied voltage. The ability of a conductor to generate a voltage with a change in current is called self-inductance or simply inductance.
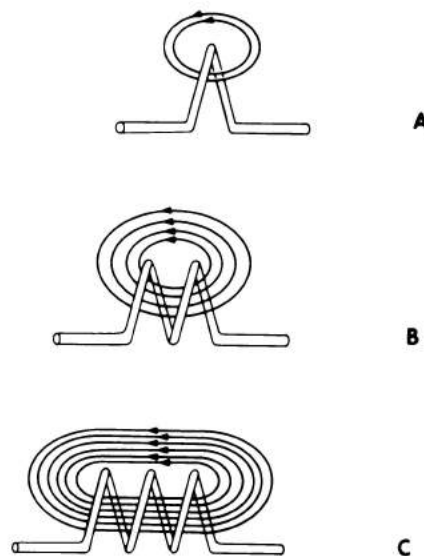
Figure 2-21
Forming a conductor into a coil increases inductance.

Inductance is defined as the ability of a device or circuit to oppose any change in current flow. The conductor we have been discussing exhibits the property of inductance because it opposes changes in current flow. For example, if the current through the conductor suddenly increases, a counter EMF will be induced which opposes the applied voltage. The counter EMF will attempt to cancel the applied voltage, and in turn will tend to hold back the increase in current.

While the conductor in the example above does exhibit inductance, much larger values of inductance are normally required. By forming the wire into a coil, we increase its inductance. Figure 2-21 shows that by increasing the number of turns, we increase the strength of the magnetic field. Since it is the magnetic field which induces the counter EMF: the stronger the magnetic field becomes, the greater the inductance will be. Therefore, the more turns of wire in a coil, the greater is its inductance. These coils of wire are called **inductors**.

Figure 2-22 shows several types of fixed inductors and their schematic symbols. Note the difference between the air core and iron core inductor's schematic symbols. An iron core greatly increases the inductance due to its higher permeability, which concentrates the magnetic flux.
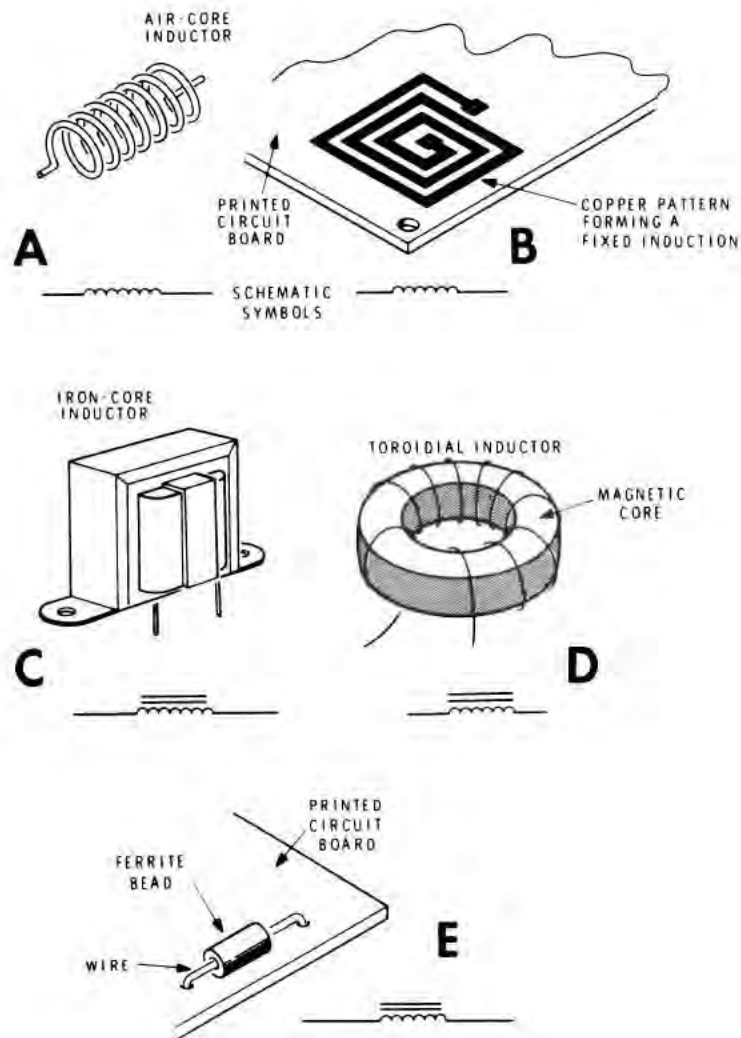
Figure 2-22

Typical fixed inductors and their schematic symbols: (A) air core coil, (B) printed circuit inductor, (C) iron core choke, (D) toroidal inductor, and (E) ferrite bead inductor.

Inductors are designed to have a specific value of inductance. The unit of inductance is the henry (H). One **henry** is defined as the amount of inductance which induces an EMF of 1 volt into a conductor when the current changes at a rate of 1 ampere per second. The henry is normally too large a unit of measurement for most circuits. We can use metric prefixes for smaller inductors. These are the millihenry (mH) and the microhenry (μH).

The symbol for inductance is L. Thus, the statement "the inductance is 10 millihenrys" can be written as an equation:

$$L = 10 \text{ mH}$$

Mutual Inductance   So far, we have been discussing self-inductance. Self-inductance occurs when the magnetic field of an inductor induces voltage into the same inductor. However, the varying magnetic field of an inductor can also induce voltage into a nearby inductor. This is known as mutual inductance.

An example of mutual inductance is shown in Figure 2-23. Here, the varying magnetic field of L_1 cuts L_2 and induces a voltage into L_2. If we were to connect a load across L_2, current would flow as a result of the induced voltage.
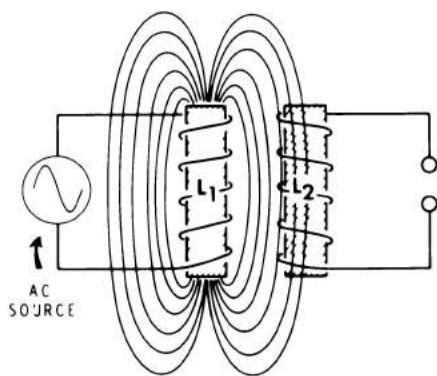
Figure 2-23
Mutual inductance.

Transformers

A common application of mutual inductance is the transformer. As shown in Figure 2-24A, a transformer consists of a primary winding (L_1) and a secondary winding (L_2). Note that both primary and secondary are wound on a single iron core, which increases the inductance and very efficiently couples the magnetic field of the primary to the secondary. The schematic symbol for a transformer is shown in Figure 2-24B.

A

L_1
PRIMARY

L_2
SECONDARY

B
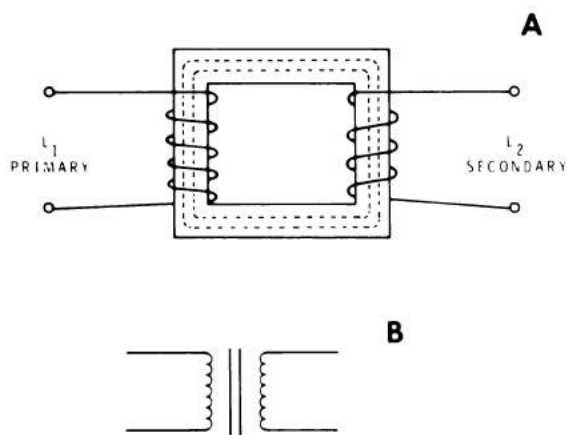
Figure 2-24
A transformer.

One purpose of a transformer is to transform voltages. By varying the ratio between the turns of wire in the primary and secondary, the amount of voltage induced into the secondary can be varied. In this manner, AC voltages can be increased or decreased by using the appropriate **turns ratio**. Transformers cannot be used to transform DC voltages, since the magnetic field must be changing to induce a voltage into the secondary. With DC, the magnetic field will not change, and no voltage will be induced into the secondary.
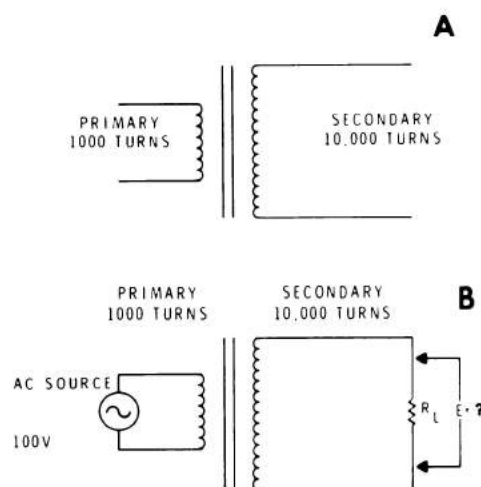
Figure 2-25
A step-up transformer.

Figure 2-25A illustrates the turns ratio of a typical transformer. The primary consists of 1000 turns of wire and the secondary consists of 10,000 turns. The turns ratio can be defined as the ratio of the number of turns in the secondary (N_S) to the number of turns in the primary (N_P):

$$\text{Turns Ratio} = \frac{N_S}{N_P}$$

Thus, in Figure 2-25A, the turns ratio is:

$$\text{Turns Ratio} = \frac{N_S}{N_P}$$

$$\text{Turns Ratio} = \frac{10,000}{1,000}$$

$$\text{Turns Ratio} = 10$$

This is usually expressed as a turns ratio of "10 to 1" or "10:1".

When the secondary has more turns, the voltage is "stepped-up" in proportion to the turns ratio. As an example, if the turns ratio is 10:1, the secondary voltage will be 10 times the primary voltage. Thus, the **voltage ratio** is equal to the turns ratio:

$$\frac{E_S}{E_P} = \frac{N_S}{N_P}$$

Where E_S = secondary voltage

E_P = primary voltage

You can use this formula to find the secondary voltage when the turns ratio and primary voltage are known. Figure 2-25B shows 100 VAC applied to the primary. The secondary voltage is found by rearranging the formula:

$$\frac{E_S}{E_P} = \frac{N_S}{N_P}$$

Multiplying both sides by E_P, we have:

$$E_P \times \frac{E_S}{E_P} = \frac{N_S}{N_P} \times E_P$$

Canceling like terms:

$$\cancel{E_P} \times \frac{E_S}{\cancel{E_P}} = \frac{N_S}{N_P} \times E_P$$

$$E_S = \frac{N_S}{N_P} \times E_P$$

$$E_S = \frac{10,000}{1,000} \times 100 \text{ V}$$

$$E_S = 10 \times 100\text{V}$$

$$E_S = 1,000 \text{ V}$$

By choosing the proper turns ratio, the voltage can be stepped-up to any value required.
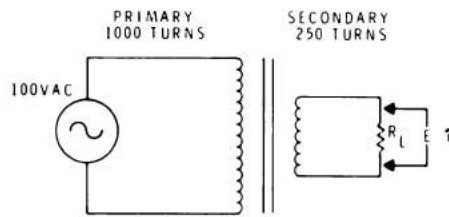
Figure 2-26
A step-down transformer.

A transformer can also be used to "step-down" a voltage. Figure 2-26 shows a transformer with 1000 turns in the primary and 250 turns in the secondary. Let's find the secondary voltage.

$$E_S = \frac{N_S}{N_P} \times E_P$$

$$E_S = \frac{250}{1000} \times 100 \text{ V}$$

$$E_S = 0.25 \times 100 \text{ V}$$

$$E_S = 25 \text{ V}$$

If any losses in a transformer are ignored, the power in the secondary will be the same as the power in the primary. Thus, in an ideal transformer the **power ratio** is 1. Although the transformer can step-up voltage, it cannot step-up power. Thus, when a transformer steps-up a voltage, it steps-down the current so that the output power is the same as the input power. Expressed as an equation:

$$P_P = P_S$$

Where P_P = power in the primary

P_S = power in the secondary

A transformer which steps-up voltage must step-down current. Otherwise, the secondary would have more power than the primary. Therefore, the **current ratio** is the opposite of the voltage or turns ratio. Thus, the current ratio is:

$$\frac{I_P}{I_S} = \frac{N_S}{N_P}$$
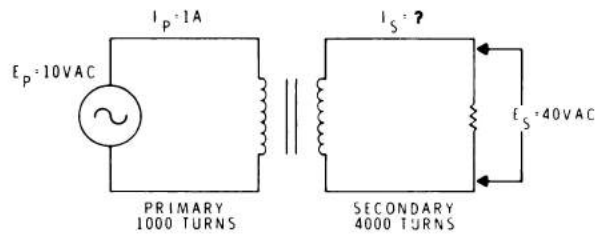
Figure 2-27
When voltage is stepped up, current is stepped down.

Figure 2-27 shows a transformer with a turns ratio of 4:1. The voltage is stepped up from 10 VAC to 40 VAC. The secondary current can be found by rearranging the current ratio formula:

$$\frac{I_P}{I_S} = \frac{N_S}{N_P}$$

By cross multiplying, we have:

$$I_S \times N_S = I_P \times N_P$$

Dividing by N_S, gives us:

$$\frac{I_S \times N_S}{N_S} = \frac{I_P \times N_P}{N_S}$$

Canceling like terms:

$$\frac{I_S \times \cancel{N_S}}{\cancel{N_S}} = \frac{I_P \times N_P}{N_S}$$

$$I_S = \frac{I_P \times N_P}{N_S}$$

$$I_S = \frac{1 \text{ A} \times 1000}{4000}$$

$$I_S = 0.25 \text{ A}$$

$$I_S = 250 \text{ mA}$$

The current is stepped down from 1 A to 250 mA, while the voltage is stepped up from 10 V to 40 V.

Figure 2-28 shows a step-down transformer. The voltage is stepped down from 100 V to 10 V. The primary current is 1 A. Let's find the secondary current.

$$I_S = \frac{N_P}{N_S} \times I_P$$

$$I_S = \frac{3000}{300} \times 1A$$

$$I_S = 10 \times 1A$$

$$I_S = 10\ A$$
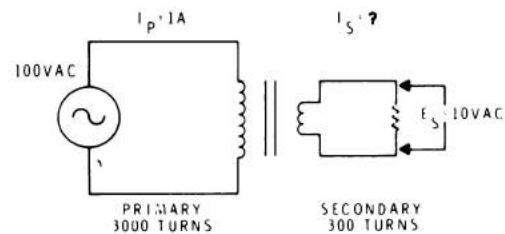
Figure 2-28
When voltage is stepped down, current is stepped up.

The current is stepped up from 1 A to 10 A, while the voltage is stepped down from 100 V to 10 V.

There is also an **impedance ratio** for transformers. Impedance is the total opposition to current in an AC circuit. The unit of impedance is the ohm, just as for DC resistance. The symbol for impedance is Z. We will be discussing impedance in detail later in this unit.

Impedance matching is one of the most important applications of trans-
formers. Maximum power is transferred from a generator to load when
their impedances match. For example, consider the AC generator shown
in Figure 2-29A. It has an internal impedance or resistance of 5 ohms. For
maximum output power from this generator, the load must also be 5
ohms. This can be proven using the power formula. For the circuit shown
in Figure 2-29B:

$$R_T = R_L + R_i$$

$$R_T = 5 \text{ ohms} + 5 \text{ ohms}$$

$$R_T = 10 \text{ ohms}$$

$$I = \frac{E}{R_T}$$

$$I = \frac{10 \text{ V}}{10 \text{ ohms}}$$

$$I = 1 \text{ A}$$

$$P_{RL} = I^2 \times R_L$$

$$P_{RL} = (1 \text{ A})^2 \times 5 \text{ ohms}$$

$$P_{RL} = 5 \text{ W}$$
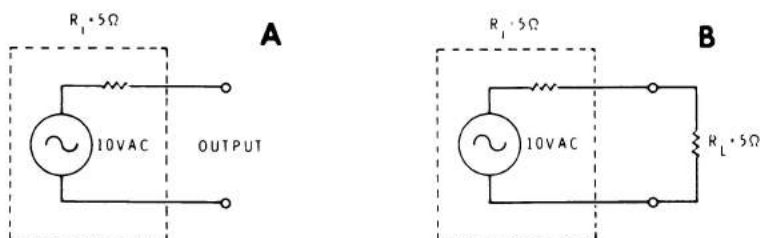
Figure 2-29
Input and output impedances must match
for maximum output power.

The output power across R_L is 5 watts. Choosing R_L values higher or lower than 5 ohms, we can prove that output power will decrease. For example, if R_L is decreased to 4 ohms, the output power across R_L decreases to 4.94 watts. If R_L is increased to 6 ohms, the output power across R_L decreases to 4.96 watts. The output power will decrease for any value of R_L other than 5 ohms. You can prove this yourself using the power formula.

A transformer can be used to match unequal impedances. For example, suppose we have a transistor amplifier whose output impedance is 100 ohms. However, we want to connect a 4 ohm loudspeaker to it. We can use a transformer with the correct turns ratio to match these impedances.

The impedance ratio of a transformer is equal to the turns ratio squared. That is:

$$\frac{Z_P}{Z_S} = \left(\frac{N_P}{N_S}\right)^2$$

Where Z_P = primary impedance

Z_S = secondary impedance

The formula can be rearranged:

$$\frac{N_P}{N_S} = \sqrt{\frac{Z_P}{Z_S}}$$

We can use this formula to solve the impedance matching problem discussed earlier. The problem is to find a turns ratio which will match a 100 ohm generator (transistor amplifier) to a 4 ohm load (speaker). Using the formula, we find:

$$\frac{N_P}{N_S} = \sqrt{\frac{Z_P}{Z_S}}$$

$$\frac{N_P}{N_S} = \sqrt{\frac{100 \ \Omega}{4 \ \Omega}}$$

$$\frac{N_P}{N_S} = \sqrt{25}$$

$$\frac{N_P}{N_S} = 5$$

Thus, the primary to secondary turns ratio must be 5:1. When the transformer is connected between the generator and load, the generator will "see " a load impedance of 100 ohms, while the load will "see" a generator impedance of 4 ohms. The impedances will be matched and maximum power will be transferred to the load.

Earlier we discussed the power ratio of transformers. While we stated that this ratio is ideally 1:1, in practical transformers it is somewhat less. This is due to losses within the transformer. The fact that the iron core gets warm indicates that some of the power supplied to the transformer is used up as heat. The two main sources of loss are **eddy currents** and **hysteresis loss**.

Eddy currents are currents which are induced, by the transformer or any inductance, into the iron core itself. These currents draw power from the primary and are lost to the secondary. The eddy current is what causes the core to heat up. One way to reduce eddy current losses is to laminate the iron core. A laminated core is made up of many thin sheets of iron rather than a large block. The laminations are insulated from one another by a thin coat of varnish. The result is a small eddy current in each lamination rather than the large current which would occur in a solid core.

Hysteresis losses occur because power is required to reverse the magnetic field in the iron core. The higher the frequency of the alternating current, the higher the hysteresis losses become. It is for this reason that radio frequency transformers use air cores.

Self-Review Questions

21. Define inductance. _____

_____

22. What is the unit of inductance? _____

_____

23. What is mutual inductance? _____

_____

24. What is the turns ratio of a transformer which has a primary voltage of 100 VAC and a secondary voltage of 500 VAC?

25. If a transformer has a primary to secondary turns ratio of 4:1, and the primary current is 3 A, what is the secondary current?

26. What is the turns ratio of a transformer whose impedance ratio is 100:1?

27. Why is it important to match impedances? _____

_____

28. Name two sources of loss in a transformer. _____

_____

Self-Review Answers

21. Inductance is the ability of a device or circuit to oppose any change in current flow.

22. The unit of inductance is the henry.

23. Mutual inductance occurs when the varying magnetic field of one inductor induces a voltage into a nearby inductor.

24. $$\frac{N_S}{N_P} = \frac{E_S}{E_P}$$

    $$\frac{N_S}{N_P} = \frac{500 \text{ V}}{100 \text{ V}}$$

    $$\frac{N_S}{N_P} = 5 \text{ or } 5:1$$

25. $$I_s = \frac{N_P}{N_S} \times I_P$$

    $$I_s = \frac{4}{1} \times 3 \text{ A}$$

    $$I_s = 12 \text{ A}$$

26. $$\frac{N_P}{N_S} = \sqrt{\frac{Z_P}{Z_S}}$$

    $$\frac{N_P}{N_S} = \sqrt{\frac{100}{1}}$$

    $$\frac{N_P}{N_S} = 10 \text{ or } 10:1$$

27. When input and output impedances are matched, maximum power is transferred from generator to load.

28. Two sources of loss in transformers are eddy currents and hysteresis loss.

INDUCTORS IN AC CIRCUITS

When an inductor is used in an AC circuit, it offers opposition to the current flow. The varying applied voltage causes a counter emf to be continuously induced into the coil. The counter emf opposes the applied voltage, and the effect is to limit the amount of current flowing in the circuit. Thus, an inductor opposes the flow of alternating current.

Inductive Reactance

The counter emf induced into an inductor by a varying current opposes the applied voltage. As a result, the total effective voltage in the circuit is the difference between the applied voltage and the induced voltage. Because the induced voltage is less than the applied voltage, the effect of the inductance is to reduce the current flow. The greater the inductance is, the greater the counter emf and therefore, the opposition to current flow is higher. The opposition to current flow offered by an inductor in an AC circuit is called **inductive reactance**. Like resistance, inductive reactance is measured in ohms.

The amount of inductive reactance offered by a coil·is directly proportional to the inductance and the frequency of the applied voltage. The greater the inductance is, the greater the magnetic field and the higher the induced voltage for a given current. The higher the induced voltage is, the greater the inductive reactance.

As the frequency of the applied AC voltage increases, the rate of change of current also increases. Thus, the magnetic lines of force expand and collapse at a faster rate. This induces a higher counter emf. The result is an increasing opposition to current flow, as the frequency increases.

Inductive reactance is represented by the symbol X_L. Inductive reactance is expressed by the formula given below:

$$X_L = 2\pi \text{ fL}$$

$2\pi = 6.28$, therefore:
$$X_L = 6.28 \text{ fL}$$

Where: X_L = inductive reactance in ohms
f = frequency in hertz
L = inductance in henries

As an example, find the inductive reactance of a 250 mH coil or inductor at 400 Hz.

$$X_L = 2\pi \, fL$$

$$X_L = 6.28 \times 400 \text{ Hz} \times 250 \text{ mH}$$

$$X_L = 6.28 \times 400 \text{ Hz} \times 0.25 \text{ H}$$

$$X_L = 628 \; \Omega$$

Let's try another example. Find the inductive reactance of a 2.5 mH inductor at 7 MHz.

$$X_L = 2\pi \, fL$$

$$X_L = 6.28 \times 7 \text{ MHz} \times 2.5 \text{ mH}$$

$$X_L = 6.28 \times 7 \times 10^6 \text{ Hz} \times 2.5 \times 10^{-3} \text{ H}$$

When multiplying powers of 10, add the exponents algebraically, then multiply the base numbers. Therefore, adding the exponents we have:

$$X_L = 6.28 \times 7 \times 2.5 \times 10^3$$

$$X_L = 109.9 \times 10^3$$

$$X_L = 109.9 \text{ k}\Omega$$

Inductors in Series and Parallel

When inductors are placed in series or parallel, they behave much like resistors. The circuit of Figure 2-30A shows three inductors in series. The total inductance is the sum of the individual inductances:

$$L_T = L_1 + L_2 + L_3$$

When inductors are placed in parallel, as shown in Figure 2-30B, the total inductance can be computed using the equation below:

$$L_T = \frac{L_1 \times L_2}{L_1 + L_2}$$
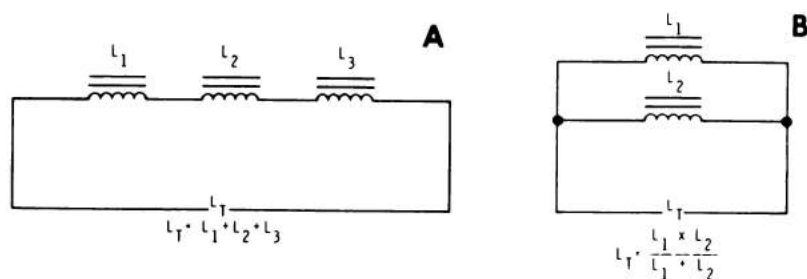
Figure 2-30
Inductive circuits.

To find total inductance when three or more inductors are in parallel, use the formula:

$$L_T = \cfrac{1}{\cfrac{1}{L_1} + \cfrac{1}{L_2} + \cfrac{1}{L_3} \text{ etc.}}$$

Note that these formulas are very similar to the ones used to compute the total resistance of a series or parallel circuit.

Ohm's Law in Inductive Circuits

Ohm's Law applies equally to inductive AC circuits as it does to resistive circuits. The basic Ohm's Law equation becomes:

$$I = \frac{E}{X_L}$$

One example of using Ohm's Law in an inductive AC circuit is to calculate the current flowing in a 2.5 mH inductor when a 10 volt, 100 kHz sine wave is applied. We must first compute the inductive reactance:

$$X_L = 2\pi \, fL$$

$$X_L = 6.28 \times 100 \text{ kHz} \times 2.5 \text{ mH}$$

$$X_L = 6.28 \times 100 \times 10^3 \text{ Hz} \times 2.5 \times 10^{-3} \text{ H}$$

$$X_L = 1570 \ \Omega \text{ or } 1.57 \text{ k}\Omega$$

Ohm's Law can now be used:

$$I = \frac{E}{X_L}$$

$$I = \frac{10 \text{ V}}{1570 \,\Omega}$$

$$I = 0.00637 \text{ A or } 6.37 \text{ mA}$$

Phase Angle in Inductive Circuits

A unique relationship exists between the applied voltage, the induced voltage, and the current in an inductive circuit. Because of this unique relationship, the applied voltage and current will be out of phase with one another. The current in an inductor lags the voltage by 90 degrees. This phase shift or lagging of the current is due to the inductive action which opposes changes in current flow. This inductive action, which is due to the induced voltage, actually causes the current to lag behind the applied voltage. This phase relationship can also be stated: The voltage across an inductor **leads** the current by 90 degrees. This is shown in Figure 2-31.
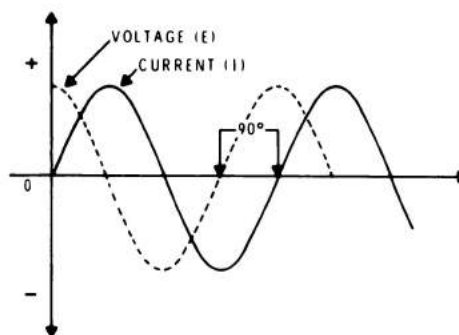
Figure 2-31
Voltage leads current by 90° in an inductive circuit.

The phase relationship in an inductive circuit can also be shown using phasors or vectors. A phasor diagram for an inductive circuit is shown in Figure 2-32. Phasors represent the values of current and voltage and also show the phase relationship which exists. Note that the current (I) phasor is drawn on the 0 degree line. Since we are dealing with a series circuit, and the current is the same everywhere in a series circuit, we use the current phasor as our reference. The voltage phasor (E) is shown at the +90 degree point because it is leading the current phasor by 90 degrees. The length of each phasor represents amplitude.
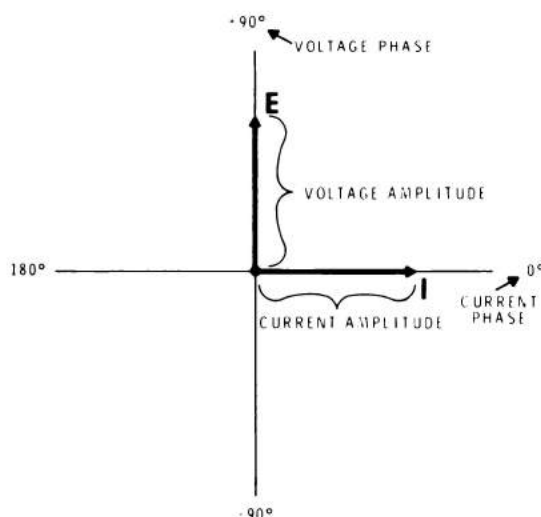
Figure 2-32
Phasor diagram for an inductive circuit.

Series RL Circuits

Circuits which contain both inductance and resistance are referred to as RL circuits. An example of a **series RL circuit** is shown in Figure 2-33. Note that both the inductive reactance and the resistance limit the circuit current.
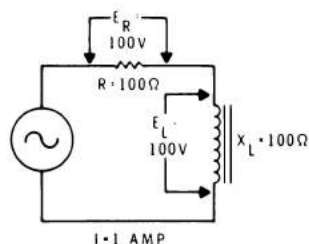
Figure 2-33
A series RL circuit.

The phasor diagram for this circuit is shown in Figure 2-34A. The current phasor (I) is drawn on the 0° line. Since the current is the same throughout a series circuit, we use it as the reference phase. The inductor voltage (E_L) must lead the current by 90°, therefore, the E_L phasor is drawn on the +90° line. Since there is no phase shift across a resistance, the resistor voltage (E_R) phasor is drawn on the 0° line.
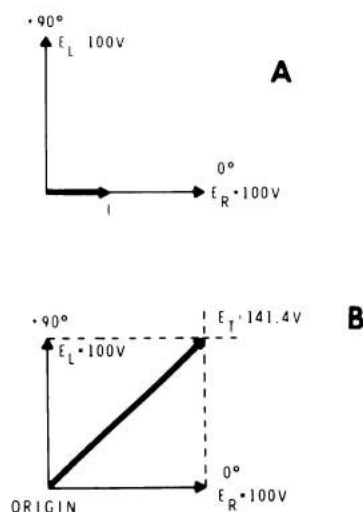
Figure 2-34
Finding total voltage in an RL circuit.

When combining these voltage drops to determine the total applied voltage (E_T), we must take the phase difference into account. One way to find E_T is to use a phasor diagram. This is shown in Figure 2-34B. Drawing lines parallel to the E_L and E_R phasors, we form a rectangle. The length of the phasor drawn from the origin diagonally across the rectangle indicates the value of E_T.

While a phasor diagram can be used to find E_T, it is not an accurate method. The primary use of phasor diagrams is in visualizing what occurs in an AC circuit. You will also find them a valuable aid when solving complex AC circuit problems.

The accurate method of determining E_T is to use the equation:

$$E_T = \sqrt{(E_R)^2 + (E_L)^2}$$

Using this equation to solve the example problem, we find:

$$E_T = \sqrt{(E_R)^2 + (E_L)^2}$$

$$E_T = \sqrt{(100 \text{ V})^2 + (100 \text{ V})^2}$$

$$E_T = \sqrt{10,000 + 10,000}$$

$$E_T = \sqrt{20,000}$$

$$E_T = 141.4 \text{ V}$$

We can also solve for the total opposition to current flow or **impedance** (Z) of an RL circuit. Impedance is the total opposition to current flow in an AC circuit. In a series RL circuit, impedance is the phasor sum of inductive reactance and resistance. The formula for impedance in a series RL circuit is:

$$Z = \sqrt{R^2 + X_L^2}$$

Where Z =   impedance in ohms

R =   resistance in ohms

X_L =   inductive reactance in ohms
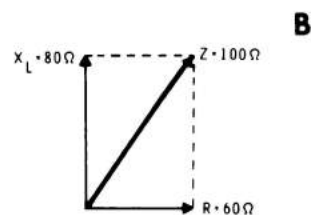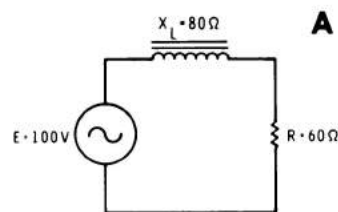
Figure 2-35
Finding impedance in an RL circuit.

We can use this equation to find the impedance of the circuit shown in Figure 2-35A.

$$Z = \sqrt{R^2 + X_L^2}$$

$$Z = \sqrt{(60\ \Omega)^2 + (80\ \Omega)^2}$$

$$Z = \sqrt{3600 + 6400}$$

$$Z = \sqrt{10{,}000}$$

$$Z = 100\ \Omega$$

The phasor diagram for impedance is shown in Figure 2-35B.

Now that we know the total impedance, we can find the current by using Ohm's Law:

$$I = \frac{E}{Z}$$

$$I = \frac{100\ V}{100\ \Omega}$$

$$I = 1\ A$$

Self-Review Questions

29. What is inductive reactance? _____

_____

30. Write the formula for inductive reactance.

_____

31. What is the inductive reactance of a 1 μH inductor at 10 MHz?

32. What is the total inductance of the circuit shown in Figure 2-36A?

33. What is the total inductance of the circuit shown in Figure 2-36B?
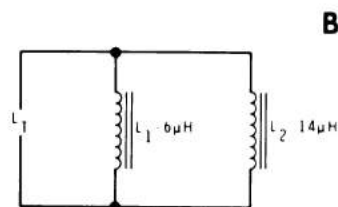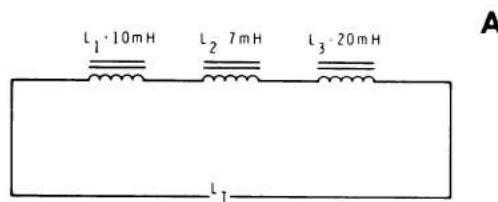
Figure 2-36
Find total inductance.

34. In an inductor, the _____ **leads** _____ by 90°.

35. What is impedance? _____

_____

36. What is the impedance of the circuit shown in Figure 2-37?

37. What is the current of the circuit shown in Figure 2-37?
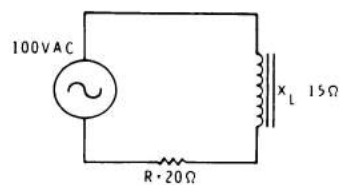
Figure 2-37
Find total impedance.

Self-Review Answers

29. Inductive reactance is the opposition to current flow offered by an inductor in an AC circuit.

30. The formula for inductive reactance is:

$$X_L = 2\pi \, fL$$

31. $X_L = 2\pi \, fL$

$X_L = 6.28 \times 10 \text{ MHz} \times 1\mu\text{H}$

$X_L = 6.28 \times 10 \times 10^6 \text{ Hz} \times 1 \times 10^{-6} \text{ H}$

$X_L = 62.8 \, \Omega$

32. $L_T = L_1 + L_2 + L_3$

$L_T = 10 \text{ mH} + 7 \text{ mH} + 20 \text{ mH}$

$L_T = 37 \text{ mH}$

33. $L_T = \dfrac{L_1 \times L_2}{L_1 + L_2}$

$L_T = \dfrac{6 \, \mu\text{H} \times 14 \, \mu\text{H}}{6 \, \mu\text{H} + 14 \, \mu\text{H}}$

$L_T = \dfrac{6 \times 10^{-6} \times 14 \times 10^{-6}}{6 \times 10^{-6} + 14 \times 10^{-6}}$

$L_T = \dfrac{84 \times 10^{-12}}{20 \times 10^{-6}}$

$L_T = 4.2 \times 10^{-6} \text{ H}$

$L_T = 4.2 \, \mu\text{H}$

34. In an inductor, the **voltage** leads **current** by 90°.

35. Impedance is the total opposition to current flow in an AC circuit.

36. $Z = \sqrt{R^2 + X_L^2}$

$Z = \sqrt{(20\ \Omega)^2 + (15\ \Omega)^2}$

$Z = \sqrt{400 + 225}$

$Z = \sqrt{625}$

$Z = 25\ \Omega$

37. $I = \dfrac{E}{Z}$

$I = \dfrac{100\ V}{25\ \Omega}$

$I = 4\ A$

CAPACITANCE

Capacitance is the property of a circuit or device that enables it to store electrical energy by means of an electrostatic field. A device especially designed to have a certain value of capacitance is called a capacitor. The capacitor has the ability to store electrons and release them at a later time. The number of electrons that it can store for a given applied voltage is a measure of its capacitance.

Capacitors have a number of important uses in electronics. In fact, it is virtually impossible to find an electronic circuit that doesn't use capacitors. In this section, you will learn what a capacitor is and how it behaves in DC and AC circuits.

Types of Capacitors

A capacitor is an electronic component that can store electrical energy in the form of an electric field. The simplest form of a capacitor is shown in Figure 2-38. It is composed of two parallel metal plates which are separated by an insulator called the **dielectric**. For practical applications, a capacitor may be constructed quite differently than this. The construction of a paper capacitor is shown in Figure 2-39. Two thin sheets of metal foil are separated by a sheet of paper, and additional sheets of paper are placed on the top and bottom of the foil sheets. This is then wound into a compact cylinder, with leads attached to the metal sheets, and the unit is sealed with wax or plastic.
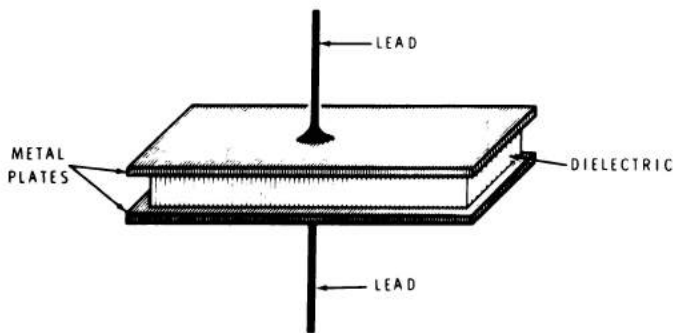
Figure 2-38
The capacitor.
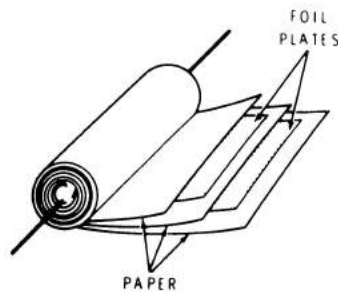
Figure 2-39
Construction of the paper capacitor.

SCHEMATIC
SYMBOL
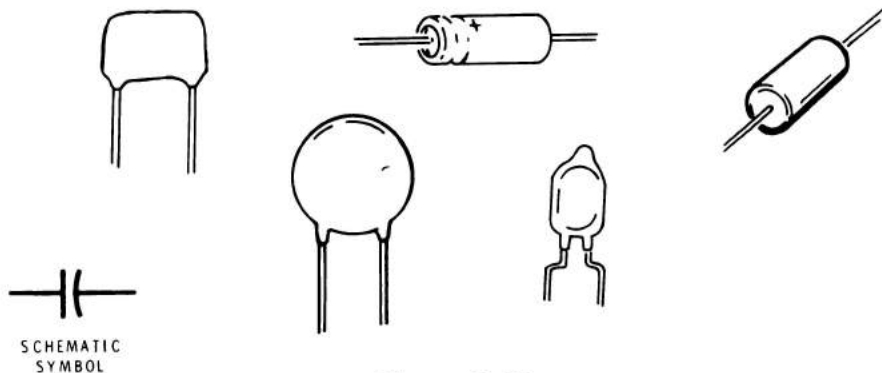
Figure 2-40
Types of capacitors.

Capacitors are generally classified by their dielectric: paper, ceramic, mica, polystryrene, tantalum, and electrolytic. Some of the types of capacitors are shown in Figure 2-40.

A variable capacitor is shown in Figure 2-41, along with its schematic symbol. Its capacitance varies when the shaft is rotated. You may have seen this type of capacitor in an AM or FM receiver. It has an air dielectric.

ROTATING PLATES
(ROTOR)
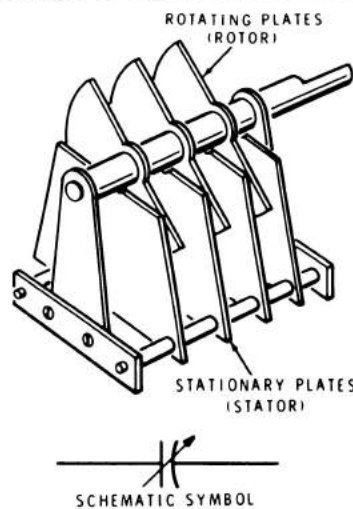
Figure 2-41
The variable capacitor.

STATIONARY PLATES
(STATOR)

SCHEMATIC SYMBOL

Charging and Discharging a Capacitor

When we apply a DC voltage to a capacitor, current will flow as the capacitor is charged. When the capacitor has reached its full charge, current will no longer flow in the circuit. This can be seen in Figure 2-42. In Figure 2-42A, no voltage has been applied to the capacitor and there is an equal number of free electrons on each plate. There will be no potential difference across the plates of the capacitor. When the switch is closed, as

in Figure 2-42B, the battery will draw the free electrons from the top plate of the capacitor to its positive terminal and the free electrons from the battery's negative terminal will be attracted to the bottom plate of the capacitor. This produces current flow, which is limited by resistor R_1.

The absence of free electrons on the top plate give it a positive charge, and the excess of free electrons on the bottom plate give it a negative charge. This results in a potential difference across the plates of the capacitor. The capacitor will continue to charge until the difference of potential between the two plates is the same as the voltage across the battery. In the example shown, current will flow until the charge on the capacitor builds up to 10 volts. Once the capacitor has the same voltage as the battery, no additional current can flow because there is no longer a difference of potential between the battery and the capacitor.

Figure 2-42C shows that once the capacitor is charged, the switch can be opened and the capacitor will retain its charge. It will retain this charge in the form of an electric field which exists between the capacitor's plates.

Although a charge current does flow in the circuit, it does **not** flow **through** the capacitor. Electrons flow out of the positive plate and into the negative plate. They can not flow through the capacitor, due to the insulating dielectric.

The capacitor can be discharged by shorting the plates together. Discharge current will flow, which will equalize the number of free electrons on each plate. If each plate has the same number of free electrons, there will be no potential difference across the plates and the capacitor will be completely discharged.
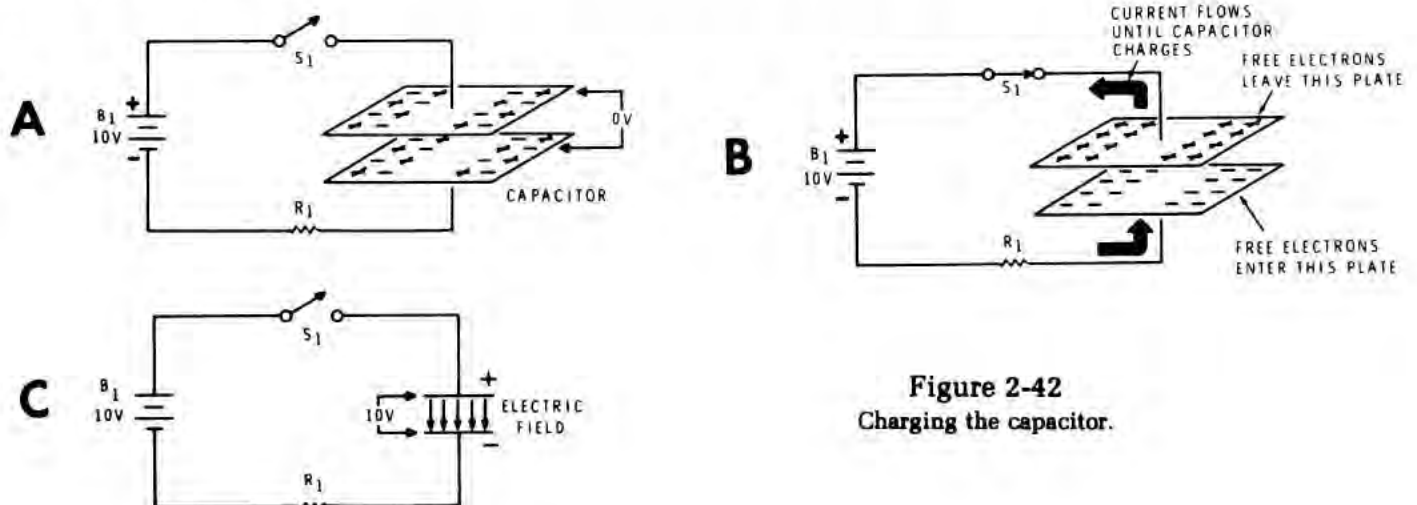
Figure 2-42
Charging the capacitor.

Unit of Capacitance

Capacitance is a measure of the amount of charge a capacitor can store for a given amount of applied voltage. The unit of capacitance is the **farad** (F or fd). One farad is the amount of capacitance that will store a charge of one coulomb when an emf of one volt is applied. A coulomb is the unit of electrical charge, and is equal to 6.25×10^{18} electrons. A one farad capacitor would be enormous. For this reason, metric prefixes are used to describe practical values of capacitance. The commonly used values of capacitance are the microfarad (μF) and the picofarad (pF).

Factors Affecting Capacitance

Capacitance is determined by three factors:

1. The area of the metal plates;

2. The spacing between the plates;

3. The nature of the dielectric.

This is illustrated in Figure 2-43. Figure 2-43A shows a capacitor made up of two 1-inch square plates separated by 0.001 inch. The dielectric is air. This component has a capacitance of 225 pF (picofarad). We will use this capacitor as our reference and see what happens when we change the area of the plates, the spacing of the plates and the dielectric material.

Figure 2-43B shows what happens when the area of the plates is doubled but all other factors are held constant. Notice that there is now twice as much area in which the electrostatic field can exist. This doubles the capacity or capacitance to 450 pF. Thus, capacitance is directly proportional to the area of the plates.

Figure 2-43C shows that the capacitance can also be doubled by reducing the spacing between the plates to one half its former distance. The path of the electrostatic lines of force is cut in half. This doubles the strength of the field which, in turn, doubles the capacitance. Thus, the capacitance is inversely proportional to the spacing between the plates.

Figure 2-43D shows that the capacitance can be increased by using a better dielectric than air. Air is actually a very poor dielectric. Most insulators support electrostatic lines of force more easily than air. The ease with which an insulator supports electrostatic lines of force is
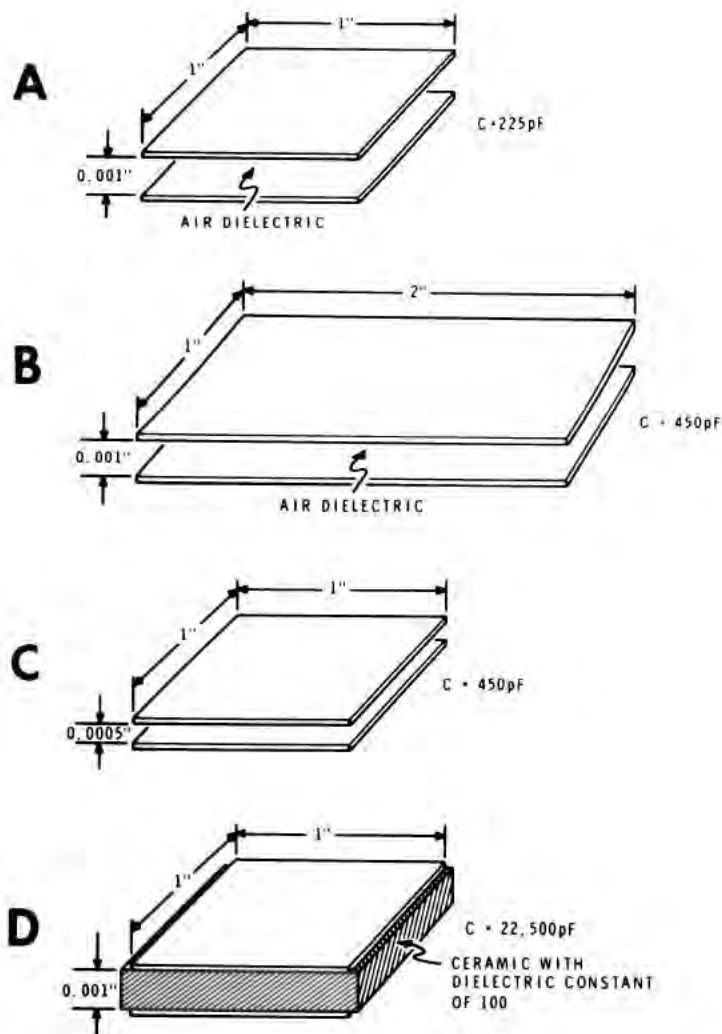
Figure 2-43
Factors which determine capacitance.

indicated by its **dielectric constant**. Air has a dielectric constant of 1, and is used as the reference. The ceramic shown in Figure 2-43D has a dielectric constant of 100, which increases the capacitance by a factor of 100. Thus, capacitance is directly proportional to the dielectric constant.

The variable capacitor of Figure 2-41 changes its capacitance by changing the area of the plates. When the rotating plates are fully meshed with the stationary plates, the capacitance is at a maximum. When the plates are rotated to the fully open position, the capacitance is at a minimum.

Parallel and Series Connected Capacitors

Figure 2-44A shows the 225 pF capacitor discussed earlier. In Figure 2-44B, two of these capacitors are connected in parallel. Of the three factors which determine capacitance, only one has changed. The dielectric constant and the spacing between the plates is the same as before, but the effective area of the two plates has doubled. Recall that capacitance is directly proportional to the area of the plates. Therefore, the total capacitance is twice that of a single capacitor.
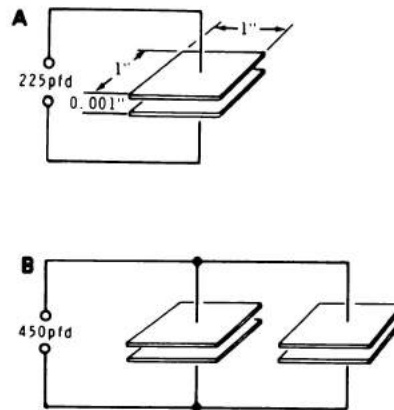
Figure 2-44
Connecting capacitors in parallel
increases capacitance.

This shows that connecting capacitors in parallel is equivalent to adding the plate areas. Consequently, the total capacitance is equal to the sum of the individual capacitance values. If three capacitors are connected in **parallel**, the total capacitance (C_T) is found by adding the individual values:

$$C_T = C_1 + C_2 + C_3$$

Notice that capacitors in parallel add like resistors in series.
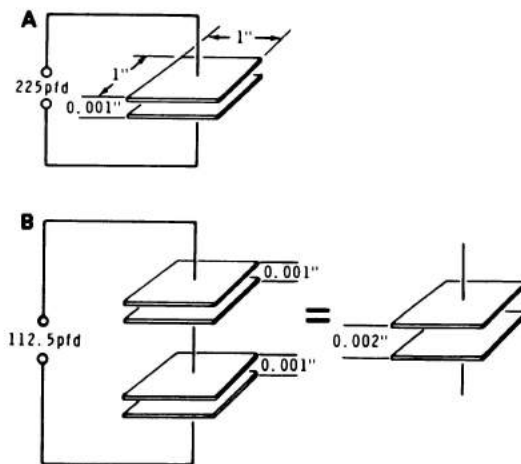
Figure 2-45
Connecting capacitors in series
decreases capacitance.

Capacitors which are connected in parallel will all charge to the same voltage. Remember the voltage is the same across each branch of a parallel network.

Figure 2-45 shows that placing these two capacitors in series is equivalent to doubling the thickness of the dielectric. That is, two series capacitors act like a single capacitor which has a dielectric thickness equal to the sum of the individual dielectric thicknesses. Since capacitance is inversely proportional to the spacing between the plates, doubling the thickness of the dielectric cuts the total capacitance value to one half that of the single capacitor.

The total capacitance of two capacitors in **series** can be found using the formula:

$$C_T = \frac{C_1 \times C_2}{C_1 + C_2}$$

Notice that this equation has the same form as the equation for resistors in parallel.

Capacitive Reactance

It is the basic nature of a capacitor to oppose changes in voltage. In a DC circuit, a capacitor will charge to the applied voltage. If the applied voltage is increased, the capacitor will then charge to the new higher voltage. If the applied voltage is decreased, the capacitor will discharge until the voltage across it equals the new lower applied voltage. However, the charge and discharge of the capacitor will take a finite time, determined by the resistance and capacitance in the circuit.

In an AC circuit, the capacitor is constantly charging and discharging due to the constantly changing voltage, and the voltage across the capacitor will be in constant opposition to the applied voltage. This constant opposition to changes in the applied voltage creates an opposition to current flow in the circuit. This opposition to the flow of AC offered by a capacitor is called **capacitive reactance**. Capacitive reactance is represented by the symbol X_C and, like resistance and inductive reactance, is measured in ohms.

The reactance of a capacitor is determined by its capacitance and the frequency of the applied voltage. In fact, capacitive reactance is inversely proportional to both capacitance and frequency. This means that increasing the capacitance or the frequency will cause the reactance to decrease.

Capacitive reactance is found using the formula:

$$X_C = \frac{1}{2\pi\, fC}$$

Where X_C = capacitive reactance in ohms

2π = 6.28

f   = frequency of the applied voltage in Hz

C   = the capacitance in farads

The following examples show how to use the capacitive reactance formula. What is the reactance of a 10 μF capacitor at 100 Hz?

$$X_C = \frac{1}{2\pi\, fC}$$

$$X_C = \frac{1}{6.28 \times 100 \text{ Hz} \times 10 \, \mu\text{F}}$$

$$X_C = \frac{1}{6.28 \times 100 \text{ Hz} \times 10 \times 10^{-6} \text{ F}}$$

$$X_C = \frac{1}{6.28 \times 10^{-3}}$$

When dividing by a negative power of 10, you must still subtract the exponents. In this case, the numerator is 1 or 10^0. The denominator is 10^{-3}. Subtracting the exponents, we have:

$$0 - (-3) = 0 + 3 = 3$$

Therefore, the power of ten in the answer will be 10^3. Getting back to the example problem, we have:

$$X_C = \frac{1}{6.28} \times 10^3$$

$$X_C = 0.1592 \times 10^3$$

$$X_C = 159.2 \, \Omega$$

What is the reactance of the series combination of a 10 μF and a 20 μF capacitor, when the applied voltage has a frequency of 1 kHz? We must first find the total capacitance.

$$C_T = \frac{C_1 \times C_2}{C_1 + C_2}$$

$$C_T = \frac{10 \, \mu\text{F} \times 20 \, \mu\text{F}}{10 \, \mu\text{F} + 20 \, \mu\text{F}}$$

$$C_T = \frac{2 \times 10^{-10}}{3 \times 10^{-5}}$$

$$C_T = 6.67 \, \mu\text{F}$$

Now we can use the equation for capacitive reactance:

$$X_C = \frac{1}{2\pi \, fC}$$

$$X_C = \frac{1}{6.28 \times 1 \text{ kHz} \times 6.67 \text{ } \mu F}$$

$$X_C = \frac{1}{6.28 \times 1 \times 10^3 \text{ Hz} \times 6.67 \times 10^{-6} \text{ F}}$$

$$X_C = \frac{1}{4.19 \times 10^{-2}}$$

$$X_C = 23.9 \text{ } \Omega$$

We could have calculated the individual reactances and then, since they are in series, added them to find the total reactance.

In an AC circuit containing only capacitive reactance, Ohm's Law can be used to find reactance, current, or voltage.

The basic Ohm's Law equation then becomes:

$$I = \frac{E}{X_C}$$

When finding total reactance in series or parallel, use the following formulas:

Series Circuit

$$X_T = X_{C1} + X_{C2} + X_{C3}$$

or $$X_T = X_{L1} + X_{L2} + X_{L3}$$

Parallel Circuit

$$X_T = \frac{X_{C1} \times X_{C2}}{X_{C1} + X_{C2}}$$

or $$X_T = \frac{X_{L1} \times X_{L2}}{X_{L1} + X_{L2}}$$

Note: Combine only like reactances with these formulas, i.e. X_C with X_C or X_L with X_L.

Let's try an example problem. Two capacitors having reactances of 10 Ω and 25 Ω are connected in series. What is the total reactance?

$$X_T \quad = X_{C1} + X_{C2}$$

$$X_T \quad = 10\ \Omega + 25\ \Omega$$

$$X_T \quad = 35\ \Omega$$

If the applied voltage is 70 V, what is the current?

$$I \quad = \frac{E}{X_C}$$

$$I \quad = \frac{70\ V}{35\ \Omega}$$

$$I \quad = 2\ A$$

Phase Angle in Capacitive Circuits

Just as in an inductive AC circuit, the voltage and current are out of phase with one another in a capacitive AC circuit. There is one important difference: in a capacitive AC circuit the current leads the voltage by 90°. This is just the opposite of an inductive AC circuit. An easy way to remember the phase shifts in AC circuits is the phrase: ELI the ICE man. ELI indicates the voltage (E) in an inductor (L) leads the current (I) by 90°. ICE indicates that the current (I) in a capacitor (C) leads the voltage (E) by 90°.
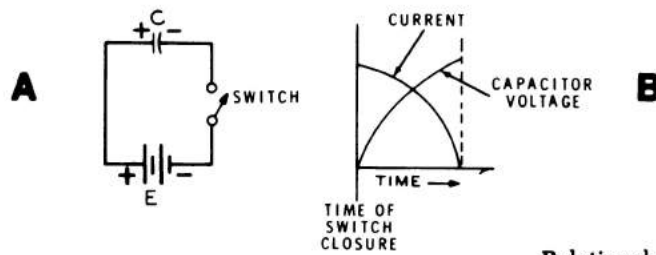
Figure 2-46
Relationship between voltage and current
in a charging capacitor.

The reason that current leads the voltage in a capacitive circuit can best be explained by referring to the DC circuit shown in Figure 2-46A. Initially, the capacitor is completely discharged. When the switch is closed, the voltage across the capacitor is zero; voltage does not appear across the capacitor until it starts to charge. When the capacitor has charged completely, the voltage across it is equal to the supply voltage. The current flow initially is at a maximum, as it is charging the capacitor. This current will gradually decrease as the capacitor charges, until finally, when the capacitor is fully charged, it is reduced to zero. This relationship between the current and voltage is shown in Figure 2-46B.
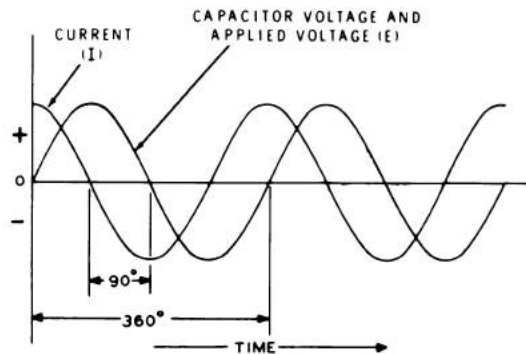
Figure 2-47
Current and voltage relationship in a
purely capacitive AC circuit.

The exact relationship between the current and voltage in a capacitive circuit when a sine-wave AC signal is applied as shown in Figure 2-47. Note that when the current is at a maximum, the voltage is at zero. This conforms to our drawing of the relationship between voltage and current in a charging capacitor (Figure 2-46B). The only difference here is that the capacitor is continually charging and discharging as the AC voltage changes.

RC Circuits

Just as with inductive AC circuits, we use phasor diagrams to help us visualize what is happening in a capacitive AC circuit.

A series RC circuit is shown in Figure 2-48A, and the phasor diagram for this circuit is shown in Figure 2-48B. Note that we have plotted the voltage phasors in this example. The current is again our reference as this is a series circuit, and the current is the same at all points in a series circuit. The phasor representing the voltage drop across the resistor is also shown on the zero degree line, as there is no phase shift across a resistor. The phasor representing the voltage drop across the capacitor is drawn at the -90 degree point, as current leads voltage in a capacitive circuit; this can also be stated: voltage lags current in a capacitive circuit.

When calculating the total applied voltage, we must take the phase difference into account. Therefore, we must use the equation:

$$E_T = \sqrt{E_R^2 + E_C^2}$$

We can now apply this equation to the circuit of Figure 2-48.

$$E_T = \sqrt{E_R^2 + E_C^2}$$

$$E_T = \sqrt{(60 \text{ V})^2 + (150 \text{ V})^2}$$

$$E_T = \sqrt{3.6 \times 10^3 + 2.25 \times 10^4}$$

$$E_T = \sqrt{2.61 \times 10^4}$$

$$E_T = 161.6 \text{ V}$$
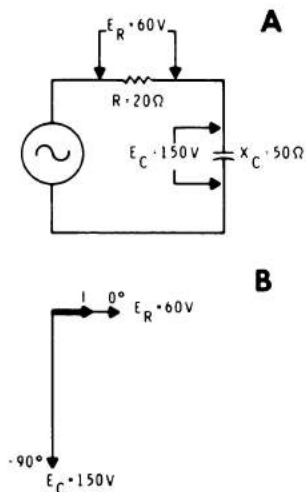
Figure 2-48
A series RC circuit.

Calculating the total impedance of an RC circuit is much the same as is done in an RL circuit. The impedance phasor diagram is shown in Figure 2-49, for the circuit of Figure 2-48, taking the phase difference into account, we use the following equation to find the impedance of a series RC circuit:

$$Z = \sqrt{R^2 + X_C^2}$$
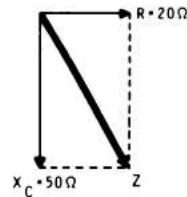
Figure 2-49
Impedance phasor diagram for a series RC circuit.

We can use this equation to find the impedance of the circuit shown in Figure 2-48.

$$Z = \sqrt{R^2 + X_C^2}$$

$$Z = \sqrt{(20\ \Omega)^2 + (50\ \Omega)^2}$$

$$Z = \sqrt{4 \times 10^2 + 2.5 \times 10^3}$$

$$Z = \sqrt{2.9 \times 10^3}$$

$$Z = 53.9\ \Omega$$

We can now use Ohm's Law to find the circuit current:

$$I = \frac{E}{Z}$$

$$I = \frac{161.6\ V}{53.9\ \Omega}$$

$$I = 3\ A$$

Self-Review Questions

38.   Define capacitance. _____

_____

39.   What is the unit of capacitance? _____

_____

40.   State the three factors which determine capacitance.

1.   _____

2.   _____

3.   _____

41.   What is the total capacitance of the circuit shown in Figure 2-50A?
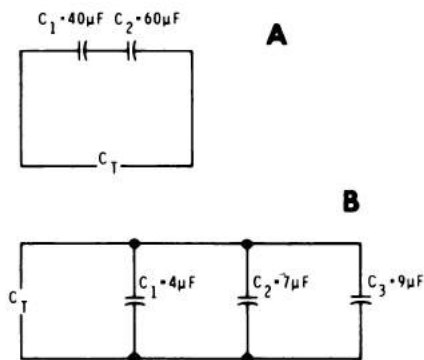
Figure 2-50
Find total capacitance.

42.   What is the total capacitance of the circuit shown in Figure 2-50B?

43.   What is capacitive reactance? _____

_____

44. Write the formula for capacitive reactance.

   _____

45. What is the capacitive reactance of a 20 μF capacitor, when the applied voltage has a frequency of 500 Hz?

46. In a capacitor, the _____ **leads** _____ by 90°.

47. A series circuit has 50 ohms resistance and 40 ohms capacitive reactance. What is the impedance?

Self-Review Answers

38. Capacitance is a measure of the amount of charge a capacitor can store for a given amount of applied voltage.

39. The unit of capacitance is the farad.

40. The three factors which determine capacitance are:

    1. Area of the plates;

    2. Spacing between the plates;

    3. Nature of the dielectric.

41. $C_T = \dfrac{C_1 \times C_2}{C_1 + C_2}$

    $C_T = \dfrac{40\ \mu F \times 60\ \mu F}{40\ \mu F + 60\ \mu F}$

    $C_T = \dfrac{2.4 \times 10^{-9}}{1 \times 10^{-4}}$

    $C_T = 24\ \mu F$

42. $C_T = C_1 + C_2 + C_3$

    $C_T = 4\ \mu F + 7\ \mu F + 9\ \mu F$

    $C_T = 20\ \mu F$

43. Capacitive reactance is the opposition to the flow of alternating current offered by a capacitor.

44. $X_C = \dfrac{1}{2\pi \, fC}$

45. $X_C = \dfrac{1}{2\pi \, fC}$

  $X_C = \dfrac{1}{6.28 \times 500 \text{ Hz} \times 20 \ \mu F}$

  $X_C = \dfrac{1}{6.28 \times 5 \times 10^2 \text{ Hz} \times 2 \times 10^{-5} \text{ F}}$

  $X_C = \dfrac{1}{6.28 \times 10^{-2}}$

  $X_C = 15.92 \ \Omega$

46. In a capacitor **current** leads **voltage** by 90°.

47. $Z = \sqrt{R^2 + X_C^2}$

  $Z = \sqrt{(50 \ \Omega)^2 + (40 \ \Omega)^2}$

  $Z = \sqrt{2.5 \times 10^3 + 1.6 \times 10^3}$

  $Z = \sqrt{4.1 \times 10^3}$

  $Z = 64 \ \Omega$

RLC CIRCUITS

RLC circuits contain resistance (R), inductance (L), and capacitance (C). In this section, we will be discussing series RLC and parallel RLC circuits. We will also be discussing filters and the important concept of resonance.

Series RLC Circuits

The circuit shown in Figure 2-51A is a series RLC circuit. We can use the phasor diagram in Figure 2-51B to help us visualize what is happening in the circuit and to find the total opposition to current flow or the impedance. Note that in the phasor diagram of Figure 2-51B that the X_L and X_C phasors are a total of 180 degrees out of phase. Whenever two components are 180 degrees out of phase, they cancel one another. In this case, X_L is larger than X_C; this means that X_C will only partially cancel X_L. The result in our example will be $200\,\Omega - 120\,\Omega = 80\,\Omega$. So we can reduce the phasor diagram to that shown in Figure 2-51C. This series RLC circuit is actually equivalent to an RL circuit of $80\,\Omega$ inductive reactance and $60\,\Omega$ resistance. We can now find the impedance using the formula:

$$Z = \sqrt{R^2 + X_L^2}$$

$$Z = \sqrt{(60\,\Omega)^2 + (80\,\Omega)^2}$$

$$Z = \sqrt{3.6 \times 10^3 + 6.4 \times 10^3}$$

$$Z = \sqrt{1 \times 10^4}$$

$$Z = 100\,\Omega$$
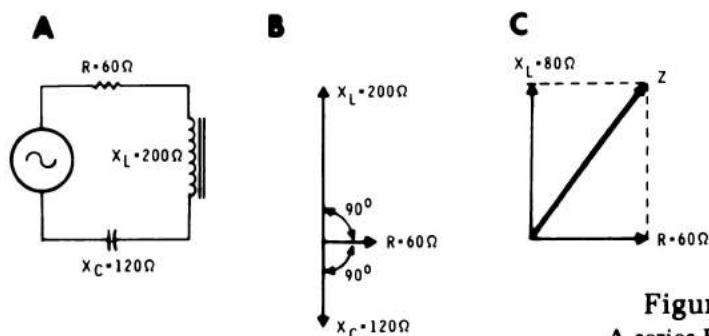
Figure 2-51
A series RLC circuit.

This series circuit acts inductively since the inductive reactance is larger than the capacitive reactance.
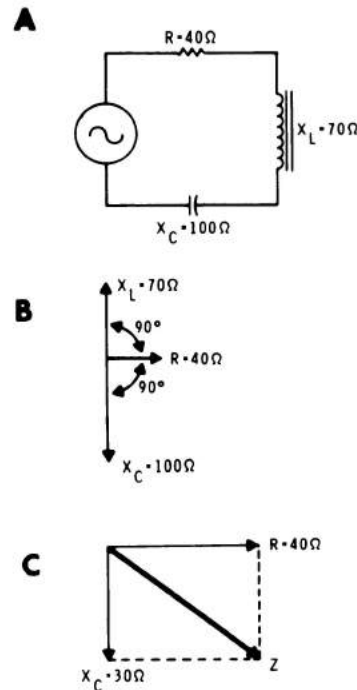
Figure 2-52
A series RLC circuit.

We will now try another example, where the capacitive reactance is larger than the inductive reactance. The circuit is shown in Figure 2-52A. The phasor diagram of the circuit is shown in Figure 2-52B. Note that X_C is greater than X_L; this means that the circuit will act capacitively. The phasor diagram in B will reduce to that shown in C by subtracting X_L from X_C, since they are 180 degrees out of phase from one another. We can now use a variation on the equation for finding impedance of a series AC circuit. This is:

$$Z = \sqrt{R^2 + X^2}$$

Where   Z   = impedance of the series AC circuit

R   = resistance

X   = effective reactance of the circuit

X is equal to $X_L - X_C$ if X_L is larger than X_C

X is equal to $X_C - X_L$ if X_C is larger than X_L

Let's find the impedance of the circuit of Figure 2-52.

$$Z = \sqrt{R^2 + X^2}$$

$$Z = \sqrt{(40\ \Omega)^2 + (30\ \Omega)^2}$$

$$Z = \sqrt{1.6 \times 10^3 + 9 \times 10^2}$$

$$Z = \sqrt{2.5 \times 10^3}$$

$$Z = 50\ \Omega$$

Parallel RLC Circuits

Analyzing parallel RLC circuits requires a different approach. In a series circuit, the current is the same at all points and is used as the reference phasor. In parallel circuits, the current is different in each branch. However, the voltage is the same across all branches. Therefore, voltage is used as the reference phasor.

A

B
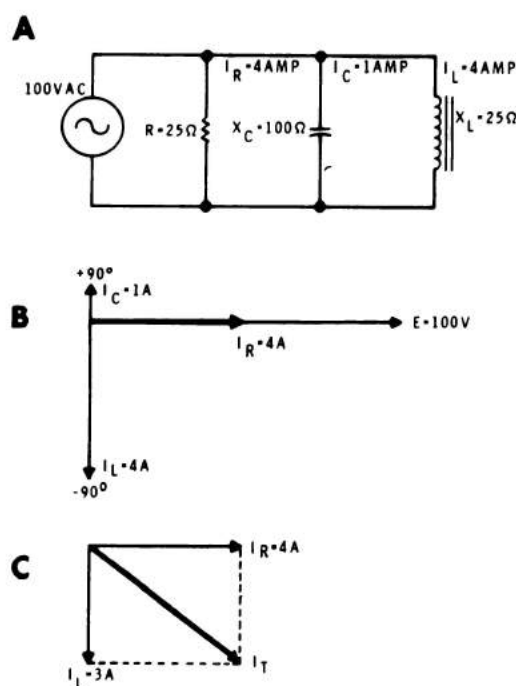
Figure 2-53
A parallel RLC circuit.

C

Figure 2-53A shows a parallel RLC circuit. The phasor diagram for this circuit is shown in Figure 2-53B. Notice that the voltage phasor (E) is drawn on the 0° line. This is our reference phasor. The I_R phasor is also drawn on the 0° line because current through a resistor is always in phase with the voltage. The I_C phasor is drawn at the +90° position because current leads voltage in a capacitor. The I_L phasor is drawn at the −90° position, because current lags voltage in an inductor. Note that this is just the opposite to the phasor diagram of a series circuit. The difference is the reference phasor. In a series circuit, current is the reference. In a parallel circuit, voltage is the reference.

To find the impedance of a parallel RLC circuit, you must first find the total circuit current. To accomplish this, the individual branch currents are found using Ohm's Law.

$$I_R = \frac{E}{R} = \frac{100 \text{ V}}{25 \text{ } \Omega} = 4 \text{ A}$$

$$I_C = \frac{E}{X_C} = \frac{100 \text{ V}}{100 \text{ } \Omega} = 1 \text{ A}$$

$$I_L = \frac{E}{X_L} = \frac{100 \text{ V}}{25 \text{ } \Omega} = 4 \text{ A}$$

The phasor diagram is then drawn as shown in Figure 2-53B. Since I_C and I_L are 180° out of phase, they subtract from one another. In this example, I_L is larger than I_C so the reactive current is $I_L - I_C$. Therefore, the phasor diagram can be reduced to that shown in Figure 2-53C.

Total current in a parallel RLC circuit is found using the following formula:

$$I_T \quad = \sqrt{I_R^2 + I_X^2}$$

Where I_T = total current in a parallel RLC circuit,

I_R = resistive current

I_X = total reactive current

I_X is equal to $I_L - I_C$ if I_L is larger than I_C

I_X is equal to $I_C - I_L$ if I_C is larger than I_L

Let's find the total current in the example problem of Figure 2-53.

$$I_T \quad = \sqrt{I_R^2 + I_X^2}$$

$$I_T \quad = \sqrt{(4A)^2 + (3A)^2}$$

$$I_T \quad = \sqrt{16 + 9}$$

$$I_T \quad = \sqrt{25}$$

$$I_T \quad = \quad 5 \text{ A}$$

Now that we know total current, we can find impedance using Ohm's Law:

$$Z \quad = \frac{E}{I_T}$$

$$Z \quad = \frac{100 \text{ V}}{5 \text{ A}}$$

$$Z \quad = \quad 20 \, \Omega$$

Series Resonance

What happens in a circuit when the inductive reactance is equal to the capacitive reactance? When this occurs, it is called **resonance**. A series resonant circuit has some véry interesting properties which we will be discussing in this section.
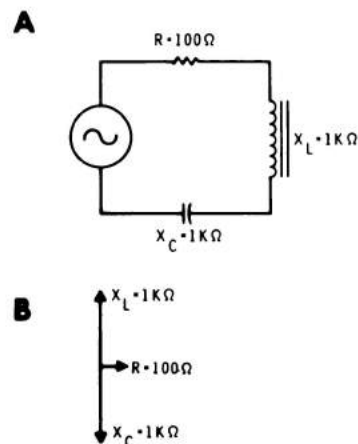
Figure 2-54
A series resonant circuit.

In the series RLC circuit of Figure 2-54, X_L is equal to X_C. This can only occur at one frequency of applied voltage. That frequency is called the **resonant frequency**. The resonant frequency can be calculated for any combination of L and C using the equation:

$$f_r = \frac{1}{2\pi \sqrt{LC}}$$

Where f_r = resonant frequency in Hz

2π = 6.28

L = inductance in henries

C = capacitance in farads

The phasor diagram for the circuit of Figure 2-54 shows that X_L and X_C equal each other, are 180 degrees out of phase, and therefore, cancel each other exactly. This means that the only opposition to current flow in the circuit at resonance is the resistance.

In the circuit of Figure 2-54, the resistance is 100 ohms and the applied voltage is 100 volts. This means that the circuit current is 1 ampere. What are the voltage drops across X_L and X_C? These can be found using Ohm's Law:

$$E_C = IX_C = 1 \text{ A} \times 1 \text{ k}\Omega = 1000 \text{ V}$$

$$E_L = IX_L = 1 \text{ A} \times 1 \text{ k}\Omega = 1000 \text{ V}$$

The voltage drops across both the inductor and the capacitor are 1,000 volts. This would seem to be contrary to the laws of series circuits which state that the sum of the voltage drops equal the supply voltage. However, in a series resonant circuit, the voltage drops across the reactances cancel one another because they are 180 degrees out of phase. In series resonant circuits, the voltage across the reactances can actually be higher than the applied voltage. Also note that the lower the series resistance, the higher the current flow. If we could reduce the series resistance to zero, the current would be infinitely high. This cannot happen, however, as the wire in the inductor and the wire in the circuit all have some resistance.

The important things to remember about a series resonant circuit are:

1. *The impedance of the circuit is low and equal to the series resistance. In an ideal circuit, the impedance would be zero.*

2. *The current flow is high and is limited only by the series resistance. In an ideal circuit, the current would be infinitely high.*

A

B
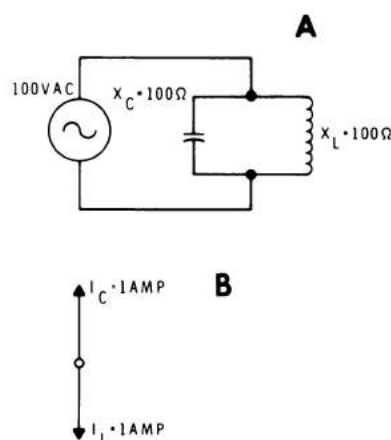
Figure 2-55
A parallel resonant circuit.

Parallel Resonance

The circuit of Figure 2-55A is a parallel resonant circuit. Note that both the inductive and capacitive reactances are equal. The current in each branch can be found using Ohm's Law.

$$I_C = \frac{E}{X_C} = \frac{100\ V}{100\ \Omega} = 1\ A$$

$$I_L = \frac{E}{X_L} = \frac{100\ V}{100\ \Omega} = 1\ A$$

The phasor diagram for the parallel resonant circuit is shown in Figure 2-55B. Since I_C is equal to I_L, and they are 180° out of phase, they cancel. The total current is zero. In a practical circuit, this would not happen, but in an ideal circuit, the current at resonance would be zero. Since current is zero, the impedance would be infinite.

The important things to remember about a parallel resonant circuit are:

1. *The impedance of the circuit is extremely high. In an ideal circuit, the impedance is infinite.*

2. *The current is very low. In an ideal circuit, the current flow is zero.*

Circuit Q

One of the most important characteristics of a resonant circuit is its **Q**. Circuit Q is a figure of **quality** for a resonant circuit. The Q for a series resonant circuit is defined as the ratio of the reactance at resonance to the series AC resistance. That is:

$$Q = \frac{X_L}{R}$$

In a resonant circuit in which $X_L = X_C = 1000$ ohms and the series resistance is 100 ohms, the Q factor is 10 because:

$$Q = \frac{X_L}{R} = \frac{1000\ \Omega}{100\ \Omega} = 10$$

Since Q is the ratio of reactance to resistance, the ohms in each term cancel. Thus, Q is simply a number without any associated units.

The Q of a resonant circuit affects the **bandwidth** of the circuit. A resonant circuit is selective in that it responds best to its resonant frequency. There is also a small band of frequencies, centered on the resonant frequency, to which the resonant circuit also responds. The width of this band of frequencies is called the bandwidth of the resonant circuit.
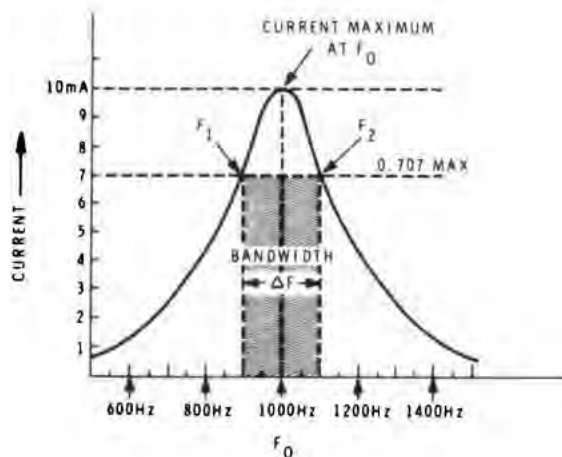
Figure 2-56
Bandwidth is measured
between half-power points.

Figure 2-56 is a graph of a series resonant circuit's current versus frequency. Note that the circuit current is at maximum at the resonant frequency of 1,000 Hz. The current then drops off on either side of this. The bandwidth includes that group of frequencies with a response of 70.7% of maximum. The bandwidth shown in Figure 2-56 is 70.7% of maximum or .707 × 10 mA = 7.07 mA. These points on the response curve are marked as F_1 and F_2. The bandwidth is $F_2 - F_1$ or $1100 - 900 = 200$ Hz.

The 70.7% points were chosen to indicate the bandwidth, because this is a very convenient point to use, since it represents the point at which the power in the circuit is exactly one-half the maximum value. Thus, the points marked F_1 and F_2 in Figure 2-56 are referred to as half-power points.

An example demonstrates that the power in a circuit drops to one-half when the current drops to 70.7%. Consider a circuit in which the resistance is 2,000 ohms and the current is 10 mA. The power is $P = I^2 \times R = 0.2$ W. Now, let's assume that the current drops to 70.7% of maximum or to 7.07 mA. The power drops to: $P = I^2 \times R = 0.1$ W. This is one half the previous power. Thus, reducing the current to 70.7% drops the power to 50%. For convenience then, the bandwidth is measured between the half-power points.

With the bandwidth measured between the half-power points, an interesting relationship exists between the bandwidth, the resonant frequency, and the value of Q. This relationship is expressed by the equation:

$$\text{BW} = \frac{f_r}{Q}$$

where   BW = bandwidth in Hz

f_r = resonant frequency in Hz

Q   = circuit Q

In the example of Figure 2-56, we can find the circuit Q by rearranging the formula.

$$BW = \frac{f_r}{Q}$$

$$Q = \frac{f_r}{BW}$$

$$Q = \frac{1000\ Hz}{200\ Hz}$$

$$Q = 5$$
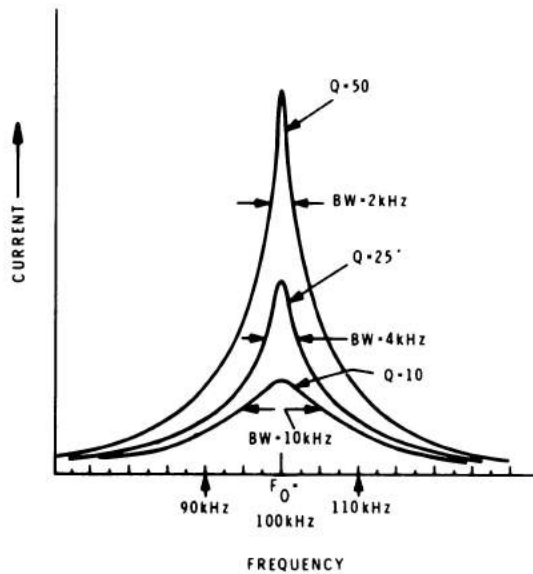
Figure 2-57
Bandwidth increases as Q decreases.

The curves shown in Figure 2-57 show that the bandwidth increases as the value of Q decreases. When the value of Q is high, the current in the circuit is relatively high and the resonant circuit responds to a very narrow band of frequencies.

The resonant frequency is 100 kHz. Thus, if the Q is 50, the bandwidth is:

$$BW = \frac{f_r}{Q}$$

$$BW = \frac{100 \text{ kHz}}{50}$$

$$BW = 2 \text{ kHz}$$

Notice what happens to the curve when the Q is reduced to 25. The current is lower and the curve is somewhat broader. The bandwidth is increased to:

$$BW = \frac{f_r}{Q}$$

$$BW = \frac{100 \text{ kHz}}{25}$$

$$BW = 4 \text{ kHz}$$

When the Q is reduced to 10, an even broader response curve results. The current is relatively low and the bandwidth is:

$$BW = \frac{f_r}{Q}$$

$$BW = \frac{100 \text{ kHz}}{10}$$

$$BW = 10 \text{ kHz}$$

Figure 2-58 shows three circuits which produce curves like those shown in Figure 2-57. The three circuits are identical except for the value of the resistance. The value of R determines the Q of the circuit. This, in turn, determines the bandwidth. Notice that the value of R does not affect the resonant frequency, only the Q bandwidth. As the value of R increases, the Q decreases and the bandwidth increases.

A

100 kHz $\sim$

$X_L = 2000\,\Omega$

$X_C = 2000\,\Omega$

$R = 40\,\Omega$

$Q = \dfrac{X_L}{R} = \dfrac{2000\,\Omega}{40\,\Omega} = 50$

B

100 kHz $\sim$

$X_L = 2000\,\Omega$

$X_C = 2000\,\Omega$

$R = 80\,\Omega$

$Q = \dfrac{X_L}{R} = \dfrac{2000\,\Omega}{80\,\Omega} = 25$

C

100 kHz $\sim$

$X_L = 2000\,\Omega$

$X_C = 2000\,\Omega$

$R = 200\,\Omega$

$Q = \dfrac{X_L}{R} = \dfrac{2000\,\Omega}{200\,\Omega} = 10$
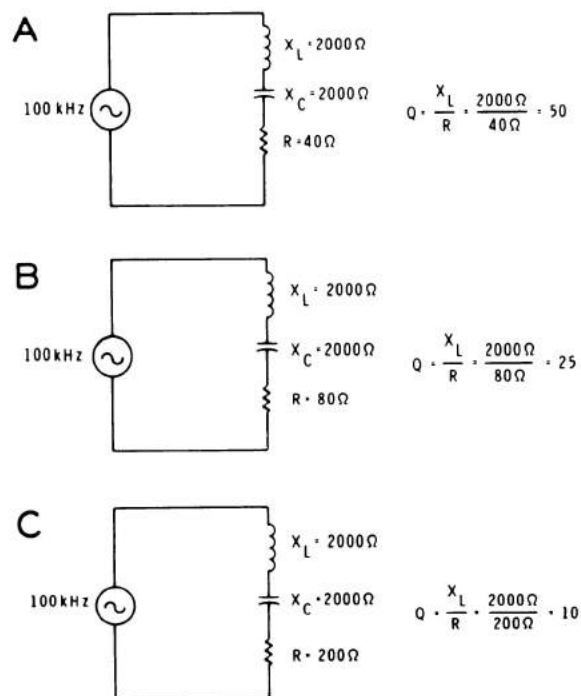
Figure 2-58
Effects of resistance.

You can calculate the Q of a **parallel** resonant circuit by using the equation:

$$Q = \frac{X_L}{R}$$

Where R = the total AC resistance.

Like the series resonant circuit, the parallel resonant circuit responds to a band of frequencies rather than a single frequency. Figure 2-59 shows a typical response curve. Notice that this curve has the same shape as the response curve shown earlier for the series resonant circuit. However, on closer examination, you will see that Figure 2-59 shows the **impedance** of the circuit rather than the current.
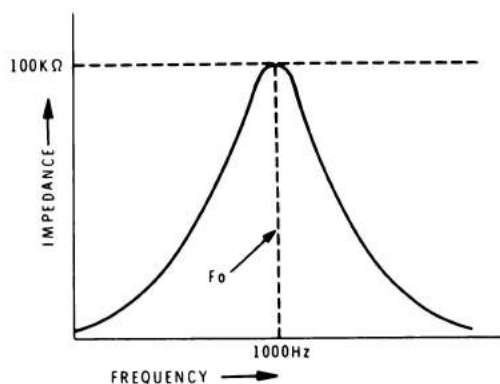
Figure 2-59
Response curve of a parallel resonant circuit (impedance vs. frequency).

At the resonant frequency, the impedance is maximum. Below resonance, the coil offers a low reactance and the impedance falls off. Above resonance, the capacitor offers a low reactance and the impedance again falls off.

Because the source current is inversely proportional to the impedance, the current response curve has the shape shown in Figure 2-60. Note that the current decreases as the resonant frequency is approached.
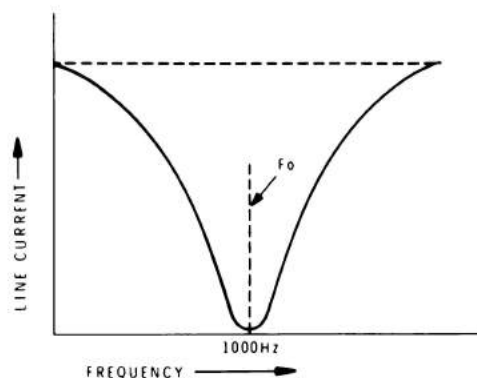
Figure 2-60
Response curve of a parallel resonant circuit (current vs. frequency).

As with the series resonant circuit, the bandwidth is determined by the Q of the circuit. The bandwidth is determined by the equation:

$$BW = \frac{fr}{Q}$$

If a parallel resonant circuit has a resonant frequency of 1000 Hz and a Q of 20, the bandwidth will be:

$$BW = \frac{f_r}{Q}$$

$$BW = \frac{1000 \text{ Hz}}{50}$$

$$BW = 50 \text{ Hz}$$

Often a parallel resonant circuit will be more selective than we would like. That is, it responds only to a very narrow band of frequencies. In these cases, we can increase the bandwidth by connecting a relatively small value resistor across the circuit as shown in Figure 2-61A. This will decrease the impedance of the circuit, or increase the source current, lowering the circuit Q. If we lower the circuit Q, we will increase the bandwidth.

Figure 2-61B shows the effects of adding a parallel resistor to the parallel resonant circuit. The lower the value of the resistor, the lower the circuit Q and therefore, the wider the bandwidth. This is often called "loading the resonant circuit".
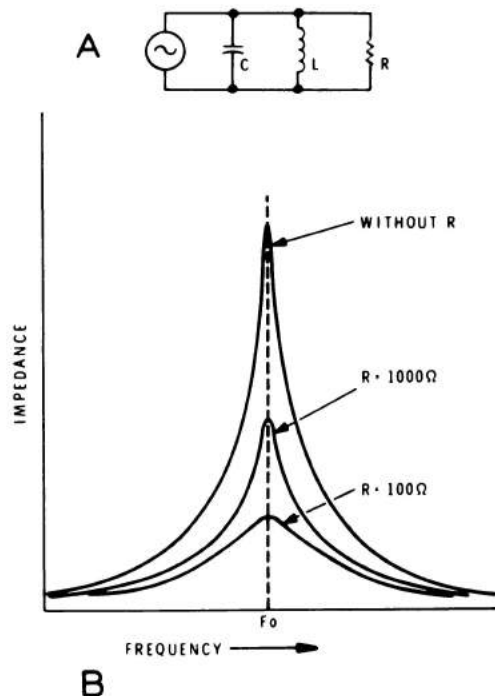
Figure 2-61
You can increase bandwidth by loading the resonant circuit.

Filters

A number of circuits in radio communications use filters. Filters are a combination of capacitors, inductors, and resistors which either pass a certain band of frequencies or stop a band of frequencies.

Bandstop Filter An example of a bandstop filter is shown in Figure 2-62A. A bandstop filter is designed to stop a certain band of frequencies while allowing all others to pass to the output.

A graph showing output voltage versus frequency is shown in Figure 2-62B. Note the very low output voltage at f_0. This is due to both Z_1 and Z_2 resonating at f_0. Z_1 offers infinite impedance at f_0, because it is in parallel resonance, while Z_2 offers zero impedance at f_0 because it is in series resonance. Since voltage drops are proportional to impedance, the majority of input voltage will be dropped across Z_1, while almost no voltage will be dropped across Z_2. Since the output voltage is taken across Z_2, very little voltage will appear at the output at f_0. At all other frequencies, much less attenuation will occur.
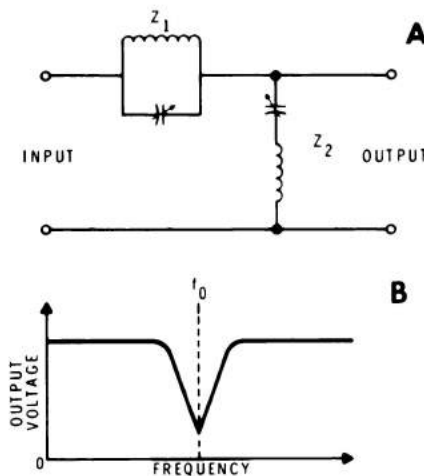
Figure 2-62
A bandstop filter.

Bandpass Filter   An example of a bandpass filter is shown in Figure 2-63A. A bandpass filter is designed to pass a certain band of frequencies and to stop all other frequencies.

The graph of output voltage versus frequency is shown in Figure 2-63B. Note that this curve is just the opposite to that of the bandstop filter, because the positions of Z_1 and Z_2 have been reversed. Now the output is taken across Z_1. At the resonant frequency (f_0), Z_1 offers infinite impedance, while Z_2 offers zero impedance. This allows the input voltage to pass to the output with very little opposition at f_0. However, at all other frequencies, Z_2 and Z_1 will oppose the input voltage.
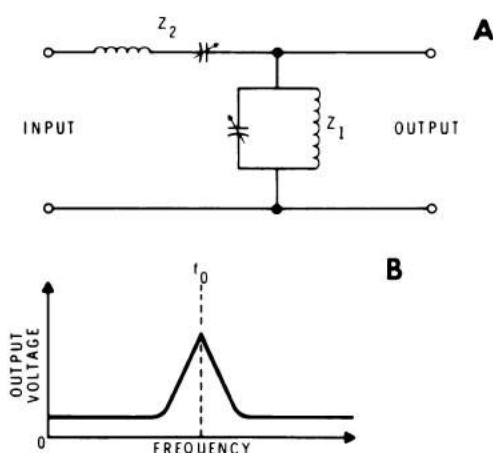
Figure 2-63
A bandpass filter.

Low Pass Filter   An example of a low pass filter is shown in Figure 2-64A. A low pass filter is designed to pass all frequencies below the cutoff frequency (f_{co}), and attenuate or greatly reduce all frequencies above f_{co}.

The output voltage versus frequency graph is shown in Figure 2-64B. Note the rapid decrease in output voltage above f_{co}. This is due to the increasing inductive reactance of L_1, and the decreasing capacitive reactance of C_1 as the frequency increases. Since the output voltage is taken across C_1, the output voltage will decrease as the frequency increases.
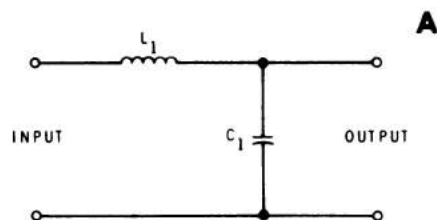
Figure 2-64
A lowpass filter.
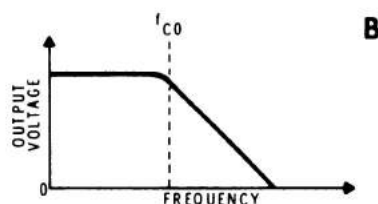
High Pass Filter An example of a high pass filter is shown in Figure 2-65A. A high pass filter is designed to pass all frequencies above f_{co}, and attenuate all frequencies below f_{co}.

The output voltage graph is shown in Figure 2-65B. Note that this curve is just the opposite of the low pass curve because the output voltage is now taken across L_1. The inductive reactance of L_1 increases and the capacitive reactance of C_1 decreases as the frequency increases. This allows the high frequencies to pass, but attenuates the low frequencies.
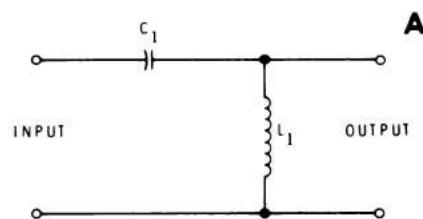
Figure 2-65
A highpass filter.
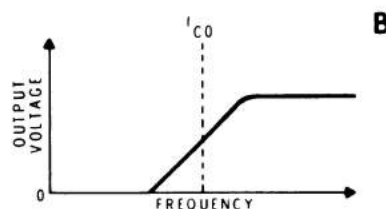
Self-Review Questions

48. Find the impedance of the circuit shown in Figure 2-66A.
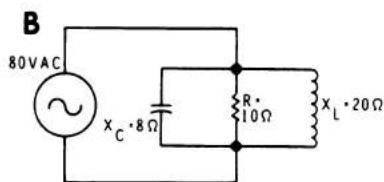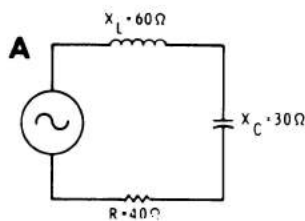
Figure 2-66
Find the impedance of these RLC circuits.

49. Find the impedance of the circuit shown in Figure 2-66B.

50. What is resonance? _____

_____

51. Write the formula for resonance.

52. State the characteristics of a series resonant circuit.

_____

_____

53. State the characteristics of a parallel resonant circuit.

_____

_____

54. What is circuit Q? _____

_____

55. How does circuit Q affect the bandwidth of a resonant circuit?

_____

_____

56. List four types of filters and state the purpose of each one.

    1. _____

       _____

       _____

    2. _____

       _____

       _____

    3. _____

       _____

       _____

    4. _____

       _____

       _____

Self-Review Answers

48. The effective reactance is $X_L - X_C = 60 \ \Omega - 30 \ \Omega = 30 \ \Omega$.

$$Z = \sqrt{R^2 + X^2}$$

$$Z = \sqrt{(40 \ \Omega)^2 + (30 \ \Omega)^2}$$

$$Z = \sqrt{1600 + 900}$$

$$Z = \sqrt{2500}$$

$$Z = 50 \ \Omega$$

49. $$I_C = \frac{E}{X_C} = \frac{80 \ V}{8 \ \Omega} = 10 \ A$$

$$I_R = \frac{E}{R} = \frac{80 \ V}{10 \ \Omega} = 8 \ A$$

$$I_L = \frac{E}{X_L} = \frac{80 \ V}{20 \ \Omega} = 4 \ A$$

$$I_L - I_C = 10 \ A - 4 \ A = 6 \ A$$

$$I_T = \sqrt{I_R^2 + I_X^2}$$

$$I_T = \sqrt{(8 \ A)^2 + (6 \ A)^2}$$

$$I_T = \sqrt{64 + 36}$$

$$I_T = \sqrt{100}$$

$$I_T = 10 \ A$$

$$Z = \frac{E}{I_T}$$

$$Z = \frac{80 \ V}{10 \ A}$$

$$Z = 8 \ \Omega$$

50. Resonance occurs when $X_L = X_C$. This occurs at only one frequency for a given combination of L and C.

51. The formula for resonance is:

$$f_r = \frac{1}{2\pi \sqrt{LC}}$$

52. A series resonant circuit has a very low impedance and very high current flow.

53. A parallel resonant circuit has a very high impedance and very low current flow.

54. Circuit Q is a figure of quality for a resonant circuit. It is equal to

$$\frac{X_L}{R}.$$

55. The higher the circuit Q, the narrower the bandwidth becomes. The lower the circuit Q, the wider the bandwidth becomes.

56. (1) A bandstop filter stops a certain band of frequencies while passing all others.

(2) A bandpass filter passes a certain band of frequencies while stopping all others.

(3) A low pass filter passes all frequencies below the cutoff frequency and stops all frequencies above the cutoff frequency.

(4) A high pass filter passes all frequencies above the cutoff frequency and stops all frequencies below the cutoff frequency.
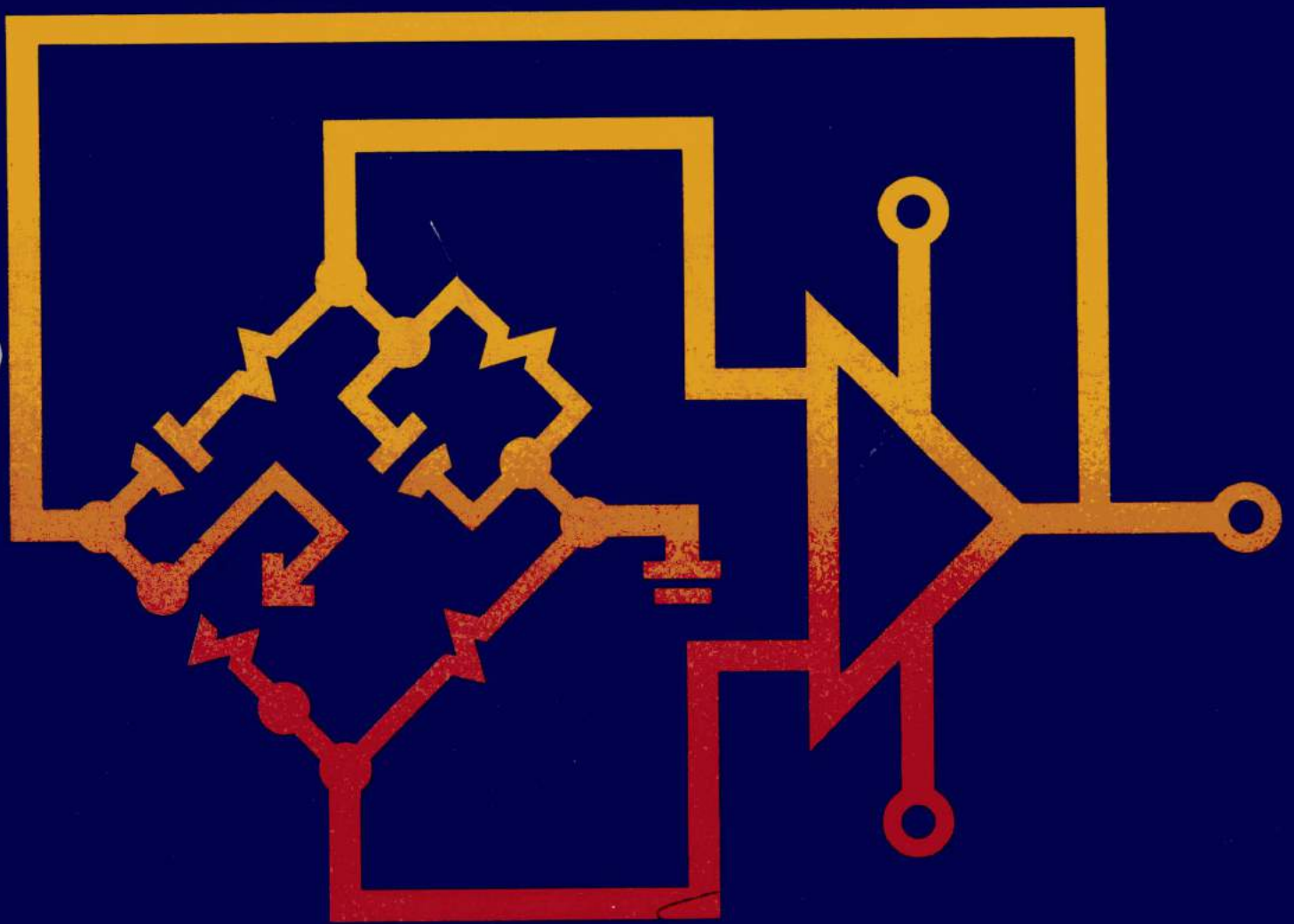
CONCEPTS OF ELECTRONICS

Heathkit

Zenith

Educational Systems

CONCEPTS OF ELECTRONICS

Book 2

Unit 3

ACTIVE DEVICES

CONTENTS

INTRODUCTION

Basically, there are two types of components used in electronic circuits, passive and active devices. **Passive** devices do not alter their resistance, impedance or reactance when constant AC signals are applied to them. Examples of passive devices are resistors, inductors, and capacitors. **Active** devices, however, change their resistance or impedance when varying voltages are applied to them. As a result of this, active devices can amplify and rectify AC signals. Examples of active devices are vacuum tubes, transistors, and diodes.

Let's take a brief look at the development of active devices. In the 1920's, the vacuum tube revolutionized the field of electronics, allowing the invention of many devices including the radio receiver. By 1950, solid-state diodes and transistors, which are smaller and more rugged, started to replace vacuum tubes. The early 1960's saw the construction of groups of transistors and diodes on a single chip of silicon, forming an integrated circuit. The result of this development has been more compact and reliable electronic circuits and equipment. You will be studying several of these solid-state devices in this unit.

The unit objectives listed on the next page state exactly what you are expected to learn from this unit. Study this list now and refer to it often as you study the text.

UNIT OBJECTIVES

When you complete this unit, you should be able to:

1. State the difference between P-type and N-type semiconductor materials.

2. Define depletion region.

3. Identify both forward and reverse-biased diodes.

4. Name the two most important solid-state diode ratings and define each one.

5. State the characteristics of zener and varactor diodes.

6. Identify the schematic symbols of a solid-state diode, a zener diode, and a varactor diode.

7. Identify the schematic symbols for NPN and PNP transistors.

8. Name the three sections of a transistor and identify them on a transistor schematic symbol.

9. State the correct bias for the emitter-base and collector-base junctions of a transistor.

10. Identify and state the characteristics of the common-emitter, common-base, and common-collector amplifier configurations.

11. Define beta, thermal runaway, and maximum power dissipation as applicable to transistors.

12. Identify the schematic symbols of a JFET, a depletion mode MOSFET, and an enhancement mode MOSFET and name the terminals.

13. State the difference between depletion and enhancement mode MOSFETs.

14. Name the three basic FET circuit configurations and state the characteristics of each circuit.

15. Define light, infrared rays, ultraviolet rays and photon.

16. State the light spectrum's frequency range.

17. Name four light sensitive devices, state the characteristics of each, and identify their schematic symbols.

18. State the basic operating principles of the light emitting diode (LED) and identify its schematic symbol.

19. Determine the necessary value of bias resistor for correct LED operation.

20. State the operating characteristics and modes of the liquid crystal display.

21. Define integrated circuit and list its advantages and disadvantages.

22. Name the two basic types of integrated circuits.

23. State the basic types and characteristics of digital ICs.

24. State the basic characteristics of linear ICs and the operational amplifier.

25. Find amplifier voltage gain when given input and output voltage.

SILICON
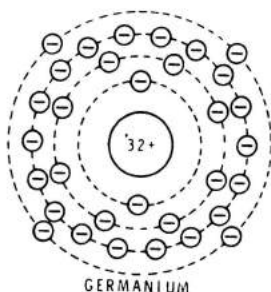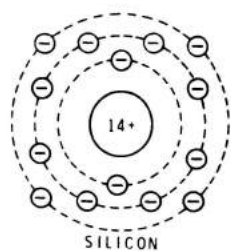
GERMANIUM

Figure 3-1

The atomic structure of silicon and germanium.

SOLID-STATE DIODES

Solid-state or semiconductor diodes have been in use since the earliest days of radio. However, recent advances have improved the semiconductor diode, and specialized diodes such as the zener diode and variable capacitance diode have been developed. In this section, you will learn the basic principles of semiconductors. You will also study the solid-state diode.

Semiconductor Materials

In an earlier unit, a semiconductor was described as being neither a good conductor nor a good insulator. This is due to the atomic structure or more importantly, the valence or outer shell electrons. Figure 3-1 shows the atomic structure of the two semiconductors used in most solid-state devices: silicon and germanium. Note that both have 4 valence electrons. Compare this to the best conductors, which have one valence electron, or to the best insulators which have a full valence shell of eight electrons.

Silicon and germanium, in their pure form, are actually worse conductors than is indicated by their four valence electrons. This is due to their crystalline structure. A simplified diagram of the germanium crystal structure is shown in Figure 3-2. Note that only the valence electrons are shown. In the crystal structure, each of the four valence electrons of any one atom is shared with four neighboring atoms. Thus, each atom appears to have eight electrons in its valence shell. This fills each atom's valence shell and it becomes very difficult to free an electron for current flow.
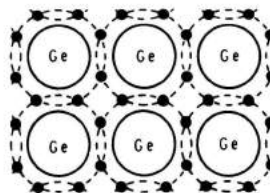
Figure 3-2
Simplified diagram of germanium crystal structure.

By adding carefully controlled quantities of certain impurities to the pure semiconductor, its conductivity can be increased. This is called **doping**. An example of an impurity atom is arsenic, which has 5 valence electrons. When it is added to a semiconductor crystal, only 4 of its 5 valence electrons can be fitted into the crystal structure. The additional fifth electron becomes free to act as a current carrier. This is shown in Figure 3-3. A crystal doped in this way is known as **N-Type** semiconductor material, since it contains additional electrons.

Another impurity atom used to dope semiconductors is gallium. However, it has only three electrons. Therefore, when it is added to a semiconductor crystal a deficiency of electrons occurs. As shown in Figure 3-4, there now exists an area or **hole** in the crystal structure that lacks an electron. These holes behave as positively charged particles which are free to drift throughout the crystal. Due to the presence of the holes, the doped material is known as a **P-Type** semiconductor.

While a hole is not actually positive, it is an area that a randomly drifting electron can "fall" into, thus completing the crystal structure. However, once an electron "falls" into a hole, another hole is created in the region from which the electron came. The movement of a hole in this way is equivalent to the movement of a positive charge equal to the negative charge carried by one electron.
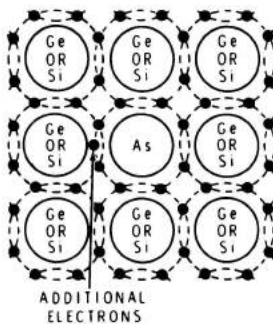
Figure 3-3
Semiconductor material doped with arsenic (N-Type).
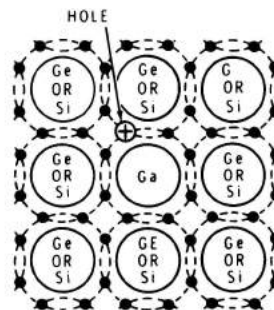
Figure 3-4
Semiconductor material doped with gallium (P-Type).

P-N Junctions

Figure 3-5
Solid-state diode operation.

Figure 3-6
Schematic symbol for a solid-state diode.
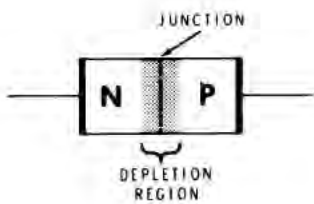
When N and P type semiconductor materials are grown together to form a single crystal, a **solid-state diode** results. This is shown in Figure 3-5A. The area where the P and N type materials join is called the junction. As soon as the junction is formed, there will be a movement of electrons across it. Electrons near the junction will move from the N-type material into the P-type and fill the holes. As a result of this, the N-type material near the junction is depleted of electrons, while the P-type material near the junction is depleted of holes. Therefore, this area is called the **depletion region**.
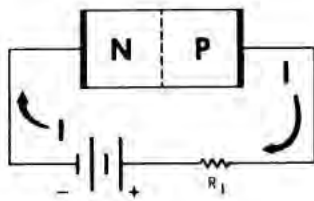
Figure 3-5B shows what happens when a battery is connected to the PN junction. In this case, a negative voltage is applied to the N-type material, while a positive voltage is applied to the P-type. In this condition, once the voltage is above 1/2 to 1 volt, the applied voltage forces the electrons to cross the depletion region and continue across the P-type material. In other words, current flows through the diode and the depletion region no longer exists. This condition is called **forward bias**. Note that resistor R₁ must be added to limit the current. This is because the voltage drop across the junction is very low.
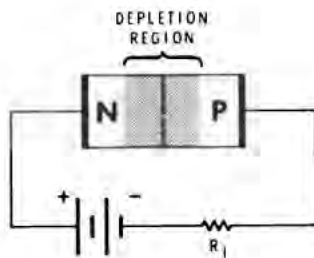
Figure 3-5C shows what happens when the battery connections are reversed. There is no current flow and the depletion region becomes larger. This is because the negative voltage on the P-type material forces electrons into the holes, thus depleting the region of still more holes. The positive voltage on the N-type material attracts the free electrons, which depletes the region of more electrons. This is called **reverse-bias**.

The PN-junction acts as a one way switch because it conducts only when it is forward biased. It is called a **solid-state diode**. And, since it has a very low forward voltage drop, it makes an excellent rectifier. You will recall that a rectifier is used to convert alternating current into pulsating direct current.

The schematic symbol for the solid-state diode is shown in Figure 3-6. The N-type material is called the **cathode**, while the P-type is the **anode**. The arrowhead in the schematic symbol points in the direction **opposite** to current flow. Note that when the solid-state diode is forward biased, current flows from cathode to anode.
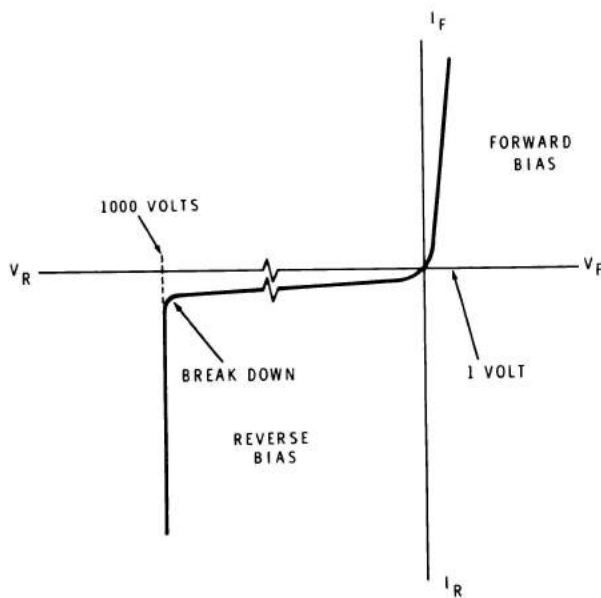
Figure 3-7
Diode characteristic curve.

Diode Ratings

Figure 3-7 shows the "characteristic curve" for a solid state diode. This curve shows diode current under forward and reverse bias conditions. In the forward bias condition, it shows that the forward voltage drop (V_F) is a relatively constant 1 volt regardless of the forward current (I_F).

An important diode rating is the maximum forward current. If this rating is exceeded, the diode will be destroyed due to excessive heat. A typical maximum forward current rating is 1 A, however, some diodes may be rated as high as 20 A or as low as 10 mA.

The curve of Figure 3-7 also shows reverse bias characteristics. Note the small amount of reverse current (I_R). Since this current is normally a few microamperes, it can usually be ignored. However, when the reverse voltage (V_R) reaches 1,000 V, the reverse current drastically increases. This is due to a **breakdown** of the depletion region. The electrons are "ripped" from the valence shell and a high current results. Normally, this will damage the diode; however, some diodes are specially designed to operate in breakdown. They are called zener diodes and are discussed in the next section.

Since a breakdown of the diode during reverse bias is possible, diodes have a **peak inverse voltage** (PIV) rating. The PIV is the maximum reverse bias voltage the diode can handle before breakdown. Typical PIV ratings are 1,000 to 2,000 V. However, diodes can be rated as high as 7,000 V or as low as 30 V.
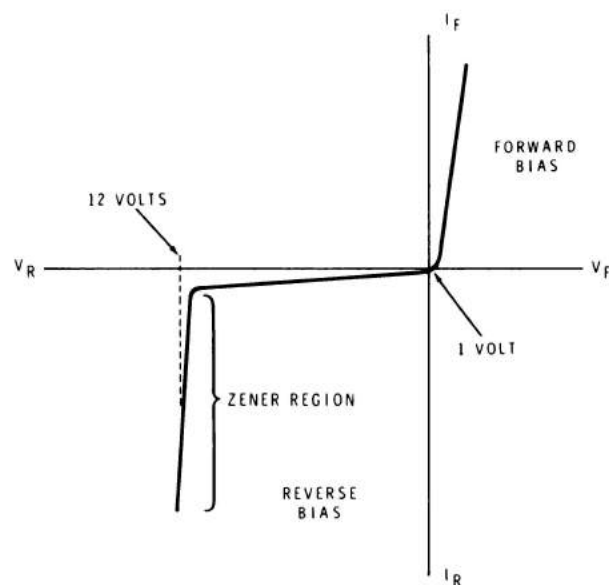
Figure 3-8
Zener diode characteristic curve.
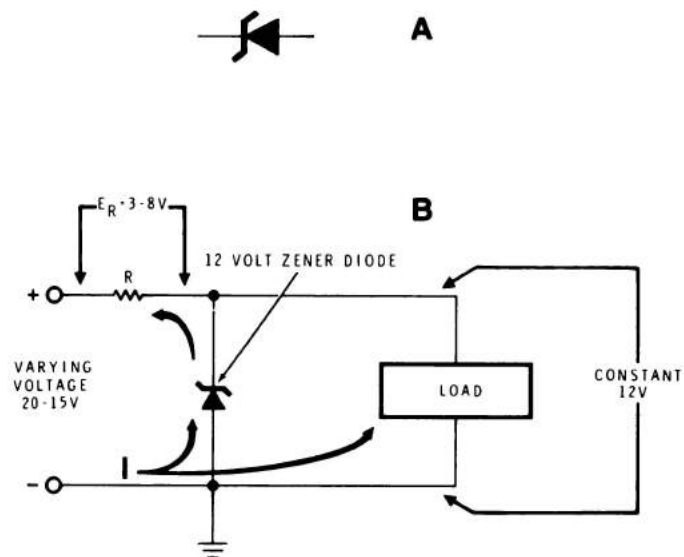
Figure 3-9
The zener diode's (A) schematic symbol, and
(B) voltage regulator circuit.

Zener Diode

Figure 3-8 shows the characteristic curve of a typical zener diode. Notice that just after breakdown, in this case at 12 V, the voltage (V_R) is constant for any value of reverse current (I_R). This is called the zener region, and once the diode is biased into this region, its voltage drop is constant. While all diodes have this characteristic, zener diodes are specially designed to dissipate the heat generated when operated in breakdown. They are also designed to break down at standard voltage values. The zener diode's schematic symbol is shown in Figure 3-9A.

The constant voltage drop across a zener diode can be put to good use by placing a load in parallel with it. Figure 3-9B shows a circuit which takes a varying voltage of 15-20 V and, using a zener diode, presents a constant 12 V to the load. This circuit is known as a voltage regulator because it supplies a constant voltage to the load, regardless of input variations. Voltage regulators are used to supply voltage to circuits which are sensitive to voltage variations.

Varactor Diode

Another special purpose diode is the varactor or variable capacitance diode. This diode uses the variable width of the depletion region in reverse bias. Figure 3-10A shows that, for varying voltages, the width of the depletion region varies proportionately. Since the depletion region acts as an insulator, it becomes the dielectric and the P and N type material become the plates of a capacitor. Now, by varying the reverse bias voltage and the plate spacing, the diode capacitance is also varied. While all diodes exhibit some junction capacitance, varactors are designed to enhance this property.

Applications of the varactor include using it as part of a resonant circuit. Then the resonant frequency can be changed by varying the diode's voltage. The schematic symbol for a varactor is shown in Figure 3-10B.
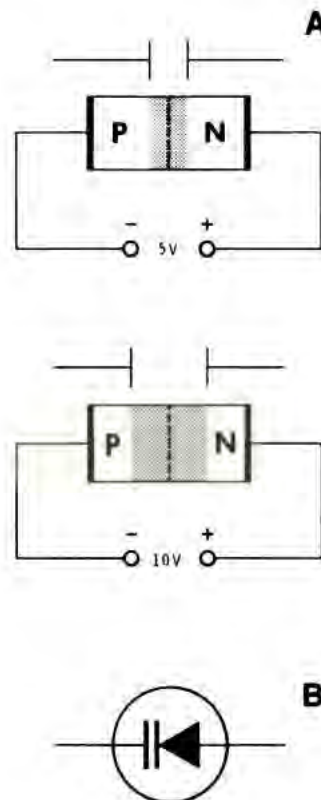
Figure 3-10
Varactor diode (A) operation,
(B) schematic symbol.

Self-Review Questions

1. A doped semiconductor material which contains additional electrons is called a/an ⎯⎯ ⎯⎯⎯⎯ semiconductor. A material which has holes or a deficiency of electrons is called a/an ⎯⎯ semiconductor.

2. In a solid-state diode, the area near the PN junction which has a deficiency of electrons and holes is called the ⎯⎯⎯⎯⎯ ⎯⎯⎯⎯ .

3. In Figure 3-11A, the diode is ⎯⎯⎯⎯⎯ biased, while in
    <u>forward/reverse</u>
    Figure 3-11B the diode is ⎯⎯⎯⎯⎯ biased.
    <u>forward/reverse</u>

4. Name the two most important solid-state diode ratings and define each one. ⎯⎯⎯

5. Identify the schematic symbols shown in Figure 3-12.

    A. ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯

    B. ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯

    C. ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯

6. A zener diode is operated in ⎯⎯⎯⎯⎯ bias and has a
    <u>forward/reverse</u>
    ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯ .

7. A varactor diode is operated in ⎯⎯⎯⎯⎯ bias and has a
    <u>forward/reverse</u>
    voltage variable ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯ .

Figure 3-11
Are these diodes forward or reverse biased?
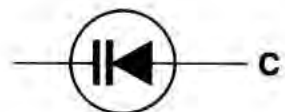
Figure 3-12
Identify these schematic symbols.

Self-Review Answers

1. A doped semiconductor material which contains additional electrons is called an **N-type** semiconductor. A material which has a deficiency of electrons or holes is called a **P-type** semiconductor.

2. In a solid-state diode, the area near the PN junction which has a deficiency of electrons and holes is called the **depletion region.**

3. In Figure 3-11A, the diode is **forward** biased, while in Figure 3-11B the diode is **reverse** biased.

4. **PIV** or **peak inverse voltage** is the maximum amount of reverse bias voltage a diode can withstand before breakdown occurs.

   Maximum Forward Current is the maximum forward bias current that a diode can handle.

5. The schematic symbols shown in Figure 3-12 are:

   A. **Solid-state diode**

   B. **Zener diode**

   C. **Varactor diode**

6. A zener diode is operated in **reverse** bias and has a **constant voltage drop**.

7. A varactor diode is operated in **reverse** bias and has a voltage variable **capacitance**.

TRANSISTORS

An extension of the solid-state diode is the transistor. However, the transistor uses two PN-junctions, rather than one, and can actually **amplify** input signals. Amplification occurs when the output signal is greater than the input signal. Therefore, the transistor has many more applications and is much more versatile than the diode.

In this section, you will be learning basic transistor operation. You will also be introduced to the three common transistor amplifier configurations.

Transistor Structure

A transistor is a single semiconductor crystal that contains two PN junctions. It is constructed so that a P-type section is "sandwiched" between two N-type sections or an N-type section is "sandwiched" between two P-type sections. Thus, there are two transistors; NPN and PNP. Figure 3-13 shows both types of transistor structure and their schematic symbols.
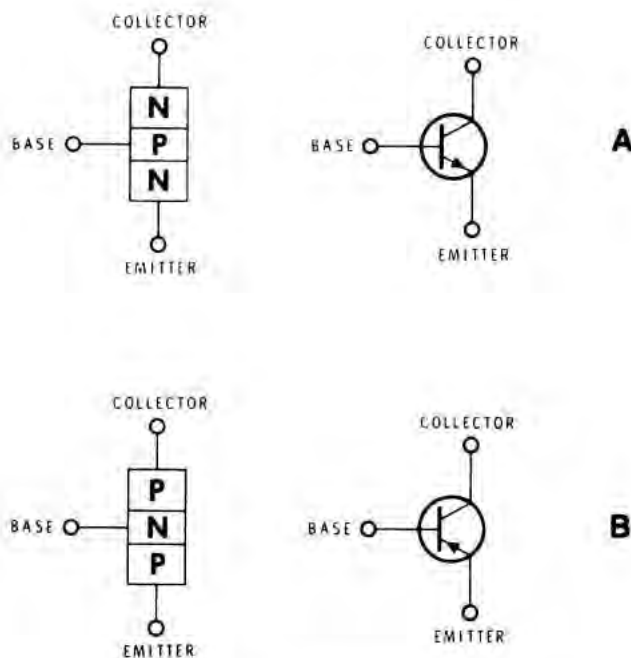
Figure 3-13
Transistor structure and schematic symbols.
(A) NPN transistor.
(B) PNP transistor.

In both NPN and PNP transistors, the three sections are called the **emitter**, the **base**, and the **collector**. In the schematic symbol, the emitter is identified by an arrowhead. Just as in the diode, the arrowhead points in the direction **opposite** to electron flow. It also points toward the N-type material. This will help you distinguish between an NPN and PNP schematic symbol.

Transistor Operation

The operation of a transistor junction is similar to the operation of a diode junction. However, the transistor is used to amplify, while the diode can only rectify. The diode is normally used as a one-way switch. Therefore, it is changing between forward and reverse bias. In a transistor, however, the emitter-base junction is forward biased and the collector-base junction is reverse biased. Figure 3-14 shows correctly biased PNP and NPN transistors.
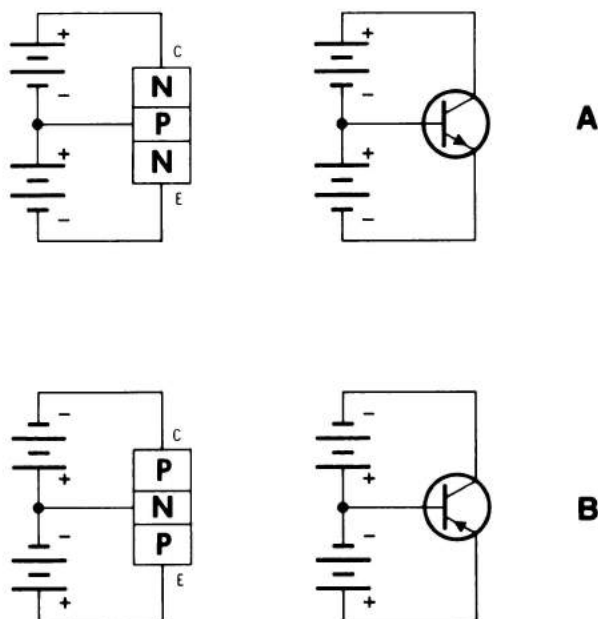
Figure 3-14
Correctly biased (A) NPN transistor, and
(B) PNP transistor.
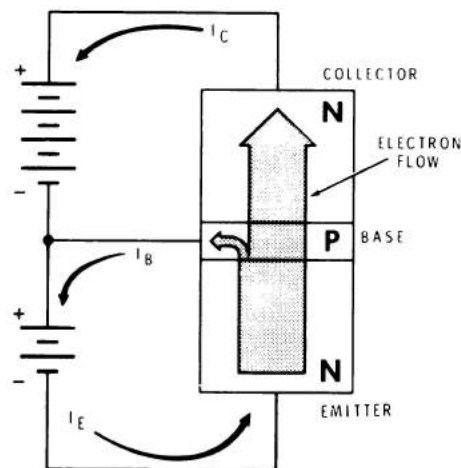
Figure 3-15
Electron flow in an NPN transistor.

Figure 3-15 shows the electron flow in a correctly biased NPN transistor. A large number of electrons enter the emitter N-region due to the forward biased emitter-base junction. Since the base region is designed to be very thin, the high positive potential on the collector causes most electrons to pass straight through the base. Of course, some will remain in the base, giving a small base current.

An important relationship exists between emitter current (I_E), base current (I_B), and collector current (I_C). It can be stated mathematically as follows:

$$I_E = I_B + I_C$$

The equation simply states that the emitter current (I_E) is equal to the sum of the base current (I_B) and the collector current (I_C). Another way of looking at this is that the collector current is equal to the emitter current less the current lost to the base.

$$I_C = I_E - I_B$$

If the base region of the transistor was not extremely thin, the action just described could not occur. The thin base region makes it possible for the emitter injected electrons to move quickly into the collector region. A large base region would minimize the interaction between the emitter and collector regions and the transistor would act more like two separately biased diodes.

It is interesting to note that a small change in I_B causes a large change in I_C. For example if I_B is decreased, fewer electrons cross the emitter-base junction. Therefore, fewer electrons cross the collector-base junction, decreasing I_C. On the other hand if I_B is increased, more electrons cross the emitter-base junction, which in turn increases I_C. Thus, I_B controls I_C. To further dramatize this point, if I_B is reduced to zero, I_C also drops to zero because there is no electron flow from emitter to base. This condition is known as **cutoff**, since there is no transistor current flow. Likewise, I_B can be increased to a point where no more electrons are available from the emitter and, therefore, I_C cannot increase. This condition is called **saturation** and is usually avoided. However, when a transistor is used as a switch, such as in digital circuits, it generally alternates between cutoff and saturation depending on the desired state of the switch.

While the preceding discussion dealt with the NPN transistor, the PNP is very similar. However, the current within a PNP transistor is in terms of "hole" flow, as shown in Figure 3-16. This is due to the P-type emitter and collector whose current or charge "carriers" are holes. Note also that, for correct bias, the batteries must be reversed. This means that the external current will be in the opposite direction to an NPN transistor.
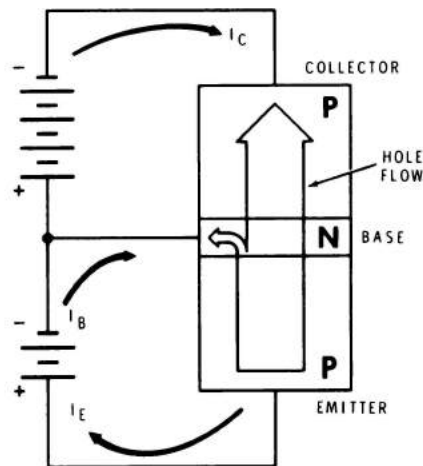
Figure 3-16
Hole flow in a PNP transistor.

Transistor Amplification

Now that you have seen basic transistor operation, the next step is transistor amplification. We will continue the discussion of the NPN transistor of Figure 3-15, however, we will use the transistor's schematic symbol as shown in Figure 3-17. Notice that the voltage sources used to forward bias the emitter junction and reverse bias the collector junction are labeled V_{EE} and V_{CC} respectively. These two symbols are standard designations for the voltages used in transistor circuits. Furthermore, V_{EE} is variable so that the forward bias voltage can be varied. The transistor is designated Q, since the Q designation is widely used to represent transistors.
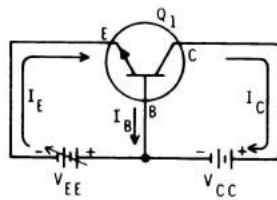
Figure 3-17
Schematic diagram of a properly biased NPN transistor.
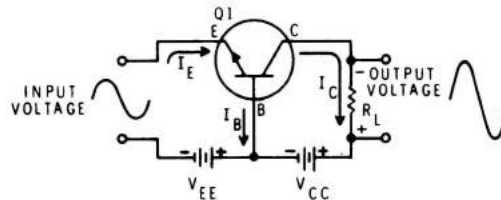
Figure 3-18
A basic NPN transistor amplifier circuit.

In order for the transistor shown in Figure 3-17 to provide amplification, the device must be capable of accepting an input signal (current or voltage) and providing an output signal that is greater in strength or amplitude. The transistor cannot perform this function if it remains connected as shown in Figure 3-17. Instead, the circuit must be changed so that it appears as shown in Figure 3-18. Notice the resistor between the collector of transistor Q_1 and the positive terminal of V_{CC}. This component is commonly referred to as a load resistor and is labeled R_L. It is used to develop an output voltage with the polarity indicated. In other words, the collector current must flow through this resistor to produce a specific voltage drop. Note also the separation between the emitter of transistor Q_1 and the negative terminal of V_{EE}. These open connections serve as the input of the amplifier circuit and allow an input signal to be inserted between the emitter of Q_1 and V_{EE}. Therefore, the input is not an open circuit, but is completed when an external voltage source is connected between the emitter and V_{EE}. Furthermore, the values of V_{EE} and V_{CC} have been selected to bias transistor Q_1 so that the values of I_E, I_B and I_C are high enough to permit proper circuit operation.

The emitter junction of transistor Q_1 is forward-biased and like a PN junction diode, has a relatively low resistance. However, the collector junction is reverse-biased and therefore has a relatively high resistance. In spite of the tremendous difference in emitter junction resistance and collector junction resistance the emitter current (I_E) is almost equal to the collector current (I_C). The value of I_C is only slightly less than I_E because a small portion of I_E flows out of the base region to become base current (I_B). The load resistor can therefore have a high value without greatly restricting the value of I_C.

If a small voltage is applied to the input terminals which aids the forward bias voltage (V_{EE}), the value of V_{EE} is effectively increased. This will cause I_E to increase and I_C will also increase. This will increase the voltage drop across the load resistor (R_L). However, when the input voltage is reversed so that it opposes V_{EE}, the value of V_{EE} is effectively reduced and both I_E and I_C will decrease. This will cause the voltage drop across R_L to decrease. A change in input voltage can, therefore, produce a corresponding change in output voltage. However, the output voltage change will be much greater than the input change. This is because the output voltage is developed across a high output load resistance (R_L) and the input voltage is applied to the low resistance offered by the emitter junction. Thus, a very low input voltage can control the value of I_E which in turn determines the value of I_C. Even though I_C remains slightly less than I_E, it is forced to flow through a higher resistance and, therefore, produce a higher output voltage.

The action just described will occur if the input signal is either a DC or an AC voltage. The important point to remember is that any change in input voltage is greatly amplified by the circuit so that a much larger but proportional change in output voltage is obtained. It is also important to realize that the input signal is not simply increased in size, but is used to control the conduction of transistor Q_1 and this transistor in turn controls the current through the load resistor (R_L). The transistor is therefore made to take energy from the external power source V_{CC} and apply it to the load resistor in the form of an output voltage whose value is controlled by a small input voltage.

The transistor shown in Figure 3-18 is therefore being used to convert a relatively low voltage to a much higher voltage. Any circuit which performs this basic function is commonly referred to as a **voltage amplifier**. Such circuits are widely used in electronic equipment since it is often necessary to raise both DC and AC voltages to higher levels so that they may be effectively used. Later in this unit you will also discover that the transistor can be connected in a different manner which will allow it to amplify currents as well as voltages. Any circuit which is used for the sole purpose of converting a low current to a higher current is commonly referred to as a **current amplifier**.

Although an NPN transistor was used in the amplifier circuit shown in Figure 3-18, the PNP transistor will perform the same basic function. However; in the PNP circuit the bias voltages (V_{EE} and V_{CC}) must be reversed and the respective currents (I_E, I_B, and I_C) will flow in the opposite directions.

Transistor Circuit Arrangements

As explained previously, the transistor is used primarily as an amplifying device. However, there is more than one way of using a transistor to provide amplification and each method offers certain advantages or benefits but also has certain disadvantages or limitations. Basically, the transistor may be utilized in three different circuit arrangements to perform its important amplifying function. In each arrangement one of the transistor's three leads is used as a common reference point and the other two leads serve as input and output connections. The three circuit arrangements are referred to as the **common-base** circuit, the **common-emitter** circuit and the **common-collector** circuit, and each arrangement may be constructed by using either NPN or PNP transistors. In many cases the common leads are connected to circuit or chassis ground and the circuit arrangements are then often referred to as grounded-base, grounded ᵊmitter, and grounded-collector circuits. Furthermore, in each circuit arrangement the transistor's emitter junction is always forward-biased while the collector junction is reverse-biased.

The common-base circuit (as explained previously) provides slightly less output current than input current because I_C is slightly lower than I_E. In other words there is a slight current loss between the input and output terminals of the circuit. However, the circuit is still useful as a voltage amplifier as explained earlier. Since the input and output currents are almost the same and because the output voltage developed across the load can be much higher than the input signal voltage, the output power produced by the circuit is much higher than the input power applied to the circuit. The common-base circuit therefore provides power amplification as well as voltage amplification.

COMMON-EMITTER CIRCUITS

Transistors (NPN and PNP) may also be connected as shown in Figure 3-20. Notice that the input signal is applied between the base and emitter of each transistor while the output signal is developed between the collector and emitter of each transistor. Since the emitter of each transistor is common to the input and output of each circuit, the arrangements are referred to as common-emitter circuits. Notice that the emitter junction of each transistor is still forward-biased while the collector junction of each device is reverse-biased. Also, the NPN circuit shown in Figure 3-20A and the PNP circuit shown in Figure 3-20B require bias voltages that are opposite in polarity. Note that the forward bias voltage (designated as V_{BB}) is now applied to the base of the transistor instead of the emitter. In other words the forward bias voltage (V_{BB}) is applied so that it controls the base current (I_B) instead of the emitter current (I_E). Only the base current flows through bias voltage V_{BB} and the AC generator (signal source). The reverse bias voltage (V_{CC}) is applied through the output load to the collector and emitter regions of the transistor and only the collector current (I_C) flows through the load and V_{CC}.
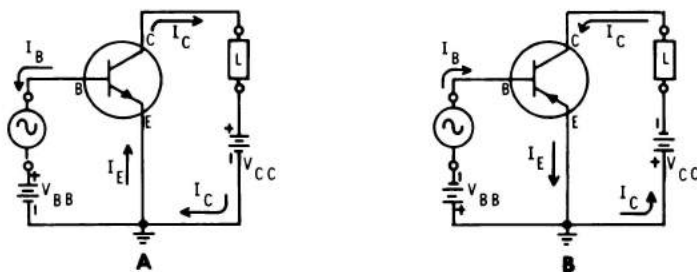
Figure 3-20
Common-emitter circuits.
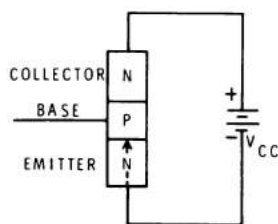
Figure 3-21
The reverse-biased collector
junction in an NPN
common-emitter circuit.

Although it may not be apparent, V_{CC} does actually reverse bias the collector-base junction of the transistor. Figure 3-21 shows how this reverse bias is obtained in an NPN common-emitter circuit. If V_{CC} is connected across the emitter and collector regions of the transistor as shown, the emitter-base junction becomes forward-biased and therefore has a low resistance. However, V_{CC} causes the N-type collector to be positive with respect to the P-type base and the collector-base junction is therefore reverse-biased. The forward-biased emitter-base junction effectively allows the negative terminal of V_{CC} to be applied to the P-type base region. V_{CC} is therefore effectively applied to the collector and base regions as shown. If V_{CC} was reversed, the collector junction would become forward-biased and the emitter junction would become reverse-biased; however, this condition would be undesirable. The same basic action also occurs in a PNP transistor, although the PNP device requires bias voltage polarities which are exactly opposite to those used with an NPN device.

The three currents (I_E, I_B, and I_C) in the common-emitter circuit still have the same relationship as before and still vary in a proportional manner. In other words I_E is still equal to $I_B + I_C$ even though I_B has now become the input current and I_C has become the output current. Any change in the input current (I_B) will therefore result in a proportional change in I_E and I_C. The difference between this circuit and the common-base circuit previously described is therefore simply one of reference. We have now connected the transistor so that its very low base current is being used to control its much larger collector current.

When an input signal voltage is applied to the common-emitter circuit, it either aids or opposes V_{BB} and therefore causes I_B to either increase or decrease. The much higher emitter and collector currents (which are almost equal) are forced to increase and decrease by the same percentage. Since a very low I_B is used to control a relatively high I_C, the transistor is now being used to provide current amplification as well as voltage amplification. This of course means that the power supplied to the load is much greater than the power applied to the input of the circuit. In fact the common-emitter arrangement provides the highest power output for a given power input when compared with the other two circuit arrangements. The common-emitter circuit is also the only circuit that provides current, voltage, and power amplification.

A transistor that is used in a common-emitter arrangement will also have a low input resistance but this resistance will not be as low as the input resistance of a common-base arrangement. This is because the input base current for the common-emitter circuit is much lower than the input emitter current of the common-base circuit. Also, the output resistance of a transistor connected in a common-emitter arrangement will be somewhat lower than when it is connected in a common-base circuit. A typical low power transistor might have an input resistance of 1000 or 2000 ohms and an output resistance of 50,000 or 60,000 ohms.

COMMON-COLLECTOR CIRCUITS

The NPN and PNP transistors shown in Figure 3-22 are connected so that the collector of each transistor is used as a common reference point instead of the emitter and the base and the emitter regions serve as the input and output connections. These circuits are known as common-collector circuits.

In the common-collector circuit the input signal voltage is applied between the base and collector regions of the transistor. The input either aids or opposes the transistor's forward bias and therefore causes I_B to vary accordingly. This in turn causes I_E and I_C to vary by the same percentage. An output voltage is developed across the load which is connected between the emitter and collector regions of the transistor. The emitter-current (I_E) flowing through the load is much greater than the base current (I_B) and the circuit therefore provides an increase in current between its input and output terminals. However, the voltage developed across the load will always be slightly lower than the voltage applied to the circuit. The slightly lower voltage will appear at the emitter of the transistor because the device tends to maintain a relatively constant voltage drop across its emitter-base junction. As explained earlier, this forward voltage drop may be equal to approximately 0.3 volts if the transistor is made of germanium or 0.7 volts if the transistor is made of silicon. The output voltage appearing at the emitter of the transistor therefore tends to track or follow the input voltage applied to the transistor's base. For this reason the common-collector circuit is often called an **emitter follower**.
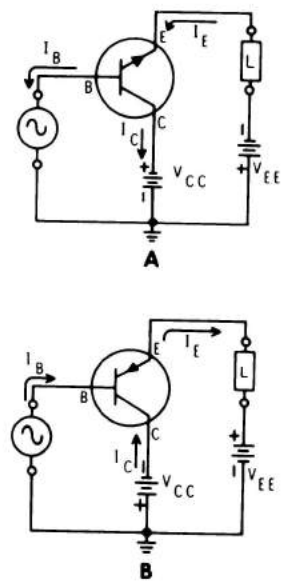
Figure 3-22
Common-collector circuits.

The common-collector circuit therefore functions as a current amplifier but does not produce an increase in voltage. However, the increase in output current (even though output voltage is slightly less than input voltage) results in a moderate increase in power. Also, the input resistance of any transistor connected in a common-collector arrangement is extremely high. This is because the input resistance is that resistance which appears across the reverse-biased collector-base junction. This input resistance can be as high as several hundred thousand ohms in a typical low power transistor. However, the output resistance appearing between the transistor's emitter and collector regions will be much lower (often as low as several hundred ohms) because of the relatively high emitter current (I_E) that flows through the output lead.

The common-collector circuit is not useful as an amplifying device. Instead, this circuit is widely used in applications where its high input resistance and low output resistance can perform a useful function. The circuit is often used to couple high impedance sources to low impedance loads and therefore perform the same basic function as an impedance matching transformer. The circuit also provides a moderate amount of power gain.

The bias voltages (V_{BB}, V_{EE}, and V_{CC}) used in the three circuit arrangements just described may appear to restrict or interfere with the flow of signal current through these circuits; however, this does not occur. The bias voltages are effectively shorted as far as the input and output signals are concerned, although in some applications large capacitors are placed across the voltage sources to insure that they offer minimum impedance to signal currents. For example, a large electrolytic capacitor could be placed across V_{CC} in Figure 3-22 to insure that the collector of the transistor is at ground potential as far as AC signals are concerned.

SUMMARY

As can be seen from the previous discussion, each transistor circuit configuration has specific characteristics. The chart shown in Figure 3-23 compares the characteristics of these three circuits. As you can tell, the common-emitter configuration is the most versatile: offering voltage, current, and power gain. It is also the most widely used configuration. However, the common-base and common-collector circuits have specific applications, such as voltage amplification and impedance matching, respectively.

	TYPICAL VALUES		
CHARACTERISTICS	**COMMON — EMITTER**	**COMMON — BASE**	**COMMON — COLLECTOR**
IMPUT IMPEDANCE	1000-2000 Ω	30-40 Ω	Up to 500 kΩ
OUTPUT IMPEDANCE	50-60 kΩ	Up to 1 MΩ	400 Ω
VOLTAGE GAIN	Up to 500	Up to 1,000	Less than 1
CURRENT GAIN	30	Less than 1	50
POWER GAIN	Up to 10,000	Up to 1,000	50

Figure 3-23
Transistor amplifier configuration characteristics.

Transistor Ratings

One of the more important transistor ratings is the forward current transfer ratio or beta (β). **Beta** is the transistor's current gain in the common-emitter configuration. It is the ratio of a small change in collector current to a small change in base current. When expressed as a formula, we have:

$$\beta = \frac{\Delta I_C}{\Delta I_B}$$

where
β = beta
ΔI_C = A small change in collector current
ΔI_B = A small change in base current

Typical values of β range from 20 to 200.

One disadvantage of transistors when compared to vacuum tubes is their sensitivity to temperature. When the semiconductor material is heated, more electrons are freed due to thermal agitation. This allows more current flow in the transistor. The additional current creates additional heat! If this cycle continues, the transistor will be in **thermal runaway** and will soon be destroyed. External bias circuitry must be designed to compensate for any heat changes to prevent thermal runaway.
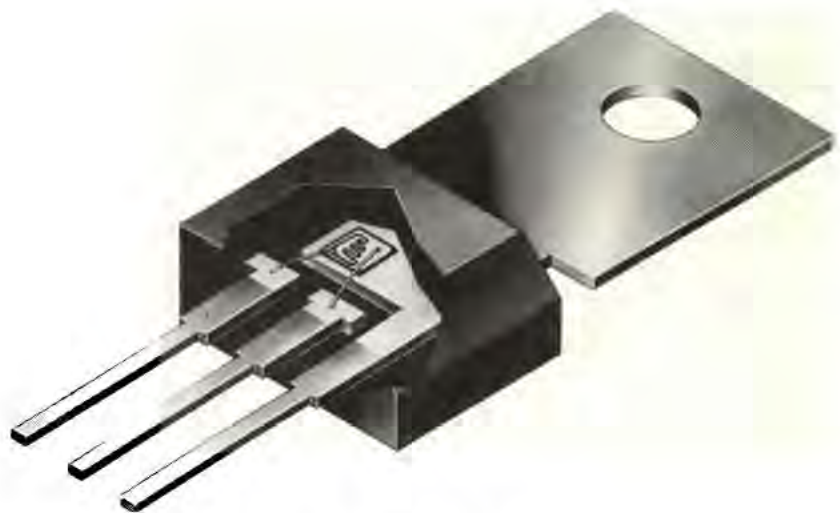
Figure 3-24
A cutaway view of a high power transistor.

A rating which is related to a transistor's thermal characteristic is its **maximum power dissipation**. This rating is determined by how much heat a transistor can safely dissipate. The input power to a transistor is determined by multiplying collector current by collector voltage. Its power dissipation (P_{dis}) is then input power (P_{in}) minus output power (P_{out}) or

$$P_{dis} = P_{in} - P_{out}$$

Therefore, any input power which is not transferred to the output or load is dissipated as heat inside the transistor.

One way to increase a transistor's maximum power dissipation is to construct it so that it can radiate heat more effectively. A transistor capable of dissipating up to 30 watts is shown in Figure 3-24. It has a special tab which can be attached to an external heat sink. The tab "conducts" heat away from the transistor to the heat sink. The heat sink has many fins which radiate the heat into the air. Figure 3-25 shows the heat sink for a transistorized 100-watt amateur radio transceiver.

HEAT SINK

Figure 3-25
The heat sink on a 100 watt solid state
amateur radio transceiver.

Self-Review Questions

8. Identify the type of transistor and the transistor leads by filling in the blanks in Figure 3-26.

_____ TRANSISTOR   _____ TRANSISTOR
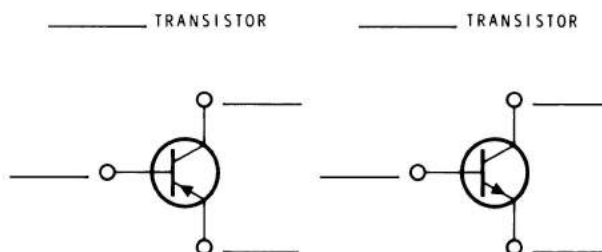
Figure 3-26
Fill in the blanks.

9. In a correctly biased transistor, the emitter-base junction is _____ biased, while the collector-base junction is
   forward/reverse
   _____ biased.
   forward/reverse

10. Name the amplifier configurations shown in Figure 3-27.

   A. _____

   B. _____

   C. _____

11. The common-collector circuit has _____ input impedance,
   low/high
   _____ output impedance, and high _____ gain.
   low/high

12. What is the primary application for the common-collector configuration? _____

   _____

13. The common-base circuit has _____ input impedance,
   low/high
   _____ output impedance, and high _____ gain.
   low/high

14. The common-_____ circuit provides the highest power amplification.

15. **Regardless** of the type of circuit **arrangement** used, the transistor's emitter junction must always be _____-biased.
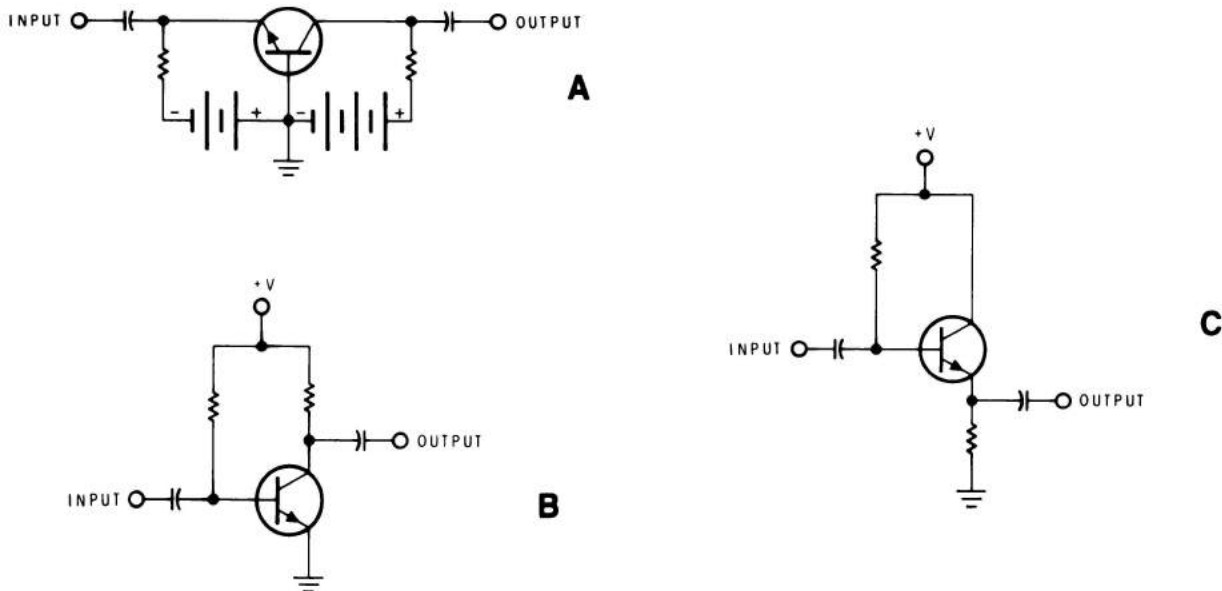
Figure 3-27
Identify these amplifier configurations.

16. The ratio of a small change in base current to a small change in collector current is called _____ .

17. What is thermal runaway? _____
_____
_____

18. A transistor's construction and its ability to effectively radiate heat determine its maximum _____ _____.

Self-Review Answers

8.   See Figure 3-28.

PNP TRANSISTOR   **NPN** TRANSISTOR
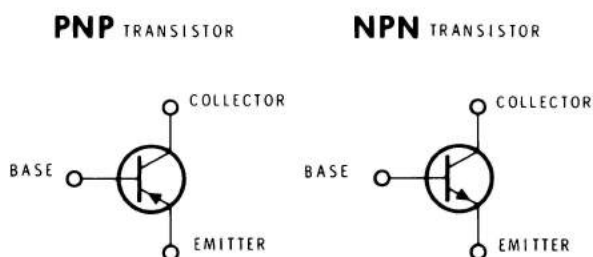
Figure 3-28

9.   In a correctly biased transistor, the emitter-base junction is **forward**-biased, while the collector-base junction is **reverse**-biased.

10. Even though you have not seen the circuits of Figure 3-27 before, you should be able to identify the configuration from the input and output connections. They are as follows:

   A.   common-base

   B.   common-emitter

   C.   common-collector.

11. The common collector circuit has **high** input impedance, **low** output impedance, and high **current** gain.

12. **Impedance matching** is the primary application for the common collector amplifier.

13. The common-base circuit has **low** input impedance, **high** output impedance, and high **voltage** gain.

14. The common-**emitter** circuit provides the highest power amplification.

15. Regardless of the type of circuit **arrangement** used, the transistor's emitter junction must always be **forward**-biased.

16. The ratio of a small change in base current to a small change in collector current is called beta or β.

17. Thermal runaway occurs when a transistor heats up, which causes an increase in current. The increased current causes additional heat. Eventually the transistor will be destroyed. Thermal runaway is prevented by using a bias network that compensates for heat changes.

18. A transistor's construction and its ability to effectively radiate heat determine its maximum **power dissipation**.

FIELD-EFFECT TRANSISTORS

The field-effect transistor (FET) is a three terminal solid-state device capable of amplification when used in the proper circuitry. Its three terminals are called the **source**, **drain**, and **gate**. These terminals roughly correspond to the emitter, collector, and base, respectively, of the junction transistor. However, in the FET the source to drain current is controlled by a **voltage** applied to the gate. Compare this to the junction transistor which is a current controlled device. Also, since no current flows into the gate, the input impedance is very high, ranging from 10^6 to 10^{14} Ω.

There are two basic types of field-effect transistors: the junction FET or JFET, and the insulated gate FET or MOSFET. You'll be studying both types in this section.

Junction FET

The basic construction of a junction FET is shown in Figure 3-29A. It consists of a block of N-type semiconductor to which the source and drain leads are attached as shown. This N-type material is called the **channel**, and the electrons flow through this material from source to drain. The gate is a slice of P-type semiconductor that forms a junction with the N-type channel.
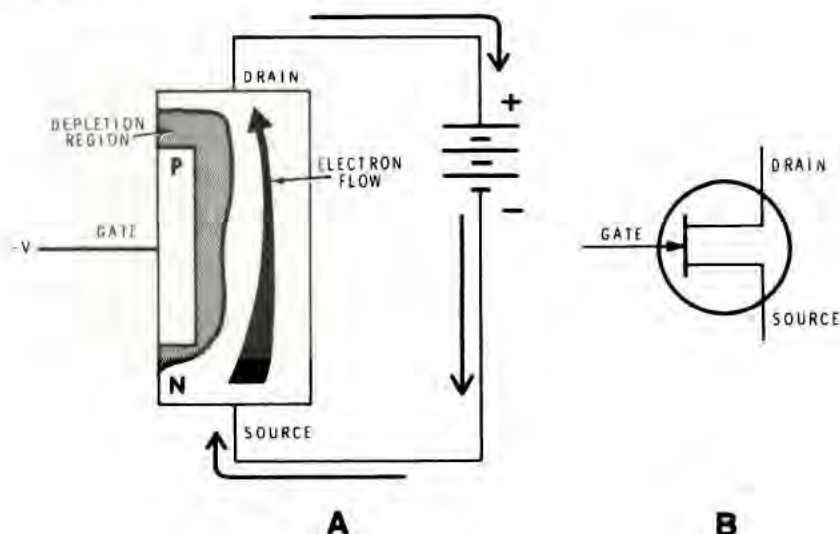
Figure 3-29
N-channel Junction FET, construction (A),
schematic symbol (B).

Under normal conditions the gate junction is reverse-biased. This is accomplished by applying a negative bias voltage to the gate. The resulting depletion region will extend across the channel and, since the depletion region is "depleted" of electrons, it will limit the current flow through the channel. Therefore, the channel current is controlled by varying the amount of reverse bias on the gate junction. In fact, the gate voltage can be increased until the depletion region covers the entire N-channel, thus, cutting-off all current flow. This condition is called **pinch-off**. Since the gate junction is always reverse-biased, the JFET's input impedance is very high.

The schematic symbol of the JFET is shown in Figure 3-29B. Note that the gate lead has an arrow. It points to the N-type material, just as in the junction transistor and diode. Figure 3-29 shows an N-channel JFET while Figure 3-30 shows a P-channel JFET. The only differences are the positive bias voltage required to reverse-bias the gate junction, and the positive charge carriers or holes in the channel. The operation is the same as the N-channel JFET. Note the schematic symbol in Figure 3-29B.
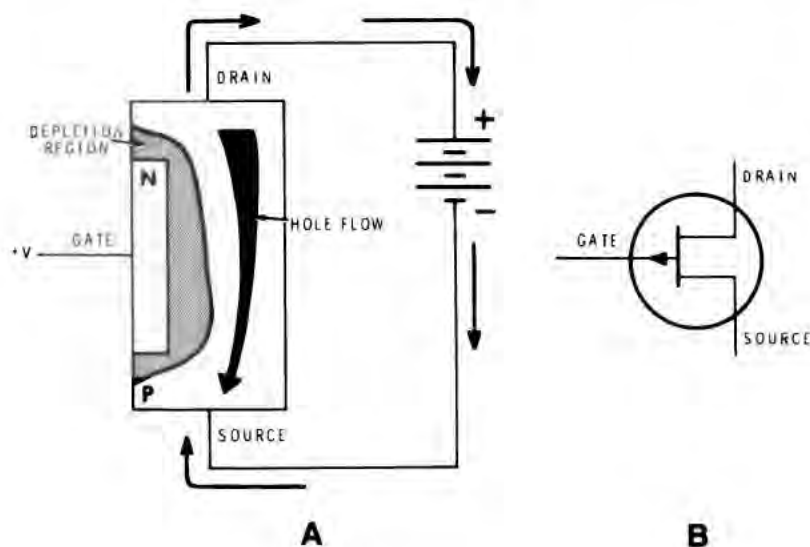
Figure 3-30
P-channel Junction FET, construction (A),
schematic symbol (B).

Insulated Gate FET

The construction of an insulated gate FET is shown in Figure 3-31A. In this type of FET the channel is an N-type semiconductor that is placed on a P-type **substrate**. The substrate is a block of semiconductor material upon which the rest of the FET is constructed. The gate is a strip of metal which is insulated from the channel by a layer of metal oxide. For this reason the insulated gate FET is also known as a metal oxide semiconductor FET or MOSFET. Since the gate is insulated from the channel the input impedance is extremely high, typically 10^{10} to 10^{14} Ω.
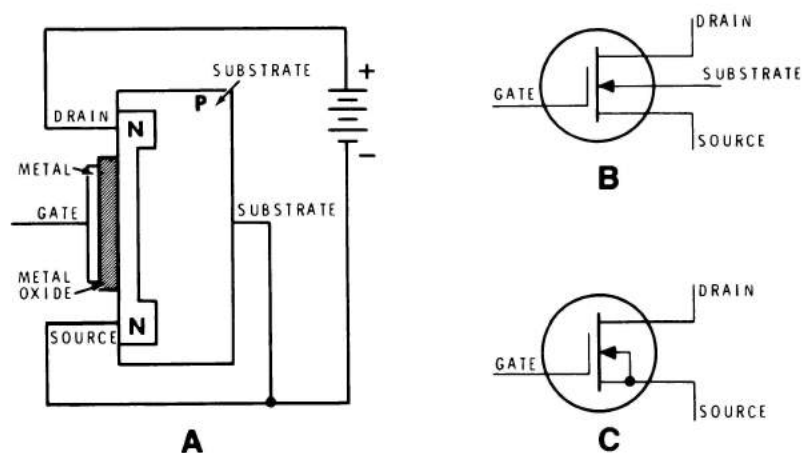
Figure 3-31
Insulated gate FET, construction (A),
schematic symbols (B and C).

During normal operation a negative voltage is applied to the source and a positive voltage to the drain. This establishes current flow through the channel from source to drain. If a negative voltage is applied to the gate, the negative charge on the gate will repel the electrons in the channel and a depletion region will form. The gate potential will control the size of the depletion region and, therefore, the channel current flow. This is very similar to the junction FET. However, when a positive voltage is applied

to the MOSFET gate, current flow in the channel is enhanced. That is, electrons are drawn toward the gate from the channel and the substrate. This increases channel conductivity and, thus current flow. This could never be done in a JFET because the gate junction would become forward biased, allowing gate current and drastically reducing input impedance. However, the MOSFET gate is insulated and its input impedance remains high regardless of gate voltage polarity. Thus, the MOSFET's source to drain current is controlled by a positive or negative voltage applied to the gate.

The insulated gate FET schematic symbol is shown in Figure 3-31B. Note that the gate is shown insulated from the channel. The substrate lead is identified by the arrow and, as always, the arrow points toward the N-type semiconductor. In this case the arrow identifies the device as an N-channel MOSFET. P-channel devices are also available and their operation is similar. In the device shown, the substrate has a separate lead, however, in many MOSFETs the substrate is connected to the source internally. This is shown in Figure 3-31C.

Another type of MOSFET is shown in Figure 3-32A. Note that the construction is similar to that just discussed except that it does not have a complete channel. The source and drain connections are there but a complete path for electron flow does not exist.
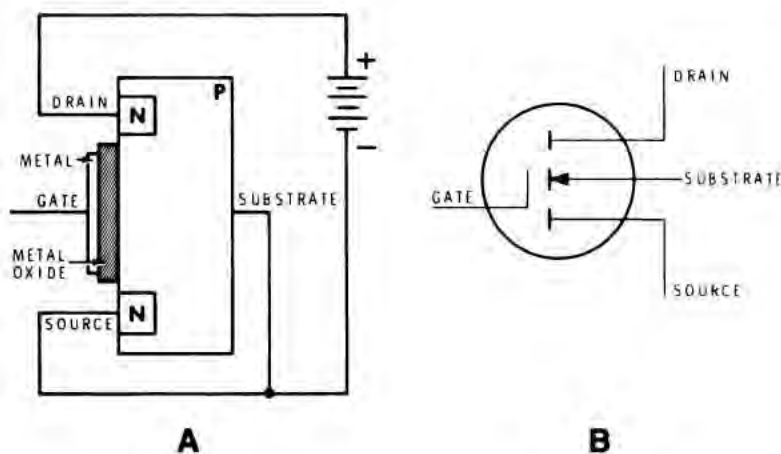
Figure 3-32
Enhancement mode MOSFET, construction (A),
schematic symbol (B).

Therefore, with no bias voltage applied to the gate, source to drain current is zero. However, when a positive voltage is applied to the gate, the substrate electrons are attracted toward it. This actually induces a conduction path or channel from the source to the drain. The amount of positive voltage on the gate determines the channel width and, therefore, the source to drain current. This type of device is known as an **enhancement mode** MOSFET since electron flow is "enhanced" by the application of a positive gate voltage. The previously discussed device is called a **depletion mode** MOSFET since electron flow is "depleted" by a negative gate voltage.

The schematic symbol for an enhancement mode MOSFET is shown in Figure 3-32B. Note the broken line indicating the type of channel and the substrate arrow showing that it is an N-channel device. P-channel units are also available.

One problem encountered with insulated gate FETs is that too high a gate voltage may actually puncture the insulating layer, thus destroying the FET. Because of the extremely high gate resistance, even static voltages from your fingertips can puncture the oxide layer. Therefore, to avoid damage, MOSFETs are shipped with their leads shorted together. Another widely used method of protection is to wire back-to-back zener diodes across the gate connection. This is done internally as shown in Figure 3-33. The zener diodes place an upper limit on the gate voltage since they will conduct when their zener voltage is reached. Thus, the voltage applied to the gate can never exceed the zener voltage.
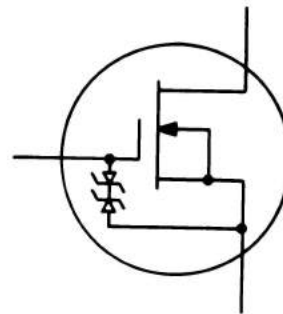
Figure 3-33
A static-voltage protected depletion mode MOSFET.

FET Circuit Arrangements

Like bipolar transistors, FET's are used primarily to obtain amplification and like bipolar transistors, FET's can be connected in three different circuit arrangements. These three configurations are commonly referred to as **common-source**, **common-gate**, and **common-drain** circuits. The circuits' connections are valid for both JFET's and IGFET's.

COMMON-SOURCE CIRCUITS

The common-source circuit is the most widely used FET circuit arrangement. This circuit configuration is comparable to the common-emitter circuit arrangement that was described in a previous section. A basic common-source configuration is shown in Figure 3-34. Notice that the input signal is applied between the gate and source leads of the FET and the output signal appears between the drain and source leads. The source is therefore common to both input and output.
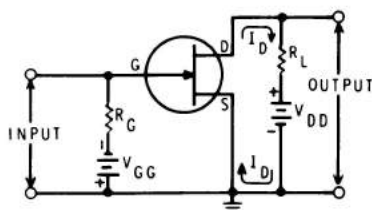
Figure 3-34
Basic common-source circuit.

An N-channel junction FET is shown in Figure 3-34 and it is therefore biased so that its gate is negative with respect to its source. The gate-to-source bias voltage is provided by an external voltage source (designated as V_{GG}) which is in series with a resistor (R_G). Therefore, V_{GG} is not applied directly to the gate and source leads. However, the FET's gate-to-channel PN junction is reverse-biased so that no current can flow through the gate lead and R_G. The voltage across R_G is therefore zero and the full value of V_{GG} is effectively placed across the gate and source leads. Normally R_G has a high value of resistance (often more than 1 megohm) so that the resistance seen at the input of the circuit will remain high. A low value of resistance could reduce the input resistance of the circuit since R_G is effectively in parallel with the FET's high input (gate-to-source) resistance as far as the input signal is concerned.

The external bias voltage (V_{GG}) is adjusted so that a specific value of drain current (I_D) will flow through the FET and the device will operate correctly. The value of I_D is also controlled (but to a lesser extent) by external voltage source V_{DD}. This voltage source supplies the necessary drain-to-source operating voltage for the FET but another resistor (R_L) is inserted between V_{DD} and the FET's drain lead. The drain current flowing through the FET therefore flows through R_L thus causing a voltage to appear across this resistor. A portion of the voltage (V_{DD}) is therefore dropped across the FET and the rest appears across R_L.

Since the FET is controlled by an input voltage (not an input current) the common-source arrangement is used to obtain voltage amplification. For example, an AC input voltage will alternately aid and oppose the input bias voltage (V_{GG}) so that the FET's gate-to-source voltage will vary with the changes in input voltage. This will in turn cause the FET to alternately conduct more and less drain current. The FET therefore acts like a variable resistor in series with a fixed resistor (R_L) and the resistance of the FET is effectively controlled by the input signal voltage. As the FET conducts more and less drain current, its drain-to-source voltage varies accordingly to produce an output voltage that changes in response to the input voltage. However; by making R_L relatively large (often more than 10 k ohms) and by biasing the FET so that its drain-to-source resistance is also high, the changes in drain current (even when small) can produce an output signal voltage that is much higher than the input signal voltage.

In addition to providing voltage amplification, the common-source circuit has another desirable feature. Due to the extremely high gate-to-source resistance of the FET, the circuit has a very high input resistance even though input resistor R_G is used. This means that the common-source circuit will usually have a minimum loading effect on its input signal source. Common-source circuits are therefore widely used in digital or computer applications where a number of circuit inputs must often be connected to the output of one circuit without affecting its operation.

The output resistance of the common-source circuit is lower than its input resistance because of the lower source-to-drain resistance of the FET. However, the circuit still has a moderately high output resistance, since the FET is usually biased to conduct a relatively low drain current which is often not more than a few milliamperes.

The common-source circuit may also be used to amplify both low frequency and high frequency AC signal voltages as well as a wide range of DC signal voltages. For example, this circuit is often used in the input rf amplifier stage of radio receivers to amplify the wide range of high frequency input signals which can vary widely in amplitude. The circuit is also used in electronic test instruments such as solid-state voltmeters or multimeters. In these applications it must amplify a wide range of DC or AC voltages and at the same time present a high input resistance to prevent undesirable loading of the circuits being tested.

Although the common-source circuit shown in Figure 3-34 is formed with an N-channel junction FET, the same basic circuit can be formed with a P-channel JFET if the polarities of the bias voltages are reversed. Also, this basic circuit can be formed with depletion-mode and enhancement-mode MOSFET's. When these insulated-gate devices are used, the additional substrate leads are generally connected to their source leads or to circuit ground. Also, it is important to note that depletion-mode MOSFET's can accept positive and negative gate voltages and may be operated with zero gate bias, while enhancement-mode devices require a gate bias voltage in order to conduct. This means that slightly different biasing arrangements must be used to accommodate the various FET types when the basic common-source arrangement is used.

COMMON-GATE CIRCUITS

The common-gate circuit may be compared to the common-base transistor circuit because the electrode which has primary control over the FET's conduction (its gate) is common to the circuit's input and output. A basic common-gate circuit is shown in Figure 3-35. An N-channel JFET is used in this circuit and it therefore requires the same basic operating voltages as the N-channel device in the common-source circuit previously described. However, in this circuit the input gate-to-source bias voltage is provided by voltage V_{SS} and resistor R_S while the output portion of the circuit is biased by the voltage V_{DD} and resistor R_L.
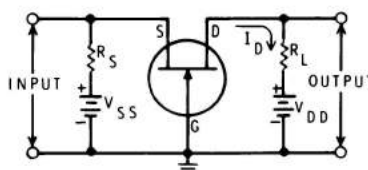
Figure 3-35
Basic common-gate circuit.

An AC input voltage will effectively vary the FET's gate-to-source voltage and cause variations in its conduction. The FET's drain current flows through R_L and variations in this current produces voltage changes across the FET which follow the input signal. The voltage developed across the drain and gate leads of the FET serves as the output.

Like the common-source circuit, the common-gate circuit provides voltage amplification; however the voltage gain of the common-gate arrangement is lower. The input resistance of the circuit is also low since current flows through the input source lead; however the output resistance of the circuit is relatively high. This makes the common-gate circuit suitable in applications where a low resistance generator must supply power to a high resistance load. When inserted between the generator and load, the common-gate circuit effectively matches the low and high resistances to insure an efficient transfer of power.

Although it has a low voltage gain, the common-gate circuit is often used to amplify high frequency AC signals. This is because of the circuit's low input resistance and because the circuit inherently prevents any portion of its output signal from feeding back and interfering with its input signal. The circuit is therefore inherently stable at high frequencies and additional components are not required to prevent interference between input and output signals. Such stabilizing components are often required with the common-source circuit arrangement previously described to insure reliable operation.

The common-gate circuit may also be formed with a P-channel JFET or with depletion-mode or enhancement-mode MOSFET's. However, when an MOSFET is used, the substrate lead is usually connected directly to the gate or to circuit ground and additional components may be required to properly bias the device.

COMMON-DRAIN CIRCUITS

The common-drain circuit is similar to the common-collector bipolar transistor circuit. This basic FET circuit arrangement is shown in Figure 3-36. The N-channel JFET in this circuit receives its gate-to-source bias voltage from V_{GG} and R_G. The output (drain-to-source) portion of the FET is biased by V_{DD} and R_L. The input signal voltage is effectively applied between the FET's gate and drain even though it appears to be applied to the gate and source leads. The output signal is effectively developed across the FET's source and drain even though it appears to be taken directly from R_L. This is because voltage source V_{DD} is effectively a short as far as the input and output signals are concerned. Therefore, V_{DD} effectively grounds the FET's drain and makes it common to both input and output signals.
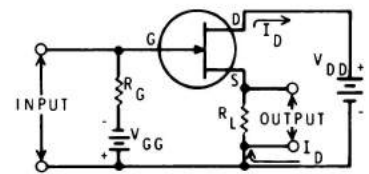
Figure 3-36
Basic common-drain or source follower circuit.

The common-drain circuit cannot provide voltage amplification. The voltage appearing at the source of the FET tends to track or follow the voltage at the gate and for this reason the circuit is often referred to as a **source-follower**. However the output source voltage is always slightly less than the input gate voltage thus causing the circuit to have a voltage gain that is less than one or unity.

The input resistance of the common-drain circuit is extremely high. In fact, this circuit has a higher input resistance than a common-source or a common-gate configuration. However, the output resistance of this circuit is low. This makes the common-drain circuit suitable for coupling a high resistance generator to a low resistance load so that an efficient transfer of power can take place.

The common-drain configuration can also be formed with P-channel JFET's as well as depletion-mode and enhancement-mode IGFET's.

Self-Review Questions

19. The junction transistor is a _____ controlled device, while the FET is a _____ controlled device.

20. In an FET electrons flow from source to drain through the _____ .

21. The major advantage of the FET is its very high _____ _____ .

22. Identify the type of FET and the terminals by filling in the blanks in Figure 3-37.

23. With no gate bias the depletion mode MOSFET is always _____ while the enhancement mode MOSFET is always
    <u>on/off</u>
    _____ .
    <u>on/off</u>
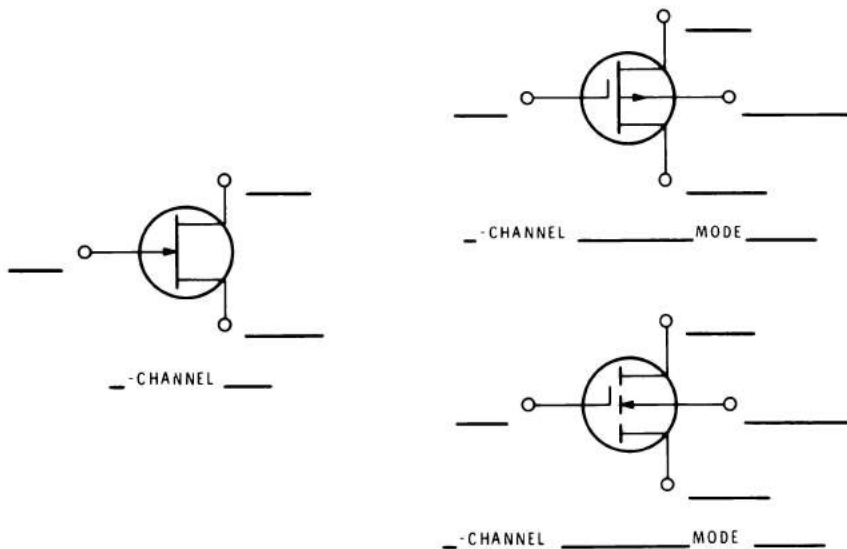
Figure 3-37
Fill in the blanks.

24. How can static voltages damage MOSFETS? _____

_____

_____

_____

25. The most widely used FET circuit arrangement is the
_____-_____ circuit.

26. The FET circuit arrangement that has the highest voltage gain is the
_____-_____ circuit.

27. The FET circuit arrangement that inherently provides the most
stable operation at high signal frequencies is the _____-
_____ circuit.

28. What FET circuit arrangement can be used to match a high impe-
dance generator to a low impedance load? _____

_____

29. What FET circuit arrangement can be used to match a low impe-
dance generator to a high impedance load? _____

Self-Review Answers

19. The junction transistor is **a current** controlled device, while the FET is a **voltage** controlled device.

20. In an FET, electrons flow from source to drain through the **channel**.

21. The main advantage of the FET is its very high **input impedance**.

22. See Figure 3-38.
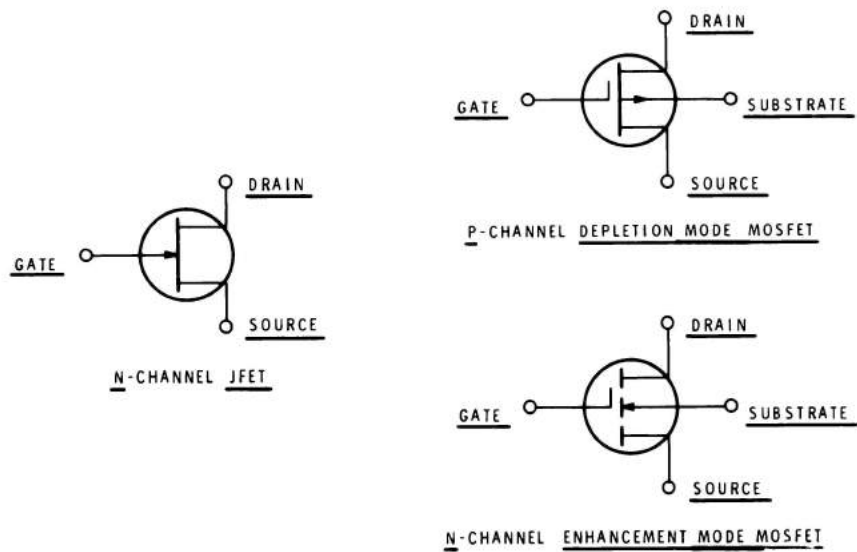
Figure 3-38

23. With no gate bias the depletion mode MOSFET is always **on** while the enhancement mode is always **off**.

24. Static voltages can puncture the oxide layer of MOSFETs and destroy the device.

25. The most widely used FET circuit arrangement is the **common-source** circuit.

26. The FET circuit arrangement that has the highest voltage gain is the **common-source** circuit.

27. The FET circuit arrangement that inherently provides the most stable operation at high signal frequencies is the **common-gate** circuit.

28. A **common-drain** or **source-follower** circuit can be used to match a high impedance generator to a low impedance load.

29. A **common-gate** circuit can be used to match a low impedance generator to a high impedance load.

OPTOELECTRONIC DEVICES

In this section you will examine a group of solid-state components which are capable of converting light energy into electrical energy or electrical energy into light energy. These components are commonly referred to as **optoelectronic** devices since their operation relies on both optic and electronic principles.

The optoelectronic devices discussed in this section are divided into two basic groups. They are classified as light-sensitive devices and light-emitting devices. The most important components found within these two categories are described in detail and various applications are considered.

In order to understand the operation of optoelectronic devices, it is first necessary to understand the basic principles of light. Therefore, we will begin this section by defining light and considering its various properties.

Basic Characteristics of Light

The term *light* is used to identify *electromagnetic radiation* which is visible to the human eye. Basically, light is just one type of electromagnetic radiation and differs from other types such as cosmic rays, gamma rays, X-rays, and radio waves only because of its frequency.

The light spectrum extends from approximately 300 gigahertz* to 300,000,000 gigahertz. It is wedged midway between the high end of the radio frequency (RF) waves, which roughly extend up to 300 gigahertz, and the X-rays, which begin at roughly 300,000,000 gigahertz. Above the X-ray region are the gamma rays and then the cosmic rays.

---

*One gigahertz, abbreviated GHz, is equal to a frequency of 1,000,000,000 Hertz or 10^9 cycles per second.

Within the 300 to 300,000,000 GHz light spectrum only a narrow band of frequencies can actually be detected by the human eye. This narrow band of frequencies appears as various colors such as red, orange, yellow, green, blue and violet. Each color corresponds to a very narrow range of frequencies within the visible region. The entire visible region extends from slightly more than 400,000 GHz to approximately 750,000 GHz. Above the visible region (between 750,000 and 300,000,000 GHz) the light waves cannot be seen. The light waves which fall within this region are referred to as **ultraviolet rays**. Below the visible region (between 300 and 400,000 GHz) the light waves again cannot be seen. The light waves within this region are commonly referred to **as infrared rays**. The entire light spectrum is shown in Figure 3-39.
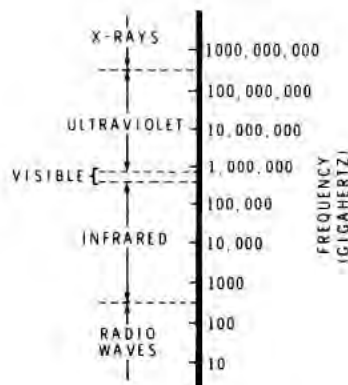
Figure 3-39
The light spectrum.

Although light is assumed to propagate or travel as electromagnetic waves, wave theory alone cannot completely explain all of the phenomena associated with light. For example, wave theory may be used to explain why light bends when it flows through water or glass. However, it cannot explain the action that takes place when light strikes certain types of semiconductor materials, and it is this resultant action that forms the basis for much of the optoelectronic theory presented in this section. In order to explain why and how semiconductor materials are affected by light it is necessary to assume that light has additional characteristics.

To adequately explain the operation of the optoelectronic devices included in this section it is necessary to consider an additional aspect of light as explained by basic **quantum** theory. Quantum theory acknowledges that light has wave-like characteristics but it also states that a light wave behaves as if it consisted of many tiny particles. Each of these tiny particles represents a discrete quanta or packet of energy and is called a **photon**.

The photons within a light wave are unchanged particles and their energy content is determined by the frequency of the wave. The higher the frequency, the more energy each photon will contain. This means that the light waves at the upper end of the light spectrum possess more energy than the ones at the lower end of the spectrum. This same rule also applies to other types of electromagnetic radiation. For example, X-rays have a higher energy content than light waves while light waves possess more energy than radio waves.

Therefore, light has a dual personality. It propagates through space like radio waves, but it behaves as if it contains many tiny particles. This particle-like aspect of light will be used in this section to explain the action that takes place in various types of optoelectronic components.

Light Sensitive Devices

Light-sensitive devices respond to changes in light intensity by changing their internal resistance or by generating an output voltage. We will now examine some of the most important of these devices.

PHOTOCONDUCTIVE CELLS

The **photoconductive cell** is one of the oldest optoelectronic components. It is nothing more than a light-sensitive resistor whose internal resistance changes with light intensity. The resistance of the device decreases nonlinearly with an increase in light intensity. In other words the resistance decreases, but the decrease is not exactly proportional to the increase in light.

Photoconductive cells are usually made from light-sensitive materials such as cadmium sulfide (Cd S) or cadmium selenide (Cd Se), although other materials such as lead sulfide and lead telluride have been used. These materials may also be doped with other materials such as copper or chlorine to control the exact manner in which the resistance of the device varies with light intensity.

Figure 3-40 shows how a typical photoconductive cell is constructed. A thin layer of light-sensitive material is formed on an insulating substrate which is usually made from glass or ceramic materials. Then two metal electrodes are deposited on the light-sensitive material as thin layers. The top view (Figure 3-40A) shows that the electrodes do not touch but leave an S-shaped portion of the light-sensitive material exposed. This allows greater contact length but at the same time confines the light-sensitive material to a relatively small area between the electrodes. Two leads are also inserted through the substrate and soldered to the electrodes as shown in the side view in Figure 3-40B. The photoconductive cell is often mounted in a metal or plastic case (not shown) which has a glass window that will allow light to strike the light-sensitive material. Also, the electrodes used with some cells may be arranged in more complicated patterns and the entire cells may be quite large (1 or more inches in diameter) or relatively small (less than 0.25 inches in diameter).
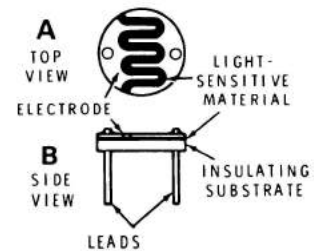
Figure 3-40
A typical photoconductive cell.
(A) topview,
(B) sideview.

Photoconductive cells are more sensitive to light than other types of light-sensitive devices. The resistance of a typical cell might be as high as several hundred megohms when the light striking its surface (its illumination) is zero (complete darkness) and as low as several hundred ohms when the illumination is over 9 footcandles. This represents a tremendous change in resistance for a relatively small change in illumination. This extreme sensitivity makes the photoconductive cell suitable for applications where light levels are low and where the changes in light intensity are small. However, these devices do have certain disadvantages. Their greatest disadvantage is the fact that they respond slowly to changes in illumination. In fact they have the slowest response of all light-sensitive devices. Also, they have a light memory or history effect. In other words, when the light level changes, the cell tends to remember previous illumination. The resistance of the cell at a specific light level is a function of the intensity, the duration of its previous exposure, and the length of time since that exposure.

Most photoconductive cells can withstand relatively high operating voltages. Typical devices will have maximum voltage ratings of 100, 200, or 300 volts DC. However, the maximum power consumption for these devices is relatively low. Maximum power ratings of 30 milliwatts to 300 milliwatts are typical.

The photoconductive cell is often represented by one of the schematic symbols shown in Figure 3-41. The symbol in Figure 3-41A consists of a resistor symbol inside of a circle. Two arrows are also used to show that the device is light-sensitive. The symbol in Figure 3-41B is similar but it contains the Greek letter λ (lambda) which is commonly used to represent the wavelength of light.

Figure 3-41
Commonly used photoconductive cell schematic symbols.
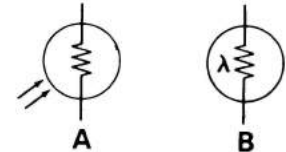
Photoconductive cells have many applications in electronics. For example, they are often used in devices such as intrusion detectors and automatic door openers where it is necessary to sense the presence or absence of light. However, they may also be used in precision test instruments which can measure the intensity of light. A simple intrusion detector circuit is shown in Figure 3-42. The light source projects a narrow beam of light onto the cell and this causes the cell to exhibit a relatively low resistance. The cell is in series with a sensitive AC relay and its 120 volt AC, 60 Hz power source. The cell allows sufficient current to flow through the circuit and energize the relay. When an intruder breaks the light beam, the cell's resistance increases considerably and the relay is deactivated. At this time the appropriate relay contacts close and apply power (from a separate DC source) to an alarm, which sounds a warning. A relay is used because it is capable of controlling the relatively high current that is needed to operate the alarm. When a large relay is used, the photocontuctive cell (because of its low power or current rating) may not be able to directly control the relay. In such a case the photoconductive cell is used to control a suitable amplifier circuit which in turn generates enough current to drive the relay.
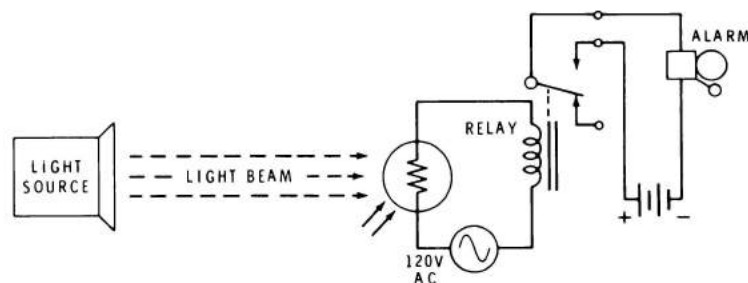
Figure 3-42
A basic intrusion detector circuit.

Since the photoconductive cell is constructed from a bulk material and does not have a PN junction, it is a bidirectional device. In other words it exhibits the same resistance in either direction and may therefore be used to control either DC or AC. Due to its bulk construction, the photoconductive cell is often referred to as a **bulk photoconductor**. However, you may also see it referred to as a **photoresistive cell** or simply a **photocell**.

PHOTOVOLTAIC CELLS

The **photovoltaic cell** is a device that converts light energy into electrical energy. When exposed to light this device generates a voltage across its terminals that increases with the light intensity. The photovoltaic cell has been used for a number of years in various military and space applications. It is commonly used aboard satellites and spacecraft to convert solar energy into electrical power which can be used to operate various types of electronic equipment. Since most of its applications generally involve the conversion of solar energy into electrical energy, this device is commonly referred to as a **solar cell**.

The photovoltaic cell is a junction device which is made from semiconductor materials. Although many different semiconductor materials have been used, the device is usually made from silicon. The structure of a silicon photovoltaic cell is shown in Figure 3-43. The device has a P-type layer and an N-type layer which form a PN junction. A metal backplate or support is placed against the N-type layer. Also, a metal ring is attached to the outer edge of the P-type layer. These pieces of metal serve as electrical contacts to which external leads may be attached.
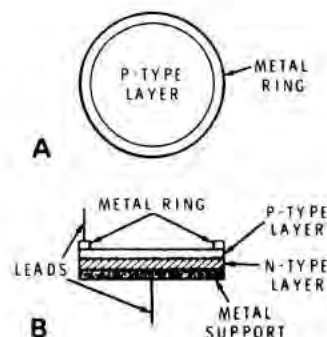
Figure 3-43
A basic silicon photovoltaic cell.
(A) topview,
(B) sideview.

The photovoltaic cell has a large surface area which can collect as much light as possible. Light strikes the top semiconductor layer within the metal ring.

Since the photovoltaic cell has a PN junction, a depletion region (an area void of majority carriers) forms in the vicinity of the junction. If the cell was forward-biased like a conventional PN junction diode, the free electrons and holes in the device would be forced to combine at the junction and forward current would flow. However, the photovoltaic cell is not used in this manner. Instead of responding to an external voltage, the device actually generates a voltage in response to light energy which strikes its surface.

In order to generate a voltage, the top layer of the photovoltaic cell must be exposed to light. The light energy striking the cell is assumed to consist of many tiny particles or photons. These photons are absorbed by the semiconductor material. A photon can intercept an atom within the semiconductor material and impart much of its energy to the atom. If sufficient energy is added to the atom, a valence electron may be knocked out of its orbit and become a free electron. This will leave the atom positively charged and a hole will be left behind at the valence site. In other words a photon can produce an electron-hole pair. This electron and hole can drift through the semiconductor material. Additional electron-hole pairs are also produced by other photons.

Some of the free electrons and holes generated by the light energy are produced within the depletion region while others are generated outside of the region but are drawn into it. The free electrons in the region are swept from the P-type to the N-type material and the holes are drawn in the opposite direction. The electrons and holes flowing in this manner produce a small voltage across the PN junction, and if a load resistance is connected across the cell's leads, this internal voltage will cause a small current to flow through the load. This current will flow from the N-type material, through the load and back to the P-type material thus making the N and P regions act like the negative and positive terminals of a battery.

All of the photons striking the photovoltaic cell do not create electron-hole pairs and many of the electrons and holes which separate to form pairs eventually recombine. The cell is therefore a highly inefficient device for converting light energy into electrical power. When this efficiency is expressed in terms of electrical power output compared to the total power contained in the input light energy, most cells will have efficiencies that range from 3 percent up to a maximum of 15 percent.

As you might expect, the output voltage produced by a photovoltaic cell is quite low. These devices usually require high light levels in order to provide useful output power. Typical applications require an illumination of at least 500 to 1000 footcandles. At 2000 footcandles, the average open-circuit (no load connected) output voltage of a typical cell is approximately 0.45 volts. When loaded, a typical cell may provide as much as 50 or 60 milliamperes of output load current. However, by connecting a large number of cells in series or parallel, any desired voltage rating or current capability can be obtained.

When used on spacecraft or satellites many photovoltaic cells are connected together, as explained above, to obtain sufficient power to operate electronic equipment or charge batteries. However, these devices are also used as individual components in various types of test instruments and equipment. For example, they are used in portable photographic light meters (which do not require batteries for operation). They are also used in movie projectors to detect a light beam which is projected through the film. The light beam is modulated (controlled) by a pattern or sound track that is printed near the edge of the film. In this way, the intensity of the light beam is made to vary according to the sounds (voice and music) that occur. The photovoltaic cell simply responds to the light fluctuations and produces a corresponding output voltage which can be further amplified and used to drive a loudspeaker which will convert the electrical energy back into sound. This application is shown in Figure 3-44.
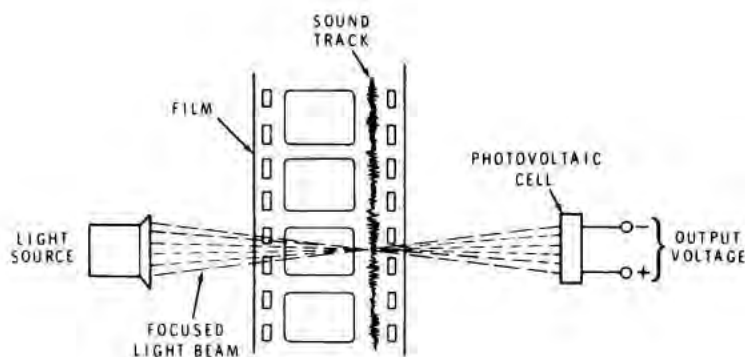
Figure 3-44
A photovoltaic cell used in a movie projector
sound reproducing system.

A schematic symbol that is commonly used to represent the photovoltaic cell is shown in Figure 3-45. This symbol indicates that the device is equivalent to a one-cell voltage source and the positive terminal of the device is identified by a plus (+) sign.

Figure 3-45
Commonly used schematic symbol for a photovoltaic cell.

PHOTODIODES

The photodiode is another light-sensitive device which utilizes a PN junction. It is constructed in a manner similar to the photovoltaic cell just described, but it is used in basically the same way as the photoconductive cell described earlier. In other words it is used essentially as a light-variable resistor.

The photodiode is a semiconductor device (usually made from silicon) and may be constructed in basically two ways. One type of photodiode utilizes a simple PN junction as shown in Figure 3-46A. A P-type region is formed on an N-type substrate as shown. This takes place through a round window that is etched into a silicon dioxide layer that is formed on top of the N-type substrate. Then a metal ring or window is formed over the silicon dioxide layer as shown. This window makes electrical contact with the P-type region and serves as an electrode to which an external lead can be attached. However, the window also accurately controls the area that will receive or respond to light. A metal base is then formed on the bottom N-type layer. This metal layer serves as a second electrode to which another lead is attached.
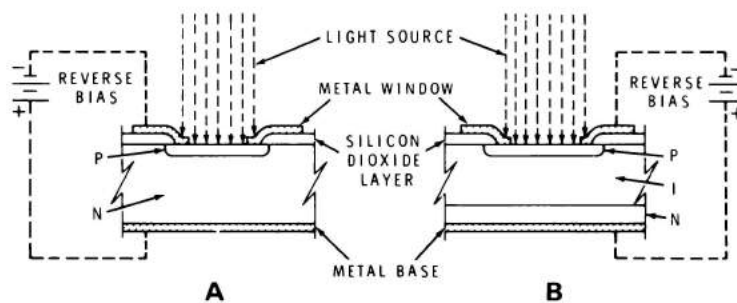
Figure 3-46
Basic construction of typical photodiodes.

The PN junction photodiode shown in Figure 3-46A operates on the same basic principle as the photovoltaic cell previously described. In fact, the photodiode may be used in basically the same manner as a photovoltaic cell. When used as a photovoltaic cell, the device is said to be operating in the **photovoltaic mode** and it will generate an output voltage (across its electrodes) that varies with the intensity of the light striking its top P-type layer. However, the photodiode is most commonly subjected to a reverse bias voltage as shown in Figure 3-46A. In other words its P-type region is made negative with respect to its N-type region. Under these conditions a wide depletion region forms around the PN junction. When photons enter this region to create electron-hole pairs, the separated electrons and holes are pulled in opposite directions because of the influence of the charges that exist on each side of the junction and the applied reverse bias. The electrons are drawn toward the positive side of the bias source (the N-type region) and the holes are attracted toward the negative side of the bias voltage (the P-type region). The separated electrons and holes therefore support a small current flow in the reverse direction through the photodiode. As the light intensity increases, more photons produce more electron-hole pairs which further increase the conductivity of the photodiode resulting in a proportionally higher current. When a photodiode is used in this manner it is said to be operating in the **photoconductive** or **photocurrent mode**.

The photodiode may also be constructed as shown in Figure 3-46B. This type of photodiode is similar to the type just described but there is one important difference. This device has an intrinsic (I) layer between its P and N regions and is commonly referred to as a **PIN** photodiode. The **intrinsic** layer has a very high resistance (a low conductivity) because it contains very few impurities. A depletion region will extend further into this I region than it would in a heavily doped semiconductor. The addition of the I layer results in a much wider depletion region for a given reverse bias voltage. This wider depletion area makes the PIN photodiode respond better to the lower light frequencies (longer wavelengths). The lower frequency photons have less energy content and tend to penetrate deeper into the diode's structure before producing electron-hole pairs and in many cases do not produce pairs. The wider depletion region in the PIN photodiode increases the chance that electron-hole pairs will be produced. The PIN photodiode is therefore more efficient over a wider range of light frequencies. The PIN device also has a lower internal capacitance due to the wide I region which acts like a wide dielectric between the P and N regions. This lower internal capacitance allows the device to respond faster to changes in light intensity. The wide depletion region also allows this device to provide a more linear change in reverse current for a given change in light intensity.

Figure 3-47
A typical photodiode package.
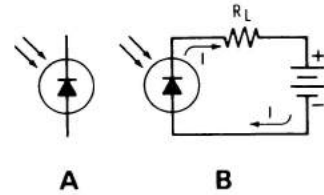
Figure 3-48
A photodiode symbol (A)
and a properly biased photodiode (B).

PN junction and PIN photodiodes are often mounted on an insulative platform or substrate and sealed within a metal case as shown in Figure 3-47. A glass window is provided at the top of the case, as shown, to allow light to enter and strike the photodiode. The two leads extend through the insulative base at the bottom of the case and are internally bonded (with fine wires) to the photodiode's electrodes.

Photodiodes have an important advantage over the photoconductive devices described earlier. A photodiode can respond much faster to changes in light intensity. In fact, the photodiode operates faster than any other type of photosensitive device. It is therefore useful in those applications where light fluctuates or changes intensity at a rapid rate. The major disadvantage with the photodiode is that its output photocurrent is relatively low when compared to other photoconductive devices.

Photodiodes and PIN photodiodes are both commonly represented by the same schematic symbol and several symbols have been used to represent these devices. A commonly used symbol is shown in Figure 3-48A. Notice that a conventional diode symbol is used with two arrows. The arrows point toward the diode to show that it responds to light. Figure 3-48B shows a properly biased photodiode. A load resistor (R_L) is also connected in series with the diode. The load resistor simply represents any resistive load which might be controlled by the photodiode as it varies its conductivity in accordance with input light intensity. The changes in the diode's conduction will cause the photocurrent (I) in the circuit to vary.

PHOTOTRANSISTORS

The phototransistor is also a PN junction device. However, it has two junctions instead of one like the photodiode just described. The phototransistor is constructed in a manner similar to an ordinary transistor, but this device is used in basically the same way as a photodiode.

The phototransistor is often constructed as shown in Figure 3-49. The process begins by taking an N-type substrate (usually silicon), which ultimately serves as the transistor's collector, and forming into this substrate a P-type region which serves as the base. Then an N-type region is formed into the P-type region to form the emitter. The phototransistor therefore resembles a standard NPN bipolar transistor in appearance. The device is often packaged much like the photodiode shown in Figure 3-47. However, in the case of the phototransistor, three leads are generally provided which connect to the emitter, base, and collector regions of the device. Also, the phototransistor is physically mounted under a transparent window so that light can strike its upper surface as shown in Figure 3-49.
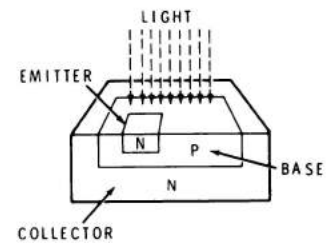
Figure 3-49
The construction of a typical phototransistor.

The operation of a phototransistor is easier to understand if it is represented by the equivalent circuit shown in Figure 3-50. Notice that the circuit shown contains a photodiode which is connected across the base and collector of a conventional NPN bipolar transistor. If the equivalent circuit is biased by an external voltage source as shown, current will flow into the emitter lead of the circuit and out of the collector lead. The amount of current flowing through the circuit is controlled by the transistor in the equivalent circuit. This transistor conducts more or less depending on the conduction of the photodiode which in turn conducts more or less as the light striking it increases or decreases in intensity. If light intensity increases, the diode conducts more photocurrent (its resistance decreases) thus allowing more emitter-to-base current (commonly referred to as base current) to flow through the transistor. This increase in base current is relatively small but due to the transistor's amplifying ability this small base current is used to control the much larger emitter-to-collector current (also called collector current) flowing through this device. The increase in input light intensity causes a substantial increase in collector current. A decrease in light intensity would correspondingly cause a decrease in collector current.
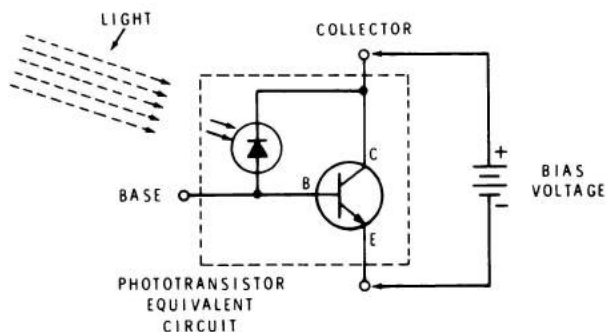
Figure 3-50
Equivalent circuit for a phototransistor.

Although the phototransistor has a base lead as well as emitter and collector leads, the base lead is used in very few applications. However, when the base is used, it is simply subjected to a bias voltage which will set the transistor's collector current to a specific value under a given set of conditions. In other words, the base may be used to adjust the phototransistor operating point. In most applications only the emitter and collector leads are used and the device is considered to have only two terminals.

An important difference between the phototransistor and the photodiode is in the amount of current that each device can handle. The phototransistor can produce much higher output current than a photodiode for a given light intensity because the phototransistor has a built-in amplifying ability. The phototransistor's higher sensitivity makes it useful for a wider range of applications than a photodiode. Unfortunately this higher sensitivity is offset by one important disadvantage. The phototransistor does not respond as quickly to changes in light intensity and therefore is not suitable for applications where an extremely fast response is required. Like other types of photosensitive devices the phototransistor is used in conjunction with a light source to perform many useful functions. It can be used in place of photoconductive cells and photodiodes in many applications and can provide an improvement in operation. Phototransistors are widely used in such applications as tachometers, photographic exposure controls, smoke and flame detectors, object counting, and mechanical positioning and moving systems.

A phototransistor is often represented by the symbol shown in Figure 3-51A and it is usually biased as shown in Figure 3-51B. As shown the phototransistor is used to control the current flowing through a load much like the photodiode shown in Figure 3-48B.

A   **B**
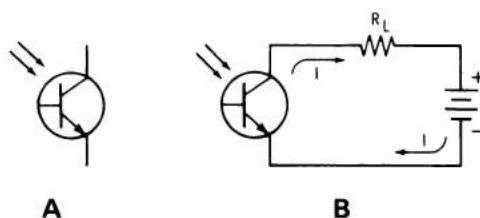
Figure 3-51
A phototransistor symbol (A)
and a properly biased transistor (B).

Light-Emitting Devices

Light-emitting devices are components which produce light when they are subjected to an electrical current or voltage. In other words, they simply convert electrical energy into light energy. For many years the incandescent lamp and the neon lamp were the most popular sources of light in various electrical applications. The incandescent lamp simply uses a metal filament which is placed inside of a glass bulb and air is drawn out of the bulb to produce a vacuum. Current flows through the filament causing it to heat up and produce light. The neon lamp utilizes two electrodes which are placed within a neon gas filled bulb. In this case current is made to flow from one electrode to the other through the gas and the gas ionizes and emits light.

The incandescent lamp produces a considerable amount of light, but its life expectancy is quite short. A typical lamp might last as long as 5000 hours. In addition to having a short life, the incandescent lamp responds slowly to changes in input electrical power. The incandescent lamp was (and still is) suitable for use as an indicator or for simply providing illumination, but due to its slow response it will not faithfully vary its light intensity in accordance with rapidly charging alternating currents. The incandescent lamp therefore cannot be effectively used to convert high frequency electrical signals (much above the audio range) into light energy which is suitable for transmission through space. The light energy produced by the incandescent lamp is not useful for carrying information which could subsequently be recovered or converted back into an electrical signal by a suitable light sensitive device.

The neon lamp has a somewhat longer life expectancy (typically 10,000 hours) than an incandescent lamp and a somewhat faster response to changes in input current. However, its output light intensity is much lower than that of an incandescent device. The neon lamp has been used for many years as an indicator or warning light and in certain applications to transmit low frequency AC signals or information in the form of light over very short distances. The neon lamp cannot be used simply for the purpose of providing illumination.

With all of their shortcomings, incandescent and neon lamps were used for many years simply because nothing better was available. However, in recent years a new type of light-emitting device was developed which has revolutionized the optoelectronics field. This newer device is a solid-state component, and it is physically stronger than the glass encased

incandescent and neon devices. Like all semiconductor devices, it has an unlimited life expectancy. This new light-emitting device is referred to as a light-emitting diode or LED. Since it is such an important solid-state component we will examine the operation and construction of an LED in detail. Then we will see how it is used in various applications.

LED OPERATION

We have seen how light energy (photons) striking a PN junction diode can impart enough energy to the atoms within the device to produce electron-hole pairs. When the diode is reverse-biased these separated electrons and holes are swept across the diode's junction and support a small current through the device. However, the exact opposite is also possible. A PN junction diode can also emit light in response to an electric current. In this case, light energy (photons) is produced because electrons and holes are forced to recombine. When an electron and hole recombine, energy may be released in the form of a photon. The frequency (or wavelength) of the photons emitted in this manner is determined by the type of semiconductor material used in the construction of the diode.

The LED utilizes the principle just described. It is simply a PN junction diode that emits light through the recombination of electrons and holes when current is forced through its junction. The manner in which this occurs is illustrated in Figure 3-52. As shown in this figure, the LED must be forward-biased so that the negative terminal of the battery will inject electrons into the N-type layer (the cathode) and these electrons will move toward the junction. Corresponding holes will appear at the P-type or anode end of the diode (actually caused by the movement of electrons) and also appear to move toward the junction. The electrons and holes merge toward the junction where they may combine. If an electron possesses sufficient energy when it fills a hole it can produce a photon of light energy. Many such combinations can result in a substantial amount of light (many photons) being radiated from the device in various directions.
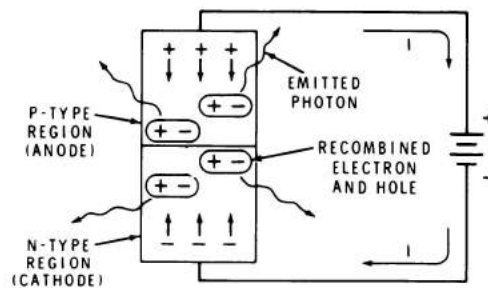
Figure 3-52
Basic operation of a light emitting diode.

At this time, you are probably wondering why the LED emits light and an ordinary diode does not. This is simply because most ordinary diodes are made from silicon and silicon is an opaque or impenetrable material as far as light energy is concerned. Any photons that are produced in an ordinary diode simply cannot escape. LED's are made from semiconductor materials that are semitransparent to light energy. Therefore, in an LED some of the light energy produced can escape from the device.

LED CONSTRUCTION

Many LED's are made of gallium arsenide (GaAs). The LED's made from this material emit light most efficiently at a wavelength of approximately 900 nanometers (10^{-9}) or 330,000 GHz which is in the infrared region of the light spectrum and is not visible to the human eye. Other materials are also used such as gallium-arsenide phosphide (GaAsP) which emits a visible red light at approximately 660 nanometers and gallium phosphide (GaP) which produces a visible green light at approximately 560 nanometers. The GaAsP device also offers a relatively wide range of possible output wavelengths by adjusting the amount of phosphide in the device. By adjusting the percentage of phosphide, the LED can be made to emit light at any wavelength between approximately 550 to 910 nanometers.

Although Figure 3-52 helps to illustrate the operation of an LED, it does not show how the device is constructed. The construction of a typical GaAsP LED is shown in Figure 3-53. Figure 3-53A shows a cross-section of the device and Figure 3-53B shows the entire LED chip. The construction begins with a gallium arsenide (GaAs) substrate. On this substrate a layer of gallium arsenide phosphide (GaAsP) is grown, however the concentration of gallium phosphide (GaP) in this layer is gradually increased from zero to the desired level. A gradual increase is required so that the crystalline structure of the substrate is not disturbed. During this growth period, an N-type impurity is added to make the layer an N-type material. The grown layer is then coated with a special insulative material and a window is etched into this insulator. A P-type impurity is then formed through the window into the layer and the PN junction is formed.

The P-type layer is made very thin so that the photons generated at or near the PN junction will have only a short distance to travel through the P-type layer and escape as shown in Figure 3-53A.
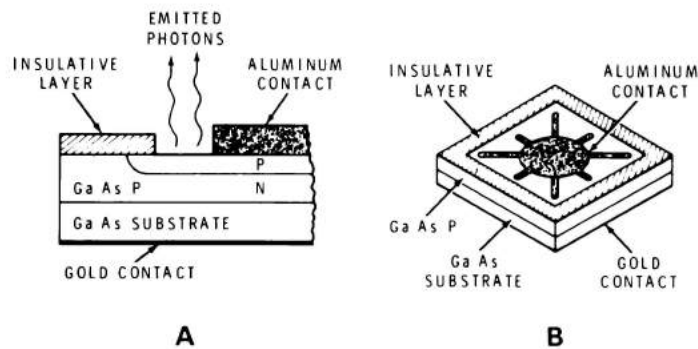
Figure 3-53
Basic construction of an LED.

The construction of the GaAsP LED is completed by attaching electrical contacts to the P-type region and the bottom of the substrate. The upper contact has a number of fingers extending outward so that current will be distributed evenly through the device when a forward bias voltage is applied across the contacts.

Once the LED is formed it must be mounted in a suitable package. Several types of packages are commonly used but all must fulfill one important requirement. All packages must be designed to optimize the emission of light from the LED. This factor is very important because the LED emits only a small amount of light. Therefore most packages contain a lens system which gathers and effectively magnifies the light produced by the LED. Also, various package shapes are used to obtain variations in the width of the emitted light beam or variations in the permissible viewing angle.

A typical LED package is shown in Figure 3-54. As shown, the package body and lens are one piece and are molded from plastic. The cathode and anode leads are inserted through the plastic case and extend up into the dome shaped top which serves as the lens. The bottom contact of the LED chip is attached directly to the cathode lead and the upper contact is connected to the anode lead by a thin wire which is bonded in place. The placement of the LED chip in this case is critical since the case serves as a lens which conducts light away from the LED and it also serves as a magnifier. In some cases the plastic lens will contain fine particles which help to diffuse the light or the entire case may be dyed or tinted with a color that enhances the natural light color emitted by the LED.

The LED package shown in Figure 3-54 is installed by simply pushing the lens through a suitable hole in a chassis or special bezel and snapping it in place. The leads are then soldered into place.

As you may suspect, the amount of light produced by an LED is small compared to an incandescent lamp. Most LED's generate a typical luminous intensity of only a few millicandelas, which is very low compared to even a miniature incandescent panel light which can produce many times that much light. However, LED's have several important advantages. First, they are extremely rugged. They also respond very quickly to changes in operating current and therefore can operate at extremely high speeds. They require very low operating voltages and are therefore compatible with integrated circuits, transistors, and other solid-state devices. They are relatively inexpensive when compared to incandescent devices. Also, they may be designed to emit a specific light color or narrow frequency range as compared to the incandescent lamp which emits a white light that contains a broad range of light frequencies.

The disadvantages associated with LED's (in addition to low light output) are similar to those which pertain to many types of solid-state components. They may be easily damaged by excessive voltage or current (beyond their maximum ratings) and their output radiant power is dependent on temperature.
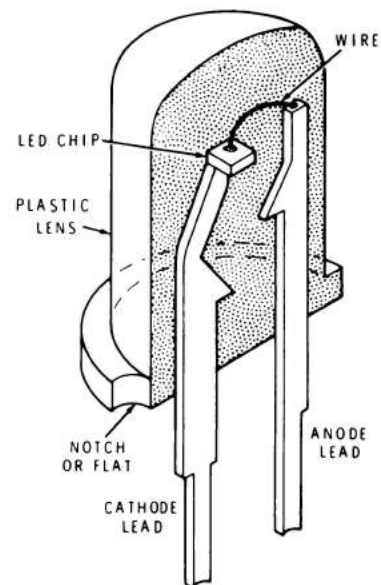
Figure 3-54
A typical LED package.

LED APPLICATIONS

In any application of an LED, the device is seldom used alone. The LED is usually connected in series with a resistor which limits the current flowing through the LED to the desired value. To operate the LED without this current limiting resistor would be risky since even a slight increase in operating voltage might cause an excessive amount of current to flow through the device. Some LED packages even contain built-in resistors (in chip form).

A schematic symbol that is commonly used to represent the LED is shown in Figure 3-55A. The correct way to bias an LED is shown in Figure 3-55B. The series resistor (R_S) must have a value which will limit the forward current (I_F) to the desired value based on the applied voltage (E) and the voltage drop across the LED which is approximately 1.6 volts. The following equation can be used to determine the required value of R_S:

$$R_S = \frac{E - 1.6}{I_F}$$

If we assume that E is equal to 6 volts and the I_F must equal 50 milliamperes (0.05 amperes) to obtain the desired light intensity, R_S must be equal to:

$$R_S = \frac{6 - 1.6}{0.05}$$

$$R_S = 88 \text{ ohms}$$

Figure 3-55
An LED symbol (A) and
a properly biased LED circuit (B).

The previous equation is useful for determining R_S as long as the required R_S value is equal to or greater than 40 ohms.

Visible light producing LED's with their respective series resistors are often used as indicator lights to provide on and off indications. Individual LED's may even be arranged into specific patterns as shown in Figure 3-56. The LED's shown in this figure each illuminate one of seven segments arranged in a special pattern. The segments can be turned on or off to create the numbers 0 through 9 and certain letters. Such devices are referred to as 7 segment LED **displays**.

Figure 3-56
Typical LED numeric displays in a standard dual in line
package. *Courtesy of Hewlett-Packard.*

LED's that emit infrared light may be used in intrusion detector systems if the light is properly focused and controlled. The infrared light cannot be seen by the human eye and is very effective in this application.

Infrared LED's are also commonly used in conjunction with light-sensitive devices, such as photodiodes or phototransistors, to form what is called an **optical coupler**. A typical optical coupler, which utilizes an LED and a phototransistor, is shown in Figure 3-57. The LED and phototransistor chips are separated by a special type of light transmitting glass, and they are coupled only by the light beam produced by the LED. An electrical signal (varying current or voltage) applied to the LED's terminals (through two of the pins on the package) will produce changes in the light beam which in turn varies the conductivity of the phototransistor. When properly biased, the phototransistor will convert the varying light energy back into an electrical signal. This type of arrangement allows a signal to pass from one circuit to another but provides a high degree of electrical isolation between the circuits. Also, the LED responds quickly to input signal changes thus making it possible to transmit high frequency AC signals through the optical coupler.
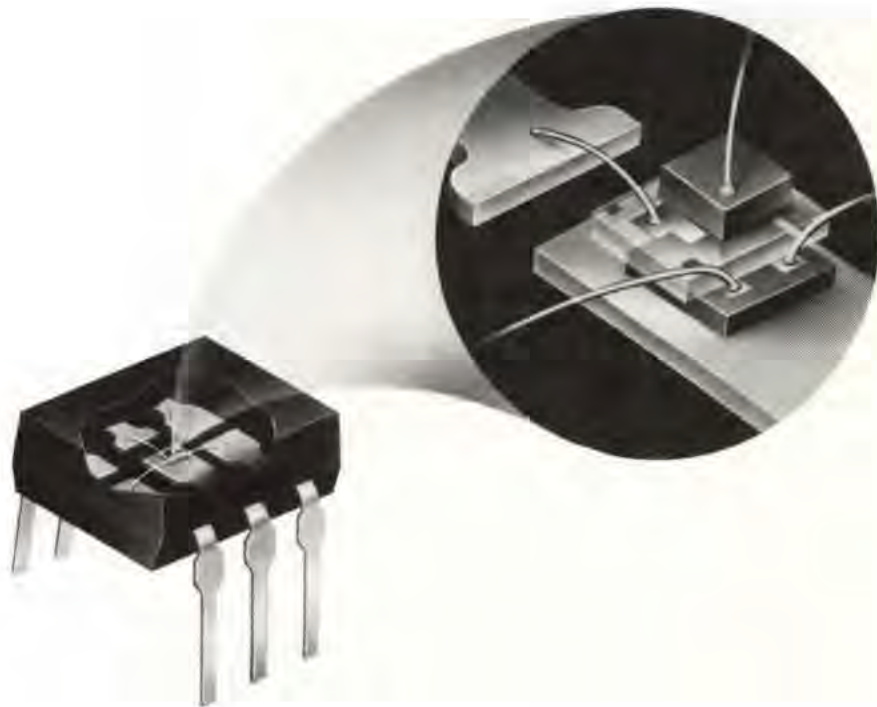
Figure 3-57
A typical optical coupler which contains an LED and a phototransistor. *Courtesy of General Electric.*

Liquid Crystals

Almost all of us have seen or, indeed, own a watch with a liquid crystal display (LCD). Not only are these devices visible in ambient light, they consume substantially less power than an LED display. This makes them ideal for watch displays. They are also used in many other devices such as the portable multimeter shown in Figure 3-58.

Figure 3-58
A portable multimeter that uses an LCD for low power consumption and high visibility.

A liquid crystal display is completely different from an LED display because the LED **generates** light, while the LCD **controls** light. That is, the LCD either blocks or permits the passage of light. Therefore, the LCD requires energy only to switch between these two states. On the other hand, the LED requires a continuous source of energy to emit light. Thus, the LCD consumes much less power than the LED.
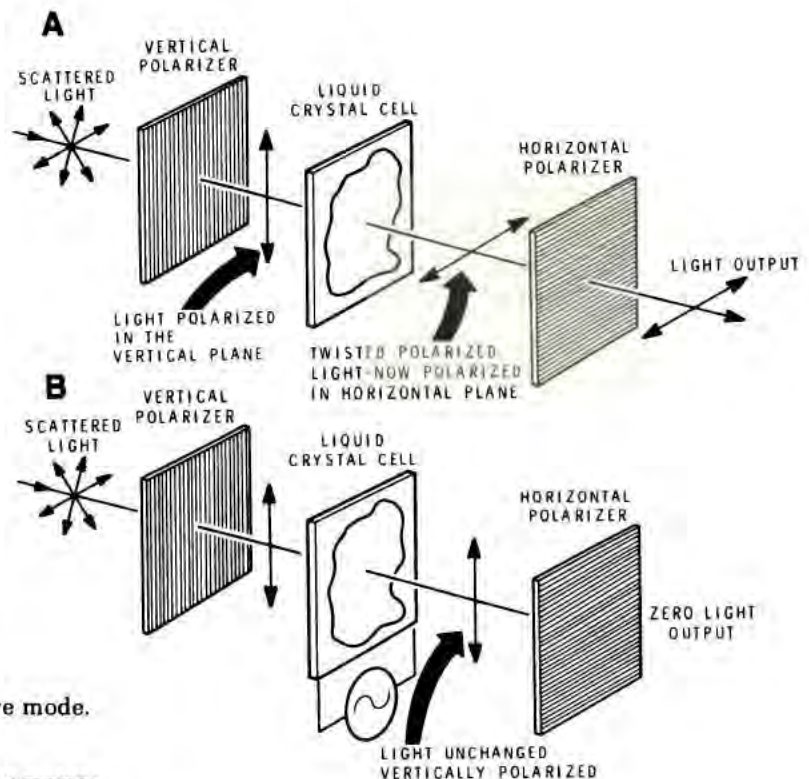
Figure 3-59
A liquid crystal display in the transmissive mode.

LCD OPERATION

LCDs have two basic modes of operation: the **transmissive** mode and the **reflective** mode.

An LCD in the transmissive mode is shown in Figure 3-59A. From the left of the diagram ambient light is traveling toward the LCD. The arrows represent the polarization of the light. Most ambient light is scattered light and, therefore, has waves that are polarized in almost all directions. However, the vertical polarizer of the LCD allows only vertically polarized light waves to pass and absorbs all other light waves. This vertically polarized light then passes through the liquid crystal. The properties of the liquid crystal are such that the light is twisted or rotated by 90°. Therefore, horizontally polarized light leaves the liquid crystal and passes uninhibited through the horizontal polarizer. Thus, the output will be some quantity of light, depending of course on the intensity of the ambient light.

However, if an AC voltage is applied across the liquid crystal, as shown in Figure 3-59B, the crystal molecules will alter their alignment in such a way that the light passing through will **not** be shifted 90° as before. Instead, it will pass through as vertically polarized light and be completely absorbed by the horizontal polarizer. Therefore, no light energy will pass to the output. Thus, the LCD can control light by either blocking it or permitting it to pass through.

REFLECTOR   HORIZONTAL   BACK
            POLARIZER   GLASS
                         PLATE
                                    LIQUID
                                    CRYSTAL   FRONT
                                               GLASS
                                               PLATE
                                                         VERTICAL
                                                         POLARIZER

METALIZED
COMMON
ELECTRODE

CHARACTER
SEGMENT
ELECTRODES
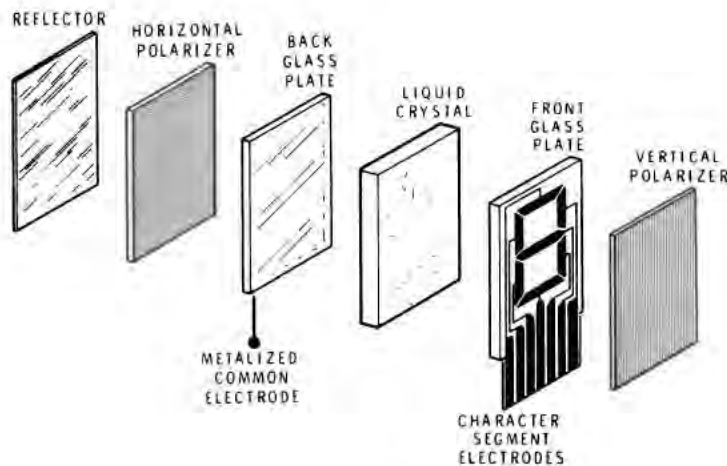
Figure 3-60
A reflective mode LCD.

An LCD operating in the reflective mode is shown in Figure 3-60. Here, with the addition of the reflector, the ambient light enters the front of the display, travels through the polarizers and liquid crystal, is reflected and then exits via the same path. In this LCD, a conductive film has been applied to the back glass plate to form a common electrode. This same conductive film is also applied to the front glass plate, except that it is applied in a special seven-segment character pattern as shown. Now, when voltage is applied to one or more of the character segments and to the common electrode, rotation of the light does not occur in the area of the activated segment(s). Therefore, the light from these segments is blocked at the horizontal polarizer. The other light passes on to the reflector and is reflected to the front of the display. The result is a dark pattern of activated character segments against a light background of reflected light.

You may have noticed that the common electrode and the character segment electrodes form the plates of a capacitor with the liquid crystal as a dielectric. Therefore, only an extremely small value of leakage current flows when a segment is activated. Thus, the drive requirements for the LCD are very low as is the power consumption.

Self-Review Questions

30. Light is considered to be _____ radiation that is visible to the human eye.

31. The light spectrum extends from _____ gigahertz to _____ gigahertz.

32. The invisible frequencies within the light spectrum are referred to as _____ and _____ rays.

33. Light has wave-like characteristics but it also behaves as if it consisted of many tiny particles known as _____ .

34. The photoconductive cell is basically a light-sensitive _____ .

35. The resistance of a photoconductive cell decreases as light intensity _____ .

36. The photovoltaic cell directly converts light into _____ energy.

37. The photovoltaic cell is sometimes referred to as a _____ cell.

38. The photodiode may operate in either the _____ or _____ mode.

39. A photodiode which has an intrinsic layer between its P and N regions is referred to as a _____ photodiode.

40. The phototransistor provides a higher output _____ than the photodiode.

41. The phototransistor's _____ and _____ leads are used in most applications.

42. Light-emitting devices convert electrical energy into _____ energy.

43. A properly biased LED emits light because of the recombination of _____ and _____ near its PN junction.

44. LED's made with gallium arsenide emit light in the _____ region.

45. Gallium phosphide LED's produce a visible _____ light.

46. An LED is to be operated at 40 mA with an applied voltage of 12 V. What value of series resistor must be used?

47. An optical coupler containing an LED and a photodiode will transmit information over a light beam but at the same time provide a high degree of electrical _____.

48. Identify the schematic symbols shown in Figure 3-61.
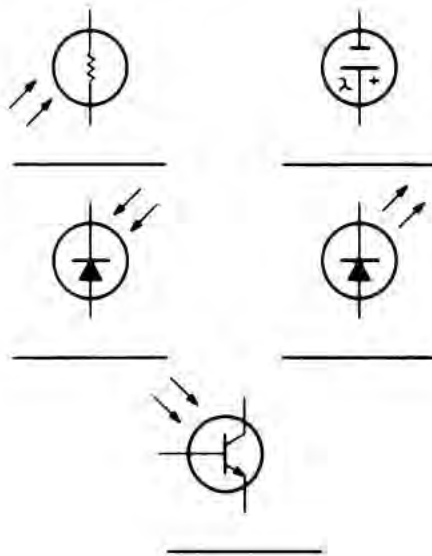
Figure 3-61
Fill in the blanks.

49. An LED display generates light, while a liquid crystal display _____ light.

50. What are the two basic modes of LCD operation? _____

51. The primary advantage of the LCD over the LED display is its low _____.

Self-Review Answers

30. Light is considered to be **electro-magnetic** radiation that is visible to the human eye.

31. The light spectrum extends from **300** gigahertz to **300,000,000** gigahertz.

32. The invisible frequencies within the light spectrum are referred to as **infrared** and **ultraviolet** rays.

33. Light has wave-like characteristics but it also behaves as if it consisted of many tiny particles known as **photons**.

34. The photoconductive cell is basically a light sensitive **resistor**.

35. The resistance of a photoconductive cell decreases as light intensity **increases**.

36. The photovoltaic cell directly converts light into **electrical** energy.

37. The photovoltaic cell is sometimes referred to as a **solar** cell.

38. The photodiode may operated in either the **photovoltaic** or **photoconductive** mode.

39. A photodiode which has an intrinsic layer between its P and N regions is referred to as a **PIN** photodiode.

40. The phototransistor provides a higher output **current** than the photodiode.

41. The phototransistor's **collector** and **emitter** leads are used in most applications.

42. Light-emitting devices convert electrical energy into **light** energy.

43. A properly biased LED emits light because of the recombination of **electrons** and **holes** near its PN junction.

44. LED's made with gallium arsenide emit light in the **infrared** region.

45. Gallium phosphide LED's produce a visible **green** light.

46. $R_S = \dfrac{E - 1.6}{I_F}$

    $R_S = \dfrac{12 - 1.6}{0.04}$

    $R_S = 260 \ \Omega$

47. An optical coupler containing an LED and a photodiode will transmit information over a light beam but at the same time provide a high degree of electrical **isolation**.

48. See Figure 3-62.

PHOTOCONDUCTIVE CELL   PHOTOVOLTAIC CELL

PHOTO DIODE   LED

PHOTOTRANSISTOR
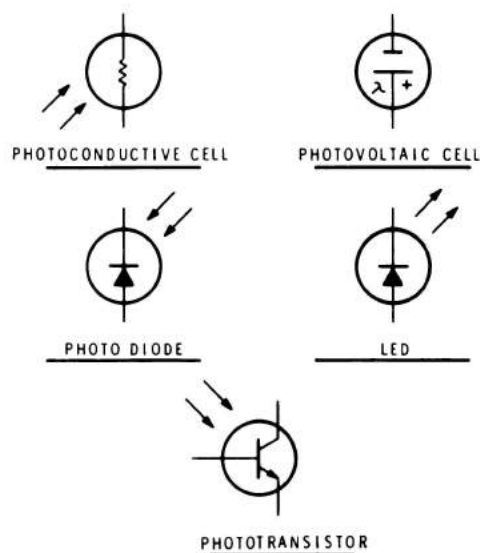
Figure 3-62

49. An LED display generates light while a liquid crystal display **controls** light.

50. The two basic modes of LCD operation are **transmissive** and **reflective**.

51. The primary advantage of the LCD over the LED display is its low **power consumption**.

INTEGRATED CIRCUITS

In this section, you will be introduced to a type of solid-state device that is known as an integrated circuit or IC. The integrated circuit is actually a group of extremely small solid-state components which have been formed within or on a piece of semiconductor material and then connected to form a complete circuit. The IC is therefore a **solid-state circuit** and not an individual component like a diode or transistor.

A magnified view of a single integrated circuit. *Courtesy of Motorola Inc.*

The Importance of IC's

Since their development in the late fifties, integrated circuits have had a tremendous effect on the electronics industry. Before IC's were developed, all electronic circuits were constructed with individual (discrete) components which were wired together. Various techniques were used to reduce the size of these discrete component circuits, but true miniaturization could not be obtained. The early vacuum tube circuits were quite large for the simple functions that they performed. The newer transistor circuits, although quite small and highly efficient when compared to their vacuum-tube counterparts, still did not offer the ultimate solution. It was the integrated circuit that finally made it possible to construct extremely small but highly efficient electronic circuits.
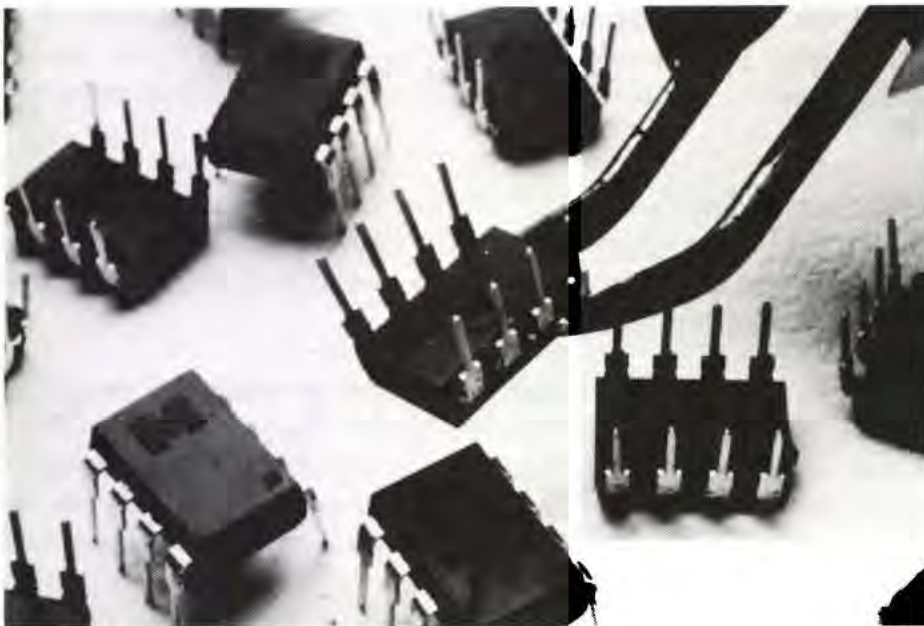
Typical miniature IC packages. *Courtesy Monsanto.*

Advantages and Disadvantages

The small size of the integrated circuit is its most apparent advantage. A typical IC can be constructed on a piece of semiconductor material that is less than one tenth of an inch square. Even after the IC has been packaged, it still occupies only a small amount of space.

The IC's small size also produces other fringe benefits. The smaller circuits consume less power than conventional circuits and therefore cost less to operate. They generate less heat and therefore generally do not require elaborate cooling or ventilation systems. The smaller circuits are also capable of operating at higher speeds because it takes less time for signals to travel through them. This is an important consideration in the computer field where thousands of decision-making circuits are used to provide rapid solutions to various types of problems.

The integrated circuit is also more reliable than a conventional circuit. This is because every component within the IC is a solid state device and because the components are permanently connected together. This eliminates any problems due to faulty connections. The IC also undergoes an extensive testing procedure which insures that it operates correctly.

Another potential advantage of IC's is low cost. However to take advantage of this, standard IC's must be used or very large quantities ordered. It can actually cost more to use IC's when a special purpose, one-of-a-kind circuit must be made.

The disadvantages of IC's are also related to their small size. Because of their size, IC's cannot handle high voltages or high power levels. High voltages can break down the insulation between the components within the IC because they are very close together. These components can also be easily damaged by the high heat levels associated with high power operation.

Another apparent disadvantage of IC's is that they cannot be repaired. However, this is offset by their low cost and the resultant easier troubleshooting. IC's are easier to troubleshoot because the problem need only be traced to a specific circuit instead of an individual component. This greatly simplifies the job of maintaining highly complex equipment.

Table I summarizes the advantages and disadvantages of integrated circuits.

ADVANTAGES	DISADVANTAGES
Small Size Consume Less Power High Speed Operation High Reliability Cheaper in Large Quantities	Cannot operate at High Power or High Voltage Cannot be repaired

Table I.
IC Advantages Vs. Disadvantages.

Applications of IC's

Integrated circuits may be placed into two general categories. They can be classified as either **digital** IC's or as **linear** IC's. Digital IC's are the most widely used devices. They are simply switching circuits which handle information and they are designed for use in various types of logic circuits and in digital computers. A linear IC provides an output signal that is proportional to the input signal applied to the device. Linear IC's are widely used to provide such functions as amplification and regulation. They are often used in television sets, FM receivers, electronic power supplies, and in various types of communications equipment.

It would be virtually impossible to consider all of the possible applications of digital and linear IC's, however, we can examine a few typical examples. First we will examine several digital IC's and then a typical linear IC.

Digital IC's

Most of the integrated circuits in use today are digital IC's. These devices are widely used in digital computers and portable electronic calculators to perform various arithmetic and decision making functions. Digital IC's are produced using both junction transistors and MOSFET construction techniques.

A typical digital IC which is formed by using junction transistor techniques is shown in Figure 3-63. A schematic diagram of the IC is shown in Figure 3-63A. Notice that only transistors, diodes, and resistors are used in the circuit and since it contains only 11 components, it is classified as a small scale integrated circuit. In this circuit, the transistors are the key elements and because of the unique manner in which they are connected, the circuit is commonly referred to as a **transistor-transistor logic (TTL)** circuit.

The TTL circuit in Figure 3-63A performs an important logic function. It is capable of comparing two input voltage levels, which must be equal to either 0 volts or approximately 3 volts, and provide an output voltage level (0 or 3 volts) depending on the input combination. The circuit

INPUTS		OUTPUT
0	0	1
0	1	1
1	0	1
1	1	0
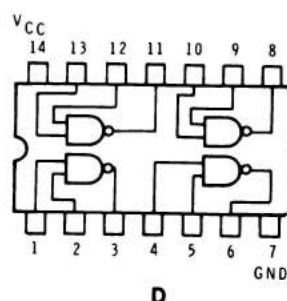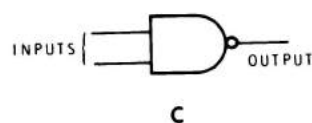
Figure 3-63
A typical TTL digital IC.

performs what is commonly referred to as the NAND function and the circuit itself is referred to as a NAND gate since it provides a gating or switching action between two voltage levels. The NAND gate, like all digital circuits, is capable of recognizing only two voltage levels (sometimes called logic levels) at each of its inputs. Instead of referring to the specific voltages involved (which can vary with different types of digital circuits), it is common practice to refer to one level as a **high** logic level or a logic 1 and the other voltage level as a **low** logic level or a logic 0. A digital circuit can therefore be thought of as a device which responds to various high and low (1 or 0) logic levels and the actual voltages involved can be ignored. The table in Figure 3-63B shows the output levels (1 or 0) produced by the NAND gate when all possible combinations (only 4 are possible) of input levels have been applied. The NAND gate is therefore capable of making a simple decision based on the combinations of logic levels at its inputs and provide a specific output logic level for each combination.

Since many thousands of NAND gates are used in digital computers and other complex digital systems, it is not practical to draw the entire circuit each time it is shown on a schematic. Therefore, the NAND gate is usually represented by the symbol shown in Figure 3-63C. Notice that only the two inputs and the output of the circuit are represented.

It is common practice to construct not one but four of these NAND gate circuits on a single IC chip and mount the chip in a single package. The dual in-line package is widely used with IC's of this type. The outline of a typical dual in-line package is shown in Figure 3-63D. This package outline drawing shows how the various NAND gates are internally connected to the package leads. Notice that the IC package has 14 leads (also called pins), which are consecutively numbered in a counterclockwise direction. The package also has a notch at one end, which serves as a key to help locate pin number 1. Notice that pins 1 and 2 serve as inputs to one gate and pin 3 provides the output connection. Power is simultaneously applied to all four circuits through pins 14 and 7.

The package outline drawing shown in Figure 3-63D is typical of those provided by IC manufacturers in their specification sheets. They may also provide a schematic of the particular circuit involved and, of course, will always provide the important electrical characteristics of the circuit. In many cases, the circuit designer or engineer is more interested in what the circuit can do and is less interested in how it does it or how it is constructed. Therefore, specification sheets are likely to contain more mechanical and electrical information which pertains to the overall performance of the IC and very little information relating to its internal construction.

A typical example of a digital IC which is formed by using MOSFET techniques is shown in Figure 3-64. This IC utilizes one of the newest and most advanced construction techniques. It contains both P-channel and N-channel enhancement mode MOS field-effect transistors and is commonly referred to as a complementary-symmetry/metal-oxide semiconductor IC or simply a CMOS IC. CMOS circuits containing P- and N-channel MOSFET's are now widely used because they have many advantages over other types of digital circuits. They consume less power than other types of digital IC's and they have good temperature stability. They can operate over a wide range of supply voltages (typically 3 to 15 volts), as compared to TTL circuits which require an accurate 5 volt supply. CMOS circuits also have a high input resistance, which makes it possible to connect a large number of circuit inputs to a single output without loading down the output and disrupting circuit operation. This is an extremely important advantage in digital equipment where thousands of circuits are used.

The circuit shown in Figure 3-64A contains four MOSFET's which are interconnected so that they can perform a useful logic function. The resulting circuit is referred to as a NOR gate, and like the NAND gate previously mentioned, it is a fundamental building block that is used to construct complex digital circuits. However, the NOR gate responds differently to various combinations of input voltage levels. The output levels (1 or 0) produced by the NOR gate for all possible combinations of input levels are shown in the table of Figure 3-64B.
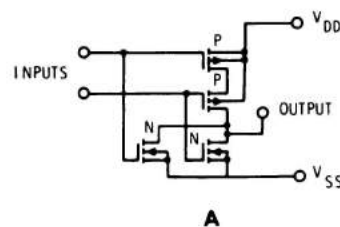
INPUTS		OUTPUT
0	0	1
0	1	0
1	0	0
1	1	0
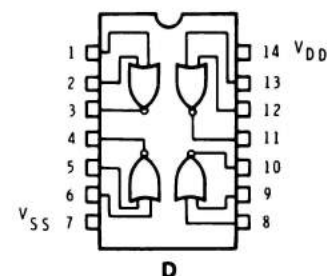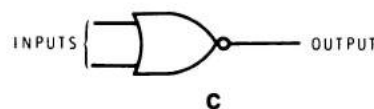
Figure 3-64
A CMOS digital IC.

The NOR gate symbol is shown in Figure 3-64C. This symbol is generally used is place of the actual schematic. Four of these NOR gates are usually formed on one IC chip and mounted in a single IC package. An outline drawing of a typical dual in-line package, which contains four NOR gates, is shown in Figure 3-64D.

Both TTL and CMOS circuits can be used to perform NAND or NOR functions or a variety of other logic functions that must be performed in a highly complex digital system. The circuits shown in Figures 3-63 and 3-64 are therefore typical of the small scale integrated circuits used in digital equipment. These circuits may be thought of as basic building blocks which can be used to construct complex digital systems that can perform useful operations. Digital ICs and their applications will be discussed in detail in a later unit.

Linear IC's

As mentioned previously, linear circuits provide outputs that are proportional to their inputs. They do not switch between two states like digital circuits. The most popular linear circuits are the types that are designed to amplify DC and AC voltages. In fact, a high performance amplifier circuit, known as an **operational amplifier**, is widely used in various types of electronic equipment.

The operational amplifier can amplify DC or AC voltages and has an extremely high gain. An operational amplifier can be constructed with discrete components, but it is more commonly produced in IC form and therefore sold as a complete package which is designed to meet certain specifications. However, the operational amplifier is designed so that it can be used in a variety of applications. Its gain (ability to amplify) can be controlled by using additional external components and it usually has built-in features which make it possible to adjust its operation in various ways.

A typical operational amplifier circuit is shown in Figure 3-65A. The circuit contains transistors, resistors, and capacitors which are interconnected to form a highly efficient amplifying circuit. The circuit has two inputs and one output as shown. One input is commonly referred to as the plus (+) or non-inverting input and the other is referred to as the minus (−) or inverting input. The circuit will amplify either DC or AC signals applied to either input. However, signals applied to the (+) input are not inverted when they appear at the output. In other words, as the input voltage goes positive or negative, the output voltage correspondingly goes positive or negative. When a signal is applied to the (−) input, inversion takes place. In other words, the polarity of the output signal is always opposite to that of the input signal. This unique feature greatly increases the versatility of the circuit.

When the input voltage is equal to zero, the output voltage should also be equal to zero. However, in practice the output voltage may be offset by a slight amount since component tolerances make it impossible to construct a perfectly balanced circuit. Therefore, two offset null terminals are provided so that the circuit can be appropriately balanced. This is done by simply connecting the opposite ends of a potentiometer to the offset null terminals, but the arm of the potentiomenter is connected to circuit ground. The potentiometer may then be adjusted to balance the circuit.
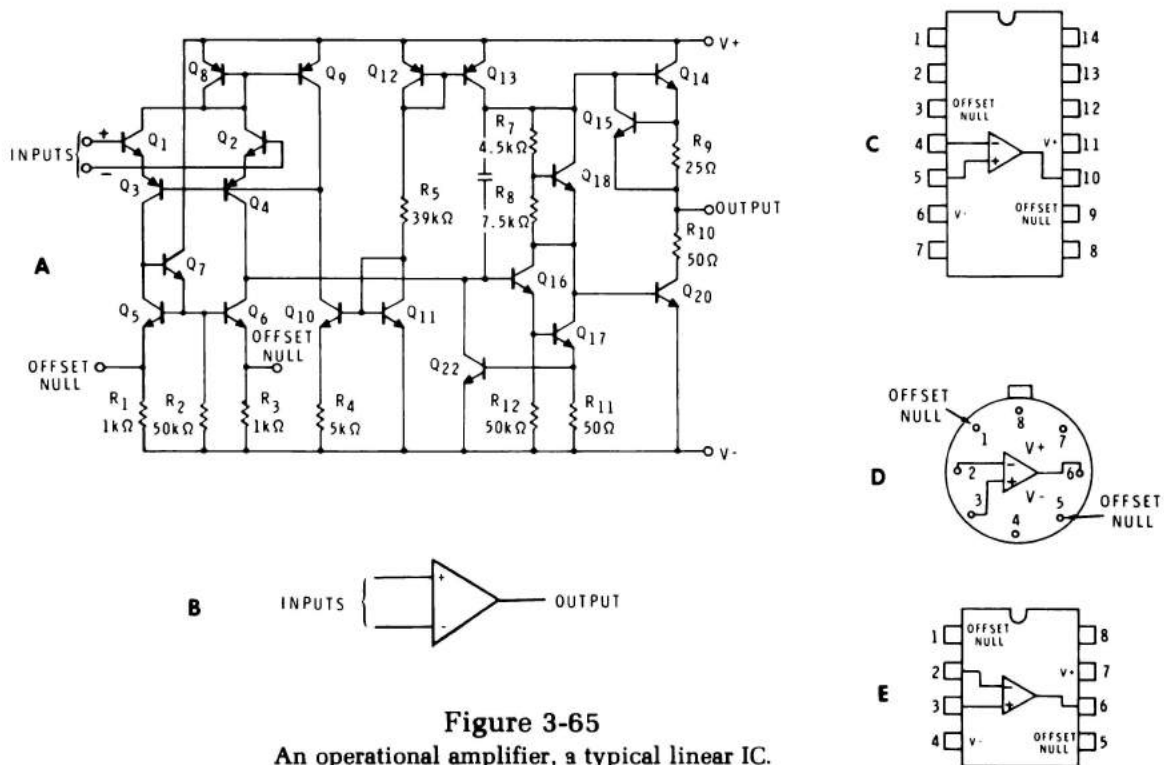
Figure 3-65
An operational amplifier, a typical linear IC.

Power is applied to the operational amplifier through terminals V+ and V−. The circuit therefore requires a positive voltage source and a negative voltage source. Most operational amplifiers can operate over a reasonably wide range of supply voltages, but these voltages should never exceed the maximum limits set by the manufacturer of the device. Operational amplifiers also consume very little power. Most units have maximum power dissipation ratings of 500 milliwatts or less.

The operational amplifier has an extremely high input resistance, but its output resistance is very low. The device also has an extremely high voltage gain. Many units are guaranteed to amplify an input voltage by at least 15,000 to 20,000 times and some of these units have typical gains that can exceed several hundred thousand or even more than a million.

In most cases, operational amplifiers are not used alone and their full amplifying capabilities are not utilized. Instead, it is common practice to connect external components to the amplifier in a manner which will allow a small portion of the output signal to return to the input and control the overall gain of the circuit. A lower gain is obtained in this manner, but circuit operation becomes more stable and predictable.

The operational amplifier is commonly represented by the symbol shown in Figure 3-65B. Notice that the (+) and (−) inputs are identified in the symbol. The operational amplifier circuit is generally packaged in a variety of ways to suit a broad range of applications. For example, the circuit in Figure 3-65A is available in a dual in-line package (DIP) as shown in Figure 3-65C or in a metal can packaged as shown in Figure 3-65D. It is even available as a mini DIP as shown in Figure 3-65E.

Operational amplifiers are used in various types of electronic equipment. They are the most important components used in electronic analog computers because their linear characteristics can be used to provide multiplying and summing operations. When used in analog computers, voltages are used to represent (are analogous to) the actual quantities to be multiplied or added. The extremely small operational amplifier IC's are suitable for use in portable electronic equipment where weight and power consumption must be held to a minimum. They are also suitable for use in portable test instruments and in communications equipment.

Various types of voltage regulator circuits are also constructed in IC form. These linear devices are used to convert an unregulated DC voltage (obtained through AC rectification) into a regulated DC output voltage which remains essentially constant while supplying a wide range of output currents. These voltage regulator IC's have replaced many of the discrete component regulators that were once widely used. Some IC regulators provide only one fixed output voltage, but other types are available which have adjustable outputs.

Special types of linear IC's are also designed for specific applications. For example, special linear IC's are designed for use in FM receivers where they are used to detect (recover the information in) FM signals. Some IC's are designed for use in solid-state color television receivers where they are used to detect, process and automatically control the chroma (color) signals. Others are used to provide extremely simple operations such as generating a signal to operate a lamp or some other indicator when an FM receiver or TV receiver is properly tuned. Some linear IC's are even used as an interface between digital circuits when digital information must be transmitted over a long transmission line. Such devices are commonly referred to as line drivers and receivers and they are used in digital systems even though they are basically linear devices.

Linear IC's and their applications will be discussed in greater detail in later units.

Self-Review Questions

52. What is an integrated circuit? _____
     _____
     _____

53. List the advantages and disadvantages of integrated circuits. ___
     _____
     _____

54. Name the two types of IC's.

     1. _____

     2. _____

55. Digital circuits deal with only _____ distinct logic levels.
     These are referred to as a _____ logic level and a
     _____ logic level.

56. The circuit shown in Figure 3-63A is known as a _____
     digital logic circuit, while that shown in Figure 3-64A is a
     _____ digital logic circuit.

57. A CMOS digital IC usually has a higher input _____
     than a TTL digital IC.

58. A linear circuit provides an output that is _____
     _____ to its input.

59. One of the most popular linear IC's is the _____
     amplifier.

60. The input resistance of an operational amplifier is
     _____ _____, the output resistance is
     _____ _____, and the voltage gain is
     _____ _____ .

Self-Review Answers

52. An integrated circuit is a solid-state circuit made up of a group of extremely small solid-state components which have been formed within or on a piece of semiconductor material.

53. Integrated circuit **advantages**:

   small size

   consume less power

   high speed operation

   high reliability

   cheaper in large quantities

   Integrated circuit **disadvantages**:

   cannot operate at high power or high voltage levels

   cannot be repaired.

54. The two types of IC's are:

   1. **Digital**

   2. **Linear**

55. Digital circuits deal with only **2** distinct logic levels. These are referred to as a **high** logic level and a **low** logic level.

56. The circuit shown in Figure 3-63A is known as a **TTL** digital logic circuit, while that shown in Figure 3-64A is a **CMOS** digital logic circuit.

57. A CMOS digital IC usually has a higher input **impedance** than a TTL digital IC.

58. A linear circuit provides an output that is **proportional** to its input.

59. One of the most popular linear IC's is the **operational** amplifier.

60. The input resistance of an operational amplifier is **very high**, the output resistance is **very low**, and the voltage gain is **very high**.

Unit 4

ELECTRONIC CIRCUITS

CONTENTS

INTRODUCTION

In this unit you will apply your knowledge of electronic devices and components. You will learn how these devices are used in power supplies, audio and radio frequency amplifiers, and oscillators. These circuits are used in all electronic equipment and are, therefore, vital to your understanding of electronics.

The "Unit Objectives," on the following page, state exactly what you are expected to learn from this unit. Study this list now and refer to it often as you study the text.

UNIT OBJECTIVES

When you have completed this unit, you should be able to:

1. Identify half-wave, full-wave, and bridge rectifier circuits.

2. State the ripple frequency and voltage polarity present at the output of a rectifier circuit.

3. State the purpose of a power supply filter and identify some common circuits.

4. Determine the peak inverse voltage a diode is subjected to in a rectifier circuit.

5. State how you can obtain a higher PIV than is available from a single diode.

6. Identify diode protection circuitry and state its operation.

7. State why electrolytic capacitors are used in power supply filters.

8. Define "capacitor working voltage rating," and state how you can increase it.

9. Identify voltage doubler circuits.

10. State the two purposes of a bleeder resistor.

11. Identify zener diodes and IC voltage regulators.

12. List the two types of audio amplifiers.

13. Find the gain or loss in decibels when given two power, voltage, or current levels.

14. Identify the four types of audio amplifier coupling circuits.

15. Define class A, B, AB, and C amplifier operation.

16. Find amplifier efficiency when given input and output power.

17. Name the basic operational amplifier terminals and state their function.

18. List the characteristics of an ideal operational amplifier.

19. Identify the three basic op amp circuit configurations and find the gain of each.

20. Identify the three types of coupling used in RF amplifiers.

21. State the requirements for an oscillator.

22. Identify the schematic diagrams of Armstrong. Hartley, Colpitts, and crystal oscillators.

23. Define piezoelectric effect.

24. State what determines the natural frequency of a crystal.

25. State the advantages and disadvantages of crystal oscillators.

POWER SUPPLIES

Most types of electronic equipment require direct current. In a few cases, such as transistor radios, this direct current is supplied by a battery. But more often, electronic devices are operated from the AC power line. In these cases, a special circuit, called a "power supply," converts the alternating current supplied by the power company to the direct current required by the electronic device. In this section you will study several circuits that are used for this purpose.

Rectifier Circuits

The heart of the power supply is the rectifier circuit, which converts the AC sine wave to a pulsating DC voltage. This is the first step in producing the smooth DC voltage required by the electronic circuits. Here, we will discuss three different types of rectifier circuits: the **half-wave, full-wave**, and **bridge** rectifier circuits.

HALF-WAVE RECTIFIER

The half-wave rectifier circuit is shown in Figure 4-1A. The load which requires the direct current is represented by a resistor. The solid-state diode connected in series with the load acts like a one-way switch that allows current to flow only in one direction.

The operation of the circuit during the positive half cycle is shown in Figure 4-1B. The anode of the diode is positive. Therefore, the diode conducts, allowing current to flow through the load. Thus, the positive half cycle of the input sine wave is developed across the load.

The next half cycle is shown in Figure 4-1C. Here, the anode is negative and the diode cannot conduct. Therefore, no current flows through the load and no voltage is developed across it.
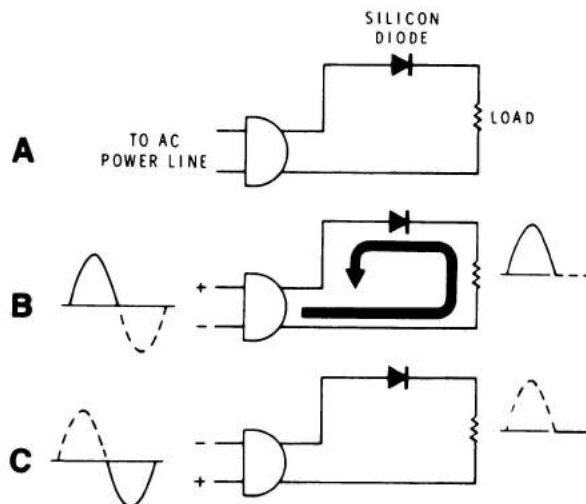
Figure 4-1
Half-wave rectifier circuit.

As you can see, the AC sine wave is changed to a pulsating DC voltage. This voltage is not suitable for most loads. However, this is only the first stage of the power supply. Later stages smooth out the ripples and change the pulsating DC to a steady DC.

The half-wave rectifier shown in Figure 4-1 will always produce the same output voltage level because it is connected across the 115 VAC line. To obtain higher or lower voltage levels, a step-up or step-down transformer is used. For example, Figure 4-2 shows a half-wave rectifier with a step-down transformer. Here the 115 VAC primary voltage is stepped down to 11.5 volts. It is then rectified and used to supply the load with approximately 11 VDC.

Ripple Frequency   The half-wave rectifier gets its name from the fact that it operates only during one half of each input cycle. Its output consists of a series of pulses. The frequency of these pulses is the same as the input frequency. Thus, when it is operated from the 60 Hz line, the frequency of the pulses will be 60 Hz. This is called the ripple frequency.
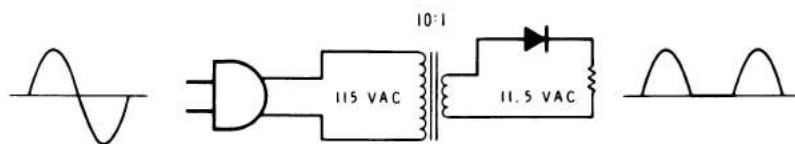
Figure 4-2
Half-wave rectifier circuit with a step-down transformer.

Output Polarity   A rectifier can produce either a negative or a positive output voltage. The output polarity depends on what point is connected to ground and which way the diode is connected. In Figure 4-3A, current can flow only in the direction indicated. This produces a positive voltage at the top of the load with respect to ground. Thus, this rectifier produces a positive output voltage.

If you need a negative output voltage, the diode can be turned around as shown in Figure 4-3B. Or, ground can be connected to a different point in the circuit as shown in Figure 4-3C. Either of these arrangements will produce a negative output voltage. However, if both changes are incorporated, the output voltage is positive again as shown in Figure 4-3D.
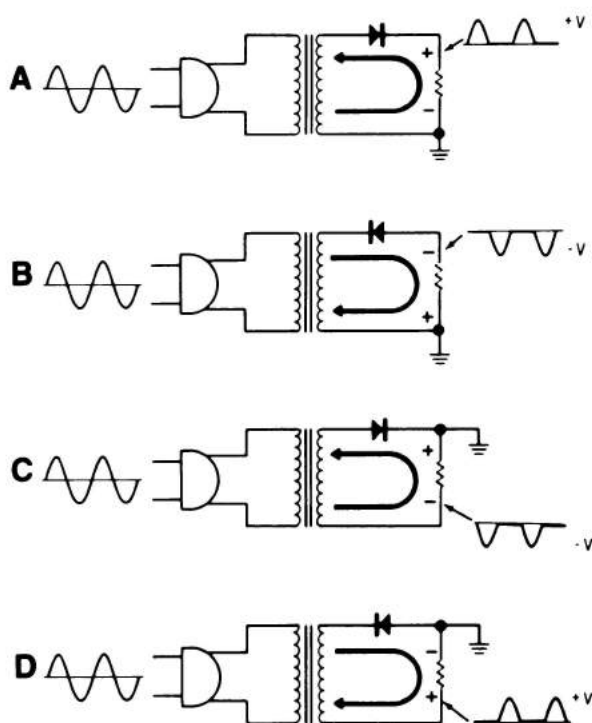
Figure 4-3
You can change the output polarity by turning the diode
around or by moving the ground connection.
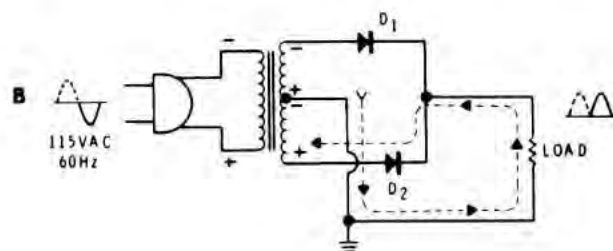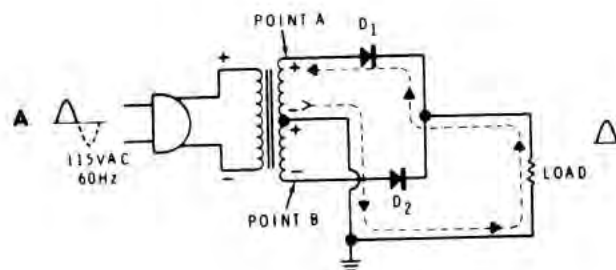
Figure 4-4
The full-wave rectifier.

FULL-WAVE RECTIFIER

Figure 4-4 shows 'a full-wave rectifier circuit. Notice that it uses two diodes and a center-tapped transformer. When the center-tap is grounded, the voltages at the opposite ends of the secondary are 180° out of phase with each other. Thus, when the voltage at point A swings positive with respect to ground, the voltage at point B swings negative. Let's examine the operation of the circuit during this half cycle.

Figure 4-4A shows that the anode of D_1 is positive while the anode of D_2 is negative. Therefore, only D_1 can conduct. Thus, current flows from the center tap up through the load and D_1 to the positive potential at the top of the secondary. The positive half cycle is therefore felt across the load.

Figure 4-4B shows what happens during the next half cycle, when the polarity of the voltage reverses. The anode of D_2 swings positive while the anode of D_1 swings negative. Therefore, D_1 cuts off and D_2 conducts. Thus, current flows from the center tap, through the load and D_2 to the positive potential at the bottom of the secondary.

Notice that current flows in the same direction through the load during **both** half cycles. This is a decided advantage over the half-wave rectifier, where current flows only during one half of the input cycle. Note also that there are two output pulses for every input cycle. Therefore, the ripple frequency is twice the input frequency. Thus, when it is operated from a 60 Hz line, the ripple frequency is 120 Hz.

Figure 4-5 illustrates a disadvantage of the full-wave rectifier. For a given transformer, it produces a peak output voltage which is half that of the half-wave rectifier. This is because the full-wave rectifier uses only one-half of the transformer secondary at one time.
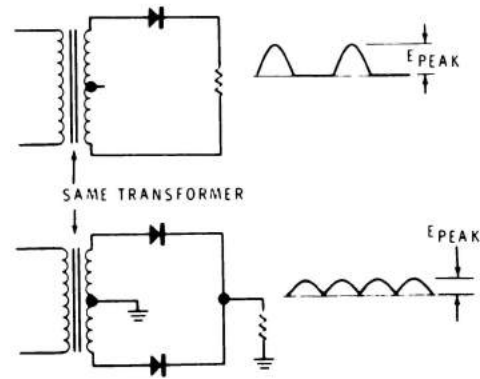
Figure 4-5
For a given transformer, E_{peak} is higher in the half-wave rectifier.

BRIDGE RECTIFIER

The bridge rectifier circuit is shown in Figure 4-6. It consists of four diodes arranged so that current can flow through the load in only one direction. This circuit does not require a center-tapped transformer as the full-wave rectifier did. In fact it does not require a transformer at all except to provide a voltage step-up or step-down.

Figure 4-6A shows how current flows on the positive half cycle of the sine wave. Current flows from the bottom of the secondary through D_1, the load, and D_2, to the positive potential at the top of the secondary. With D_1 and D_2 acting as closed switches, the entire secondary voltage is developed across the load.

On the next half cycle, the polarities reverse, as shown in Figure 4-6B. The top of the secondary is now negative and the bottom is positive. Current flows from the top of the secondary through D_3, the load, and D_4 to the positive potential at the bottom of the secondary. Notice that the current flow through the load is always in the same direction. With D_3 and D_4 acting as closed switches, the entire secondary voltage is again developed across the load.

The advantages of the bridge rectifier are that it gives full-wave rectification and the full secondary voltage appears across the load.
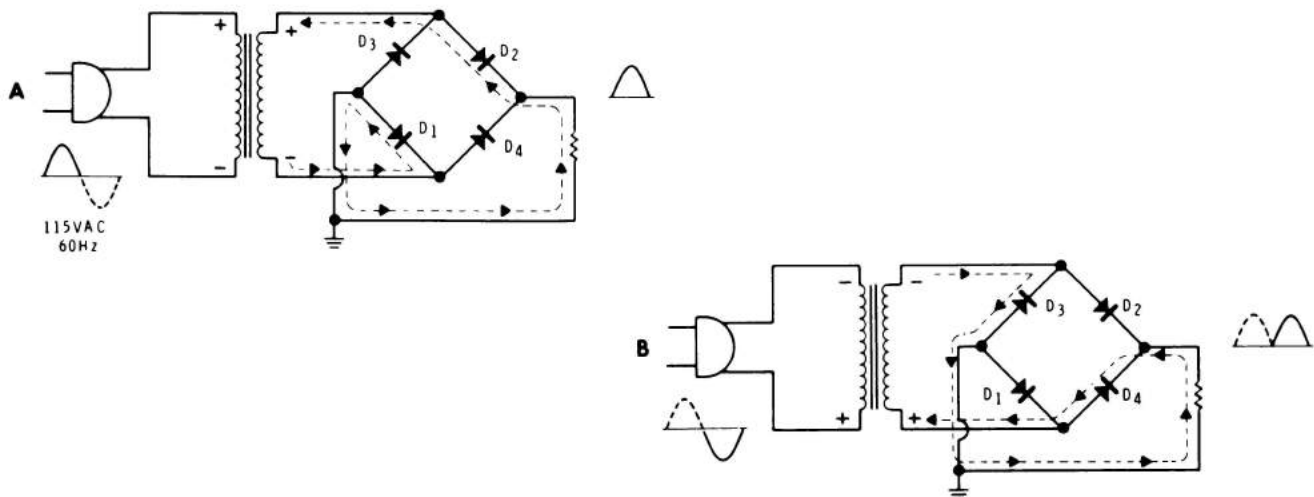
Figure 4-6
Current flow through the bridge rectifier.

Power Supply Filters

Most electronic circuits require a very smooth, constant voltage supply. Since the output of the rectifier is a pulsating DC voltage, it is unsuitable for nearly all electronic applications. Therefore, the rectifier circuit must be followed by a filter, which converts the pulsating DC into a smooth DC voltage.

CAPACITIVE FILTER

In its simplest form, the power supply filter may be nothing more than a capacitor connected across the output of the rectifier. This is shown in Figure 4-7A. Note that when the diode's anode is positive, current flows through the load. Simultaneously, a charge current flows to C_1, charging it to the peak of the input voltage.

When the diode's anode swings negative, as shown in Figure 4-7B, the current flow from the AC line is cut off. At this time, C_1 starts to discharge through the load. Before C_1 can completely discharge, the next cycle of the input AC starts, recharging C_1 to its peak voltage. Thus, C_1 supplies current to the load when the input AC is cut off.
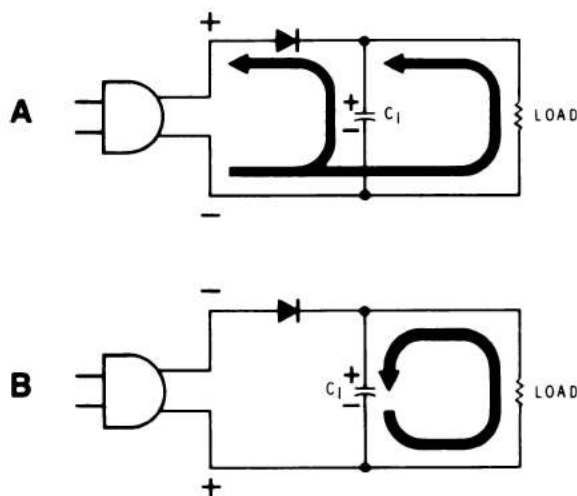
Figure 4-7
A capacitor power supply filter.

Figure 4-8A shows the output voltage waveform for an unfiltered half-wave rectifier. Figure 4-8B shows the output voltage when a small value capacitor is added. Note that the capacitor attempts to hold the voltage at a constant level. This is because of the basic nature of a capacitor to oppose changes in voltage, which it does by storing an electrical charge.

Figure 4-8C shows the output voltage when a much larger capacitor is used. The larger capacitor can store a greater electrical charge and, therefore, has a greater opposition to voltage changes.

Figure 4-8D shows the output of a full-wave or bridge rectifier with a capacitor filter. It is readily apparent that a full-wave rectified voltage is much easier to filter than a half-wave rectified voltage. This is due to the higher ripple frequency of the full-wave voltage.
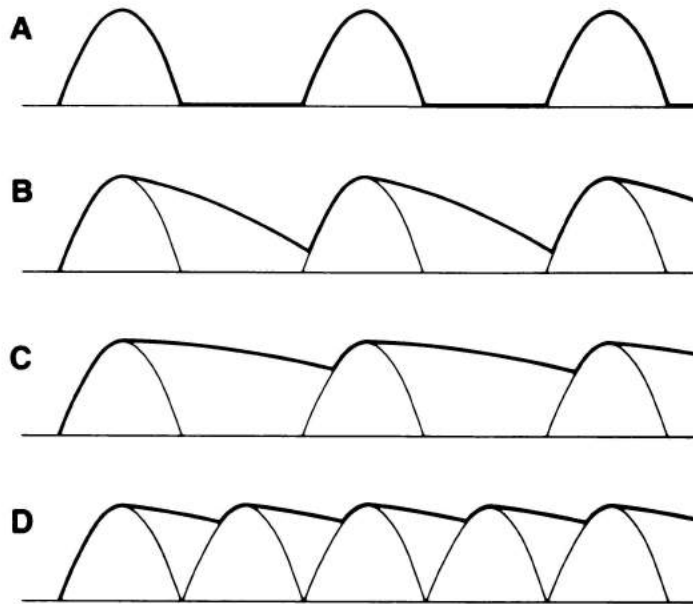
Figure 4-8

Rectifier output waveforms with varying degrees of capacitive filtering.

THE CAPACITOR'S EFFECT ON THE DIODES

Figure 4-9A shows a half-wave rectifier with a capacitor as a filter. During the half cycle in which the diode conducts, C_1 charges to the peak of the secondary voltage. Assume that this is +200 volts. The capacitor is large enough to hold the voltage across the load at approximately this level throughout the cycle.

During the next half cycle, the voltage across the secondary reverses. At the peak of the negative cycle, the anode of the diode is at −200 volts with respect to ground. At this point, the difference of potential across the diode is twice the peak value of the secondary, or 400 volts. Thus, the diode must have a peak inverse voltage (PIV) rating of at least 400V.

Figure 4-9B shows that a similar situation exists in the full-wave rectifier. With C_1 charged to the positive peak and the top of the secondary at the negative peak, D_1 experiences twice the peak secondary voltage. Here again, a suitably rated diode must be used.

Figure 4-9C shows a bridge rectifier. At first glance, it appears that D_2 is being subjected to a PIV of 400V. However, a closer look shows that D_3 is forward biased. Thus, point A is effectively at ground potential and the voltage across D_2 is only 200V, or the peak secondary voltage. The bridge rectifier therefore offers yet another advantage; diodes with lower PIV ratings can be used.
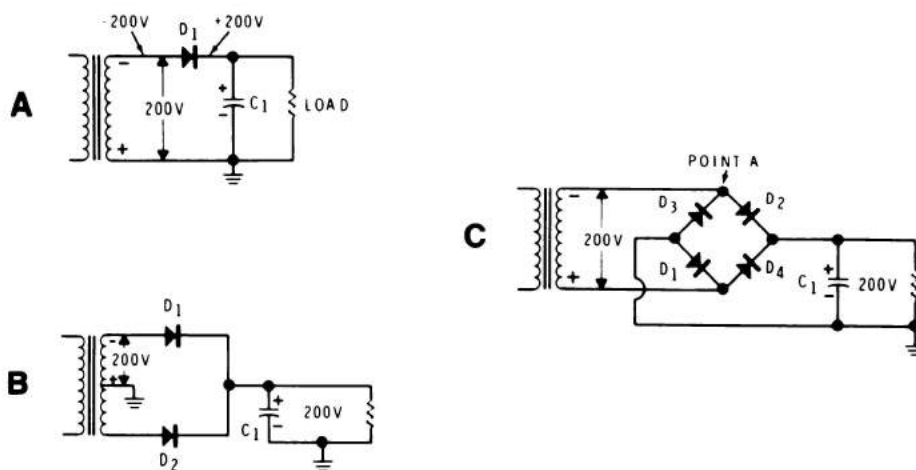
Figure 4-9
Determining the peak inverse voltage of the diodes.

When higher PIV ratings than a single diode can supply are required, several diodes can be connected in series. Then the inverse voltage will divide evenly between each diode, and the total PIV rating will be increased by the number of diodes in series.

As an example, Figure 4-10A shows three 1,000V PIV rated diodes connected in series. The PIV rating for this combination is 1,000V × 3 or 3,000V.

When you connect diodes in series, their reverse bias characteristics should match. That is, each diode's reverse "leakage" current should be the same. When this is not the case, unequal voltages will be dropped across the diodes and one diode may be subjected to a PIV greater than its maximum rating. To prevent this, **equalizing resistors** are connected in parallel with the diodes as shown in Figure 4-10B. Thus, the equalizing resistors force the reverse voltage drops to be equal.

Figure 4-10C shows an additional component which may be used on a diode "stack." The 0.01μF capacitors are used for "transient" protection. These transients, which are caused by DC switching at the load — or other sudden changes, are large voltage "spikes" that are often many times greater than the PIV rating of the diodes. These capacitors equalize and absorb any transients uniformly along the diode stack.
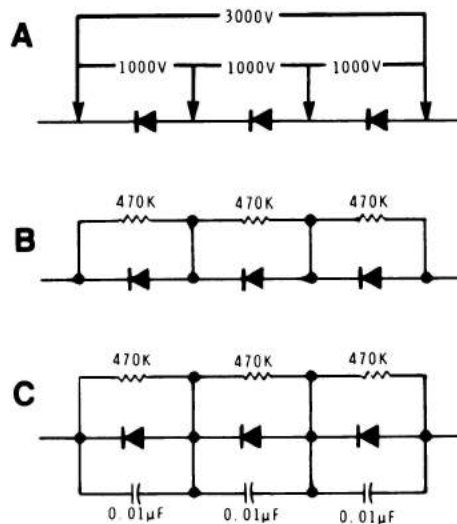
Figure 4-10
Diode protection circuitry.

ELECTROLYTIC CAPACITORS

An electrolytic capacitor is specially constructed to provide a large amount of capacitance in a small space. This makes it ideally suited for power supply circuits.

The construction of an electrolytic capacitor is shown in Figure 4-11. Sheets of metal foil are separated by a sheet of paper that is saturated with a conductive chemical paste called an electrolyte. The nonconductive dielectric is formed during the manufacturing process when a DC voltage is applied across the metal foil plates. As the DC current flows, an extremely thin layer of aluminum oxide (see Detail 4-11C) builds up on the plate connected to the positive side of the DC voltage. The upper foil then becomes the positive plate of the capacitor. The oxide, because it is a good insulator, becomes the dielectric, and the electrolyte becomes the negative plate. Note that the capacitor is **polarized**. Figure 4-11B shows how the positive lead is marked. When this type of capacitor is connected in a circuit, you must connect the positive lead to the more positive voltage. (NOTE: On some electrolytic capacitors, only the **negative** lead may be marked.)
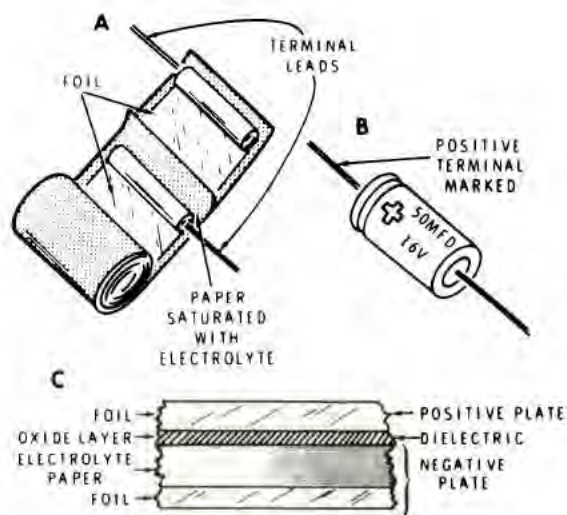
Figure 4-11
The electrolytic capacitor.

Because capacitance is inversely proportional to plate spacing, an electrolytic capacitor can have a very large capacitance due to its very thin dielectric. However, since the dieletric is so thin, a high voltage can cause the dielectric to arc over or break down. This is why capacitors have a **working voltage rating**. This is the maximum voltage the capacitor can withstand without the dielectric breaking down or arcing over.

When higher working voltage ratings are required, several capacitors can be connected in series as shown in Figure 4-12. Here, the maximum voltage rating is 450 V × 3 or 1350 V. Note the 30 kΩ "equalizing"resistors which must be used to equalize the voltage drops across each capacitor.

A "stack" of capacitors with equalizing resistors is shown in Figure 4-13. This is the power supply for a 2,000 watt amateur radio linear amplifier. The working voltage rating of the capacitor stack is 3,600 V!
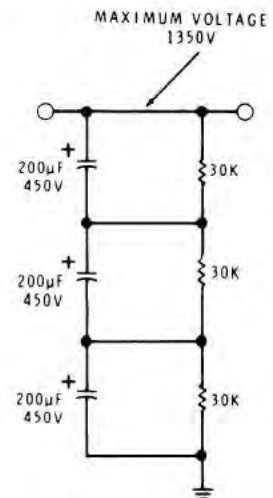
Figure 4-12

Using capacitors in series to increase the voltage rating.
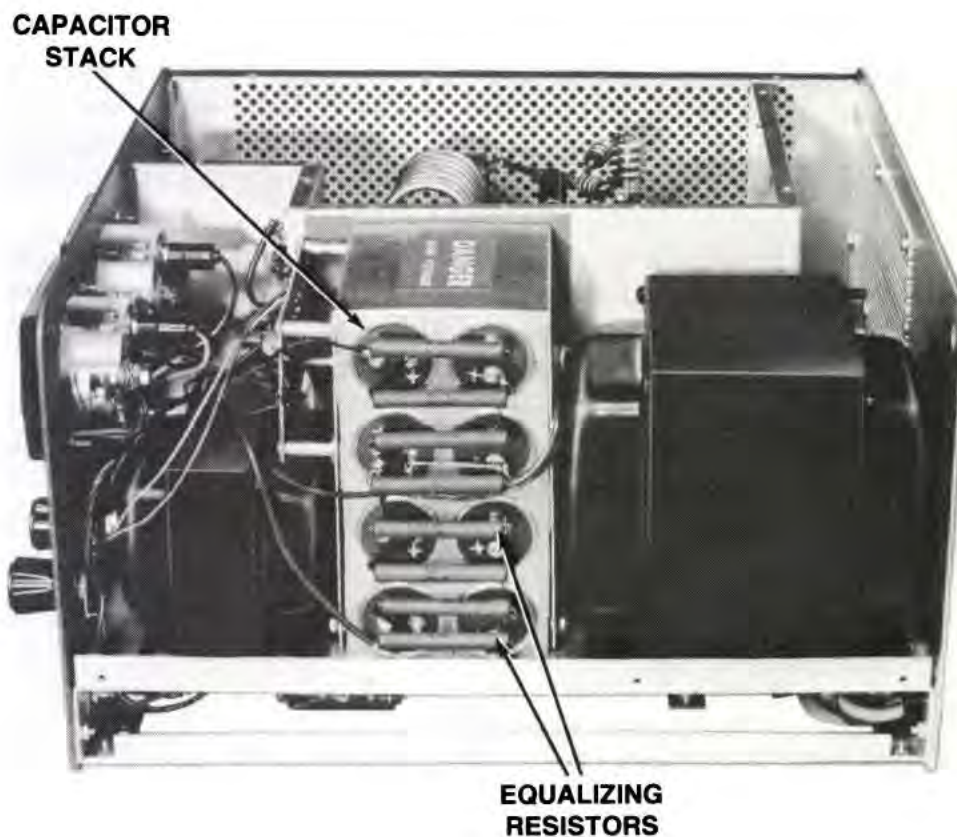
Figure 4-13

Capacitor stack in a high-power amplifier power supply.

PI FILTERS

Figure 4-14 shows a more elaborate type of power supply filter, the LC pi filter. This name is used because, in its schematic diagram, the circuit components are arranged to resemble the Greek letter pi (π). You may have recognized it as a low-pass filter. In this case, its cut off frequency is approximately 100 Hz. That is, it passes all frequencies below 100 Hz. Since the ripple frequency of the full-wave rectifier is 120 Hz, only DC is allowed to pass to the load.

Figure 4-14
Full-wave rectifier with LC filter.
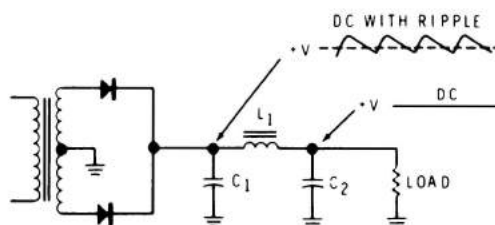
One disadvantage of the LC pi filter is the size, weight, and cost of the inductor. To obtain a cutoff frequency of 100 Hz, a large value inductor must be used along with relatively large capacitors. For a half-wave rectifier, the cutoff frequency must be 40-50 Hz to filter out the 60 Hz ripple component. This requires still larger inductors and capacitors.

The RC pi filter shown in Figure 4-15 works on the same low-pass filter principle and is a lot less expensive. However, its filter action is not as good as the LC filter. Another disadvantage is that the resistor drops voltage which could be used by the load.

Nonetheless, the RC filter is a good compromise and is found in a large number of power supplies.

Figure 4-15
Half-wave rectifier with RC filter.
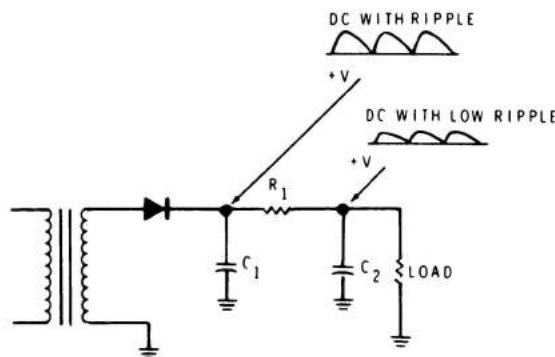
Voltage Doublers

Two rectifier circuits have been developed that can double the voltage from a transformer or the AC line. These are the half-wave and full-wave voltage doublers.

HALF-WAVE VOLTAGE DOUBLER

The half-wave voltage doubler, as shown in Figure 4-16A, consists of two diodes and two capacitors. It produces a DC output voltage which is approximately twice the peak value of the input AC sine wave. In this example, assume that the input is 115 VAC at 60 Hz.

Figure 4-16B shows how the circuit responds to the negative half cycle of the input sine wave. When the voltage at point A swings negative, D_1 conducts and current flows as shown by the arrows. This charges C_1, with the polarity shown, to the peak value of the input sine wave (about 162 volts). Since there is no immediate discharge path for C_1, the capacitor remains charged to this level until the next half cycle.

Figure 4-16C shows what happens on the positive half-cycle. At the peak of the cycle, point A is at + 162 V. Since C_1 is charged to 162 V, and the voltages are in series, they add: 162 V + 162 V = 324 V. Therefore, the voltage at point B is +324 V. This forward biases D_2 and current flows as shown, charging C_2 to +324 V. Thus, the load sees twice the peak value of the input AC, or 324 V.

Between the positive peaks of the sine waves, D_2 is cut off because C_2 holds its cathode at a high positive potential. During this period, C_2 discharges through the load. Thus, C_2 acts as a filter capacitor holding the voltage across the load fairly constant. Since C_2 is recharged only during the positive peak of the input sine wave, the ripple frequency is 60 Hz. Consequently, this circuit is called a half-wave voltage doubler.
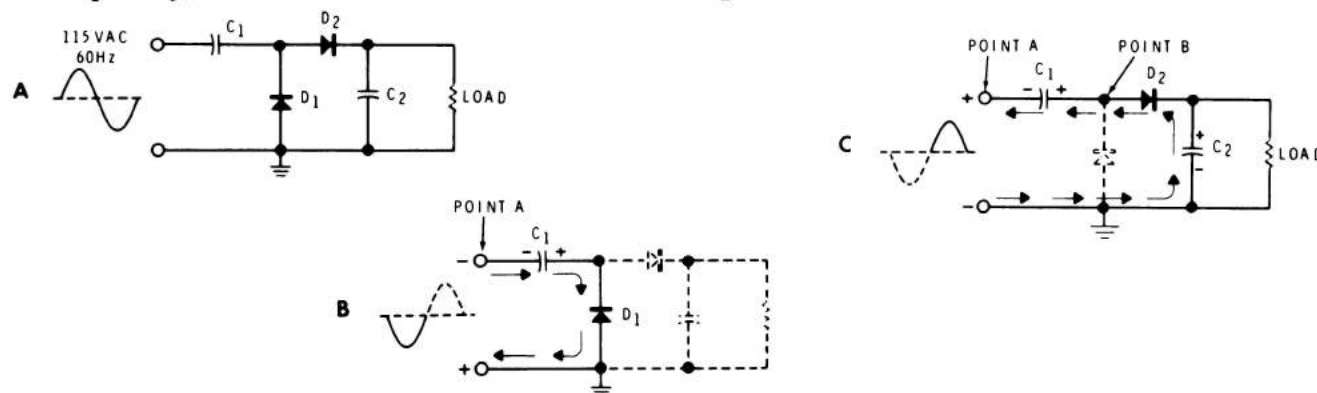
Figure 4-16
Half-wave voltage doubler.

FULL-WAVE VOLTAGE DOUBLER

The circuit for the full-wave voltage doubler is shown in Figure 4-17A. On the positive half cycle, C_1 charges through D_1 to the peak value of the AC input, as shown in Figure 4-17B. In this case C_1 charges to about 162 V.

On the negative half cycle, C_2 charges through D_2 as shown in Figure 4-17C. C_2 charges to the peak value of the AC input, or to about 162 V.

Once the capacitors are charged, D_1 and D_2 conduct only at the peaks of the AC input. Between the peaks, C_1 and C_2 discharge in series through the load as shown in Figure 4-17D. If each capacitor is charged to 162 V, then the total voltage across the load is about 324 V, or twice the peak value of the input AC. Here, the ripple frequency is 120 Hz, since the C_1-C_2 combination is recharged twice during each cycle.
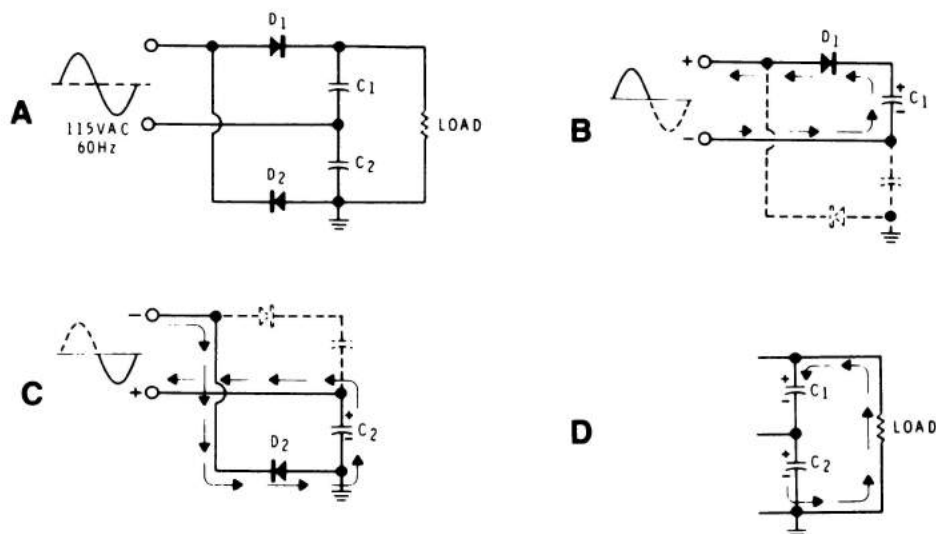
Figure 4-17
Full-wave voltage doubler.

Voltage Regulation

Ideally, the output of a power supply should be a constant voltage. Unfortunately, this is difficult to achieve. There are two factors which can cause the output voltage to change.

First, the AC line voltage is not constant. The so-called 115 VAC can vary from about 105 VAC to 125 VAC. This means that the peak AC voltage to which the rectifier responds can vary from about 148 V to 177 V. As you can see, the AC line voltage can cause as much as a 20% change in the DC output voltage.

The second thing that can change the DC output voltage is a change in the load resistance. In electronic equipment, the load can change as circuits are switched in and out. For example, in a TV Receiver, the load changes each time the video changes from white to black.

Variations in load resistance tend to change the applied DC voltage because the power supply filter capacitors are discharged more quickly. One way to decrease voltage variations is to use a bleeder resistor.

BLEEDER RESISTOR

Most power supplies use a bleeder resistor across the output, as shown in Figure 4-18. The bleeder serves two purposes. First, it discharges the filter capacitor when the power supply is turned off. This is a safety precaution, as filter capacitors can hold a charge for a very long time. Therefore, even though the power supply is turned off, they can present a very dangerous shock hazard. The bleeder eliminates this hazard by discharging the capacitors.
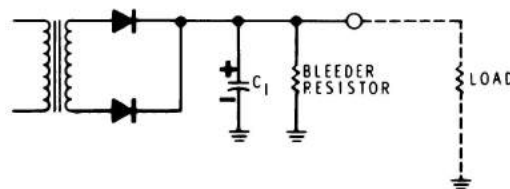
Figure 4-18
Using a bleeder resistor to improve voltage regulation.

The second purpose of the bleeder resistor is to reduce output voltage fluctuations. Figure 4-19 is a graph showing the effect of current drain on the power supply's output voltage. Note that when current is zero, the output voltage is 162 V, or the peak of the input AC. This is because the capacitors charge to the peak and hold it there. When current is drawn from the supply, the capacitors discharge and the output voltage drops rapidly, leveling off at about 120 V. This is a drop of 42 V, or a 35% change.

If a bleeder resistor that draws about 100 mA is used, the "no-load" output voltage will be approximately 130 V. Then, when full current is drawn from the supply, the voltage drops to 120 V. This is a drop of only 10 V, or an 8% change. Thus, the bleeder resistor has improved the power supply's **regulation**. That is, its no-load voltage is much closer to its full load voltage and output variations have been reduced.

Figure 4-19
Effect of current drain on output voltage.
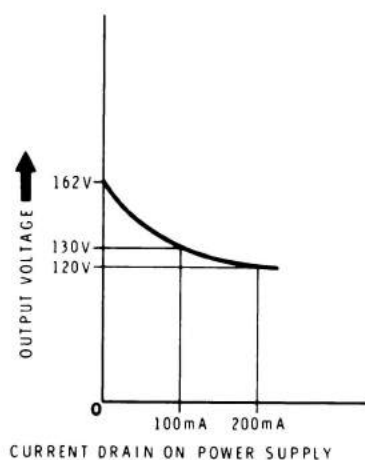
ZENER VOLTAGE REGULATOR

Ideally, the difference between no load and full load voltage would be zero. A circuit that comes close to accomplishing this ideal condition is known as a **voltage regulator**.

In an earlier unit, you studied the zener diode which, when reverse biased into breakdown, has a constant voltage drop. Thus, it can be used as a simple voltage regulator.

A zener diode regulator circuit is shown in Figure 4-20. Notice that the diode is connected in series with a resistor and that an unregulated DC input voltage is applied to these two components. The input voltage is connected so that the zener diode is reverse biased. The series resistor allows enough current to flow through the diode so that the device operates within its zener breakdown region.

In order for this circuit to function properly the input DC voltage must be higher than the zener breakdown voltage. The voltage across the diode will then be equal to the diode's zener voltage rating. The voltage across the resistor will be equal to the difference between the diode's zener voltage and the input DC voltage.

The input DC voltage is unregulated. This voltage will periodically increase above or decrease below its specified value. Therefore, it causes the current flowing through the zener diode and the series resistor to fluctuate. However, the diode is operating within its zener voltage region. Therefore, a wide range of current can flow through the diode while its zener voltage changes only slightly. Since the diode's voltage remains almost constant as the input voltage varies, the change in input voltage appears across the series resistor. Remember that these two components are in series and the sum of their voltage drops must always be equal to the input voltage.

The voltage across the zener diode is used as the output voltage for the regulator circuit. The output voltage is therefore equal to the diode's zener voltage. Since this voltage is held to a nearly constant value, it is referred to as a regulated voltage. You can change the output voltage of this regulator circuit by using a diode with a different zener voltage rating.
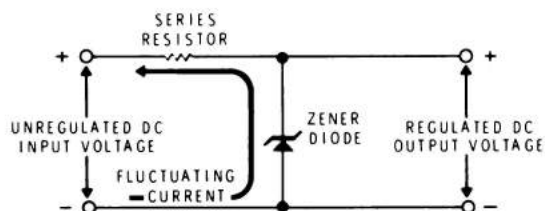
Figure 4-20
Basic zener regulator.

IC VOLTAGE REGULATORS

The trend in power supply design is toward integrated circuit (IC) regulators. These are available with a wide range of output voltages and currents. Generally, a single IC will contain a complex regulator circuit, and additional protection circuitry which makes it virtually blowout proof.

Some IC regulators are designed to deliver a fixed output voltage such as +5V, +12V, or −12V. However, others can be set up with external components to produce a range of voltages or even an adjustable voltage.

Some of the low voltage regulators can deliver 1A or more to the load without external components. Many regulators, though, are limited to 200 mA or below. If higher currents are required, additional components must be added externally.

Figure 4-21 shows a complete power supply that uses a 5V IC regulator. The regulator has only three terminals and it uses a transistor package. However, the IC contains over a dozen transistors arranged in a complex regulator circuit. Regulator IC's are therefore classified as linear IC's.

An unregulated DC voltage is supplied to the input terminal of the regulator. This voltage can be any value from about +10V to +35V. In the circuit shown, the bridge rectifier supplies +12.5V, which is filtered by C_1. The IC regulator accepts this unregulated input and produces a precise, regulated, +5V output. The circuit is protected by the 3/16 A slow-blow fuse, F1, as well as the IC's internal protection circuitry. Switch SW_1 is the on-off switch.
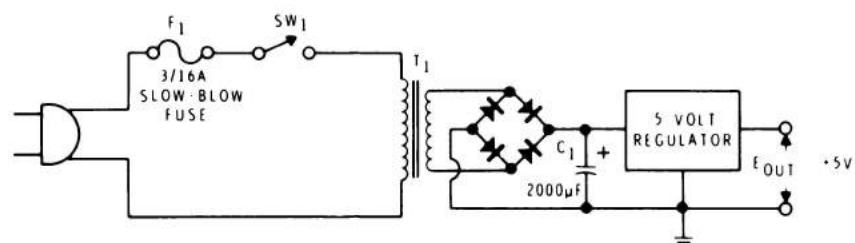
Figure 4-21
5-volt power supply using an IC voltage regulator.

Self-Review Questions

1. Identify the circuits shown in Figure 4-22.

    A. _____ .

    B. _____ .

    C. _____ .

2. What is the output ripple frequency and DC voltage polarity of the circuit shown in Figure 4-22A? _____
    In Figure 4-22B?_____
    In Figure 4-22C?_____
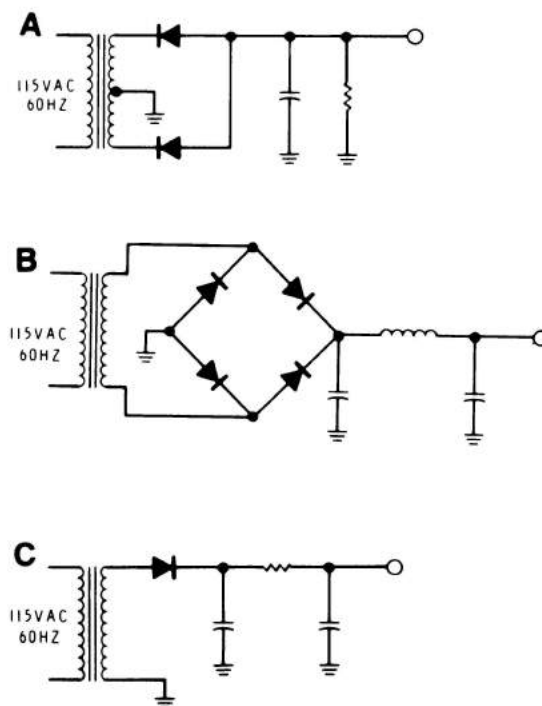
Figure 4-22
Identify these circuits.

3. What is the purpose of a power supply filter? _____
    _____
    _____

4.   What is the minimum PIV rating for D_1 in Figure 4-23? _____
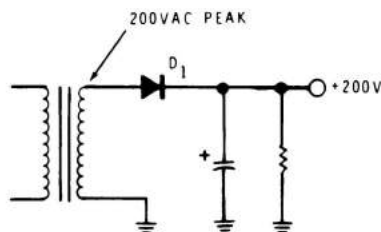
Figure 4-23
What is the PIV rating for D_1?

5.   Why are electrolytic capacitors used in power supply filters? __

6.   Identify the circuit shown in Figure 4-24. _____
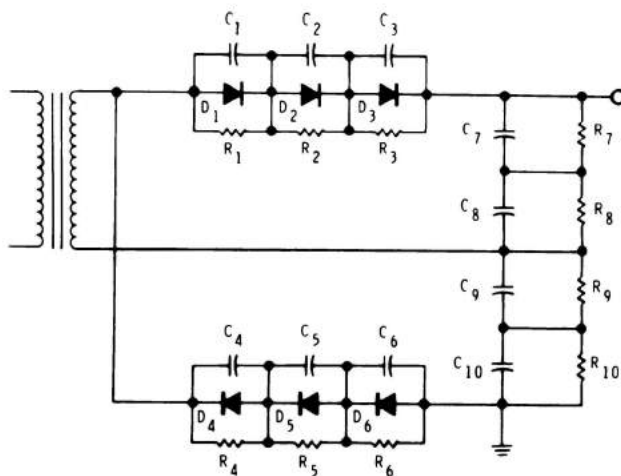
Figure 4-24
Identify this circuit.

7.   What is the purpose of connecting D_1, D_2 and D_3 in series in Figure 4-24? _____

8.   What is the purpose of C_1-C_3 and R_1-R_3 in Figure 4-24? _____

9. What is the purpose of connecting C_7 and C_8 in series in Figure 4-24? _____

    _____

10. State the two purposes of a bleeder resistor. _____

    _____

11. Identify the circuits in Figure 4-25.

    A. _____.

    B. _____.
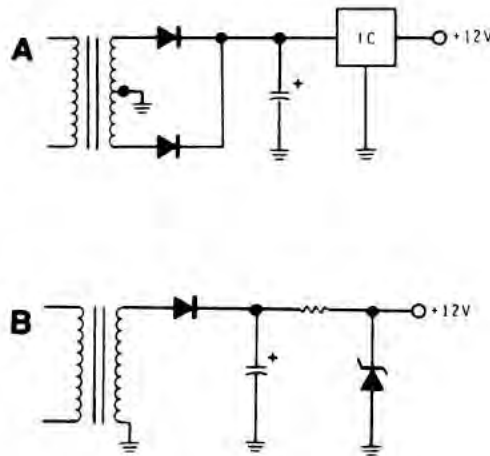
Figure 4-25
Identify these circuits.

Self-Review Answers

1. The circuits shown in Figure 4-22 are:

   A. Full-wave rectifier, capacitor filter.
   B. Bridge rectifier, LC pi filter.
   C. Half-wave rectifier, RC pi filter.

2. Figure 4-22A is a full-wave rectifier. Since the input is 60 Hz, the ripple frequency is 60 × 2 or 120 Hz. The output polarity is determined by the diode connections, in this case it is negative.

   Figure 4-22B is a bridge rectifier, so ripple frequency is 2 × 60 Hz = 120 Hz. Output polarity is positive.

   Figure 4-22C is a half-wave rectifier, so ripple frequency is 60 Hz. Output polarity is positive.

3. A power supply filter converts the pulsating DC from the rectifier into a smooth DC voltage.

4. Since Figure 4-23 has a capacitive filter, the PIV is twice the peak secondary voltage: 200 V × 2 = 400 V.

5. Electrolytic capacitors are used in power supply filters because they provide a large amount of capacitance in a small space.

6. The circuit shown in Figure 4-24 is a full-wave voltage doubler.

7. D_1, D_2, and D_3 are connected in series to increase the PIV rating.

8. The purpose of C_1-C_3 and R_1-R_3 in Figure 4-24 is diode protection. C_1-C_3 provide transient protection, while R_1-R_3 equalize the inverse voltage across the diode.

9. The purpose of connecting C_7 and C_8 in series is to increase the working voltage rating.

10. The purposes of a bleeder resistor are to discharge the filter capacitors when the power supply is turned off and to improve the voltage regulation of the supply.

11. The circuits shown in Figure 4-25 are :

    A. IC voltage regulator.
    B. Zener diode regulator.

AUDIO AMPLIFIERS

There are many applications where it is necessary to amplify AC signals that are within the audio frequency range, which extends from 20 Hz to 20,000 Hz. The circuits used to amplify these signals are generally referred to as audio frequency (AF) amplifiers, or simply audio amplifiers.

Voltage and Power Amplifiers

Audio amplifiers provide both voltage and power amplification, or gain. Many amplifiers are specifically designed to amplify low voltage audio signals and are classified as **voltage amplifiers**. In most cases, voltage amplifiers are used to increase the voltage level of an input signal to a value high enough to drive a power amplifier stage. The **power amplifier** can then supply a high output signal current to operate a loudspeaker or some other device requiring high power.

Amplifier Gain and the Decibel

As you learned in a previous unit, amplifier gain is the ratio of input to output. Therefore, amplifier power gain is:

$$A_p = \frac{P_{out}}{P_{in}}$$

Where A_p =   power gain

P_{out} =   output power

P_{in} =   input power

One of the most common ways to express gain or power ratio is with a unit of measurement called the **decibel**, which is abbreviated dB. Originally, the decibel was based on the response of the human ear to sound. At lower sound levels, the ear can detect relatively small changes. However, as the volume is increased, larger and larger sound level changes are required before the ear can detect any difference. The decibel is a unit of measurement for the smallest change that the ear can detect.

Starting with a barely discernable sound level, one milliwatt, for example, we will increase the volume in barely discernable steps and see, in Figure 4-26, how much power is required. Remember that in this case, our starting level, 0 dB, is equal to 1 milliwatt (1 mW).

A power change of approximately 26% is required before your ear can detect any change in sound level. Therefore, the first point where you would notice a change would be at the 1.26 mW level. And since 1 dB is the minimum discernable change, we have labeled this 1 on the horizontal axis.

Now, if you were to increase the sound level further until you could detect another change, it would require 26% more power than this 1.26 mW level, or approximately 1.6 mW. This is the 2 dB point.

At the 3 dB point, the sound level is up to 2 mW, or twice the original power. And as you look further up the curve on the graph, notice that a discernable (1 dB) change requires more and more power. This is called a **logarithmic** change. If you are not familiar with logarithms or would like a review, read "Appendix A."

dB	MW
0	1.0
1	1.26
2	1.6
3	2.0
4	2.5
5	3.2
6	4.0
7	5.0
8	6.3
9	7.9
10	10.0
11	12.5
12	15.8
13	20.0
14	25.1
15	31.6
16	39.8
17	50.1
18	63.1
19	79.4
20	100.0
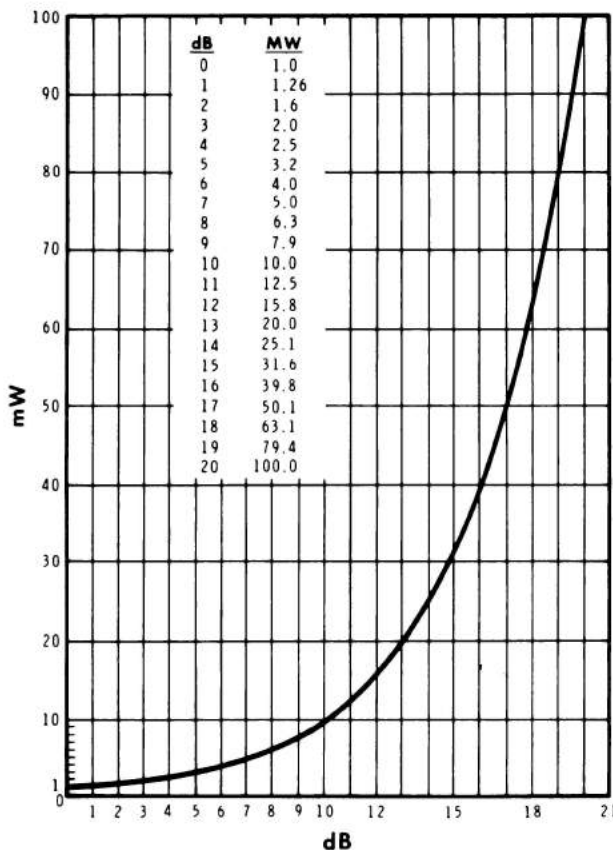
Figure 4-26
Decibels referenced to 1 milliwatt.

In our graph, you found that a power change from 1 mW to 2 mW was a 3 dB change. Notice also that a change in power from 10 mW to 20 mW is a change of 10 dB to 13 dB, a 3 dB change. Likewise, a change from 50 mW to 100 mW is a 3 dB change. Thus, you can see that dB is a measure of change, not a measure of actual power.

The formula for calculating decibels is:

$$dB = 10 \times \log \frac{P_2}{P_1}$$

To use the formula, follow this procedure:

1. Find the power ratio, making sure to use the same units for both P_1 and P_2. Always use the larger power level as P_2. Then the ratio will always be greater than 1, and you will avoid working with negative logarithms.

2. Find the logarithm of the power ratio.

3. Multiply the logarithm by 10 to find the number of decibels.

4. If the ratio represents an increase, the answer is a positive dB. If the ratio represents a loss, the answer is a negative dB.

Now let's try some examples.

The input to an amplifier is 1 mW and the output is 1W. What is the gain in dB?

$$dB = 10 \times \log \frac{P_2}{P_1}$$

$$dB = 10 \times \log \frac{1 \text{ W}}{1 \text{ mW}}$$

$$dB = 10 \times \log \frac{1 W}{0.001 \text{ W}}$$

$$dB = 10 \times \log \quad 1000$$

$$dB = 10 \times 3$$

$$dB = 30$$

Since there is an increase in power, the gain of the amplifier is +30 dB.

What is the gain or loss of a resistor network whose input is 100 mW and output is 10 mW?

$$dB = 10 \times \log \frac{100 \text{ mW}}{10 \text{ mW}}$$

$$dB = 10 \times \log \quad 10$$

$$dB = 10 \times 1$$

$$dB = 10$$

Since there is a decrease in power, the loss of the network is −10 dB.

It is also possible to express a voltage or current gain in dB. Since power is $\frac{E^2}{R}$, you can express the decibel formula as:

$$dB = 20 \times \log \quad \frac{E_2}{E_1}$$

The two voltages must be across equal impedances or you cannot use this formula. The factor of 20 in the formula is because power is proportional to voltage squared, and doubling the logarithm corresponds to squaring the number.

As an example, for an increase from 1 mV to 1V, the dB voltage gain is:

$$dB = 20 \times \log \quad \frac{E_2}{E_1}$$

$$dB = 20 \times \log \quad \frac{1 \text{ V}}{0.001 \text{ V}}$$

$$dB = 20 \times \log \quad 1000$$

$$dB = 20 \times 3$$

$$dB = 60$$

The voltage gain is +60 dB.

The formula for current gain in dB is:

$$dB = 20 \times \log \frac{I_2}{I_1}$$

Note that the factor 20 also appears in this formula. This is because power is proportional to current squared.

A graph of voltage, current, and power ratios from 1 to 10 is shown in Figure 4-27. You can use this chart to convert power, voltage, or current gain to dB, or to convert dB to power, voltage, or current gain.
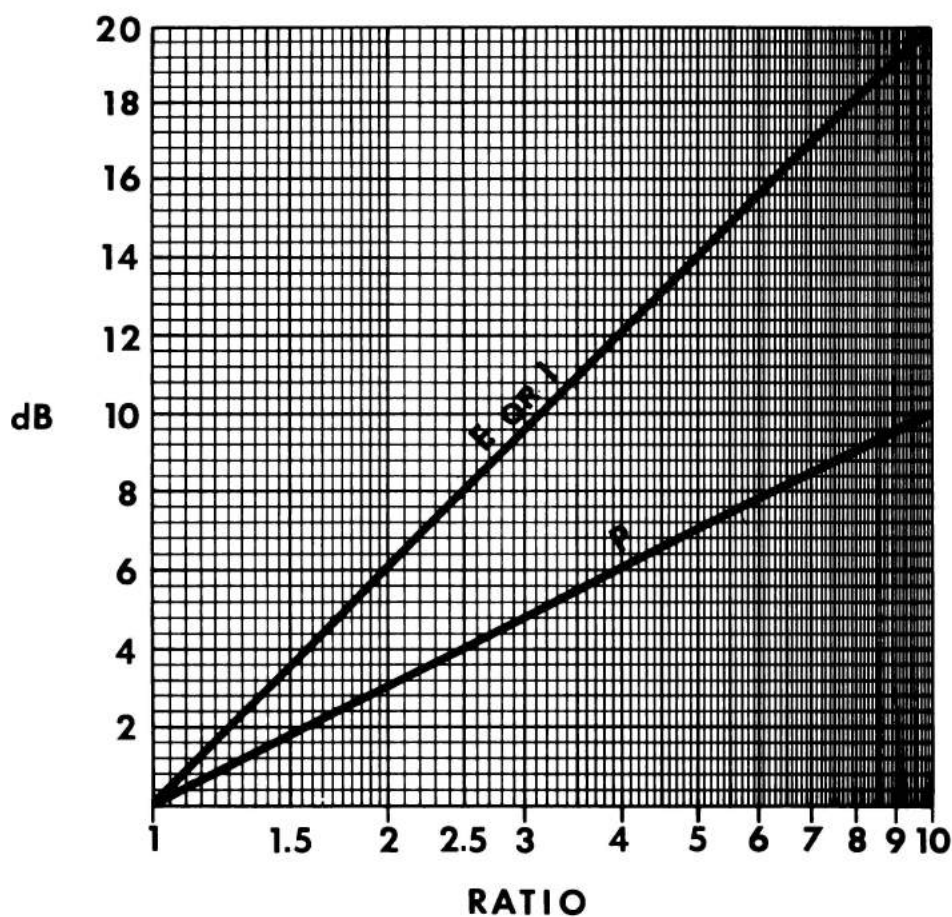
Figure 4-27
Current, voltage, and power ratios and their dB equivalents.

Coupling Circuits

In some applications, one amplifier stage (one transistor or tube) cannot provide enough amplification. Therefore, it becomes necessary to couple two or more amplifier stages together to obtain a higher overall gain.

When two amplifier stages are joined together, it must be done in a way which will not upset or disrupt the operation of either circuit. There are four basic coupling methods which are widely used. Let's examine each of these methods.

CAPACITIVE COUPLING

The capacitive, or RC, coupling technique is one of the most widely used methods. In this type of circuit, the signal must pass from one amplifier stage to the next through a coupling capacitor. The primary purpose of this coupling capacitor is to block DC voltage from the previous stage, and thus prevent bias changes. Figure 4-28 shows how transistor stages are capacitive coupled.

In order for the coupling capacitor to efficiently transfer the AC signal to the second stage, it must offer very little opposition to the AC signal current, even at the lowest frequency which must be amplified. Therefore, a coupling capacitor usually has a high capacitance value so its capacitive reactance will be very low. Typically the coupling capacitor is an electrolytic type and has a value of 10 μF or higher.
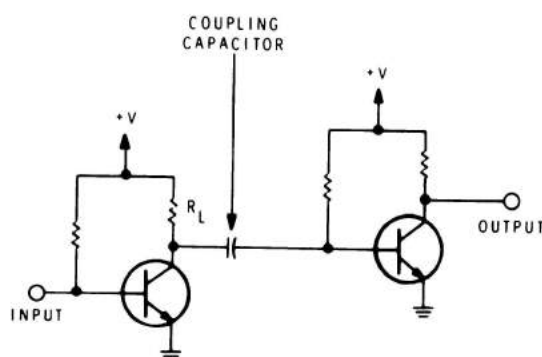
Figure 4-28
Capacitive coupling.

Capacitive-coupled amplifiers can provide amplification over a wide frequency range but there are definite upper and lower limits. The reactance of the coupling capacitor increases as frequency decreases. This means that the lower frequency limit is determined by the size of the coupling capacitor. If a higher capacitance value (thus a lower reactance) is used, it will extend the lower frequency limit. However, a lower limit of a few hertz is the best that can be obtained. The coupling capacitor will not pass DC, and it offers a large amount of reactance to extremely low frequencies.

The upper frequency limit of the capacitive-coupled stages is primarily determined by the transistor's capabilities. The gain of these devices decreases when the frequency extends beyond a certain point. Eventually, a point is reached where useful gain is no longer provided. This will be the upper frequency limit.

IMPEDANCE COUPLING

The impedance coupling technique is similar to the capacitive coupling method. However, with this method, an inductor is used in place of the the load resistor (R_L), as shown in Figure 4-29.

Impedance coupling works like capacitive coupling since the inductor (L) performs the same basic function as the load resistor, and the coupling capacitor transfers the signal from one stage to the next. However, the inductor has a very low DC resistance across its windings. The resistance of a typical inductor may be only a few hundred ohms, since the inductor is formed from low resistance wire wound around an iron core.
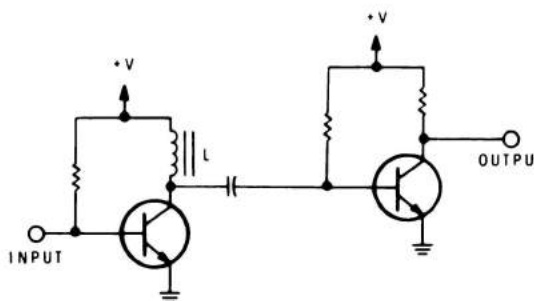
Figure 4-29
Impedance coupling.

Since the inductor has a low DC resistance, it drops only a small DC voltage. This means that the inductor itself consumes only a small amount of power. However, the inductor still offers a substantial amount of opposition to AC. Therefore, an AC signal voltage is developed across the inductor just like it would be across a load resistor; and most of this signal voltage is applied through the capacitor to the next transistor. The principle advantage that an inductor has over a resistor is that it consumes less power, thus increasing the overall efficiency of the circuit.

Unfortunately, the reactance of the inductor does not remain constant but increases with the signal frequency; and because the reactance of the inductor increases, so does the output signal voltage. This points out the main disadvantage of this type of amplifier coupling: the voltage gain gets higher as the signal frequency increases. This type of coupling is used principally in applications where only one signal frequency or a narrow range of frequencies must be amplified.

DIRECT COUPLING

Capacitive and impedance coupling techniques cannot be effectively used when very low frequencies must be amplified because the coupling capacitor cannot pass them. A technique known as "direct coupling" is used in low frequency and DC signal amplifiers.

A typical direct-coupled circuit is shown in Figure 4-30. Notice that the base of the second transistor is connected directly to the collector of the first transistor. No coupling capacitor is used between the stages.
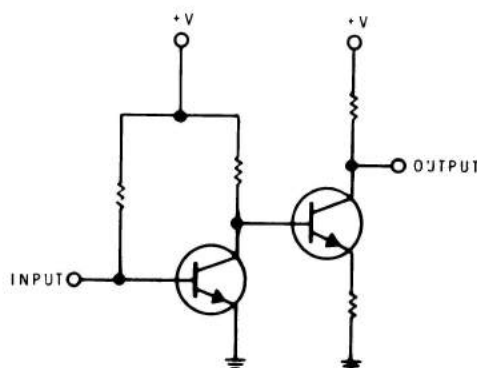
Figure 4-30
Direct coupling.

Directly coupled amplifiers provide a uniform current or voltage gain over a wide range of signal frequencies. These amplifiers may be used to amplify frequencies that range from zero (DC) to many thousands of hertz. However, they are particularly suited to low frequency applications, due to the elimination of the coupling capacitor.

Unfortunately, direct-coupled circuits are not as stable as the capacitive and impedance-coupled circuits previously described. This is because the second stage is biased by the first stage. Any changes in the output current of the first stage due to temperature variations are amplified by the second stage. This means that the operating bias of the second stage can drift extensively if the circuit is not carefully designed. To obtain a circuit that has a high degree of temperature stability, it is often necessary to use expensive, precision components or special components to compensate for temperature changes.

TRANSFORMER COUPLING

A technique known as transformer coupling is also occasionally used to couple two amplifier stages. A typical transformer coupled circuit is shown in Figure 4-31. Note that the first stage is coupled to the second stage through a transformer.

The primary advantage of transformer coupling is its ability to match differing impedances. By varying the turns ratio, the circuit designer can achieve optimum coupling between the stages. Thus, maximum power transfer takes place.

However, transformer coupling does have several disadvantages. First, the transformer is large and heavy, and therefore expensive when compared to the price of a resistor or capacitor. Also, the transformer can only pass AC signals, not DC and then the frequency range is somewhat limited. Thus, transformer-coupled amplifiers are useful only over a relatively narrow range of frequencies.

Figure 4-31
Transformer coupling.
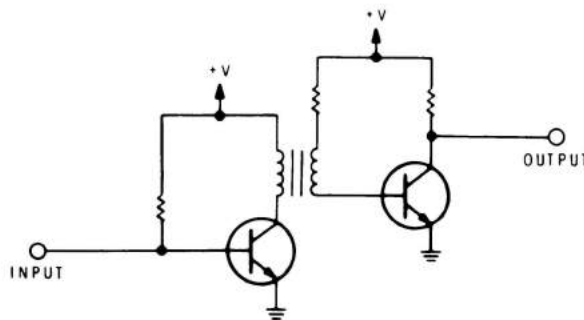
Classes of Operation

All of the amplifier circuits we have examined so far are biased so their output currents flow during the entire cycle of the input AC voltage. As the input signal goes through its positive and negative alternations, the output current increases or decreases accordingly, but the output current always continues to flow. Any amplifier that is biased in this manner is operating as a **Class A** amplifier.

The output current of a class A amplifier is shown in Figure 4-32. Note that the current increases and decreases around its no-signal value, but never drops to zero during one complete cycle. This means that the output voltage produced by this output current also varies in a similar manner.

To operate in the class A mode, the amplifier must be biased so its output current has a no-signal value that is midway between its upper and lower limits. The upper limit is reached when the transistor is conducting as hard as it can. At this time, the output current levels off to a maximum value and the device is said to be **saturated**. Saturation occurs if the input current or voltage is increased to a sufficiently high value. Once the saturation point is reached, the output current stops increasing even though the input may continue to rise.
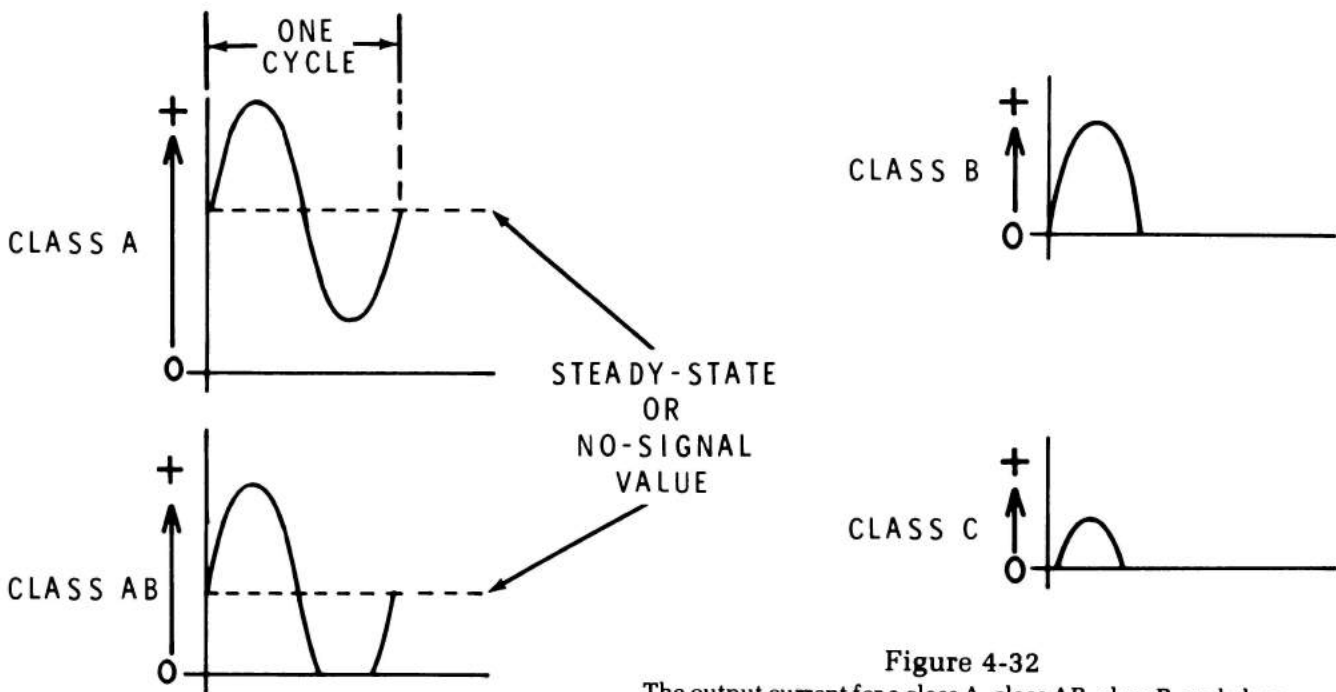
Figure 4-32
The output current for a class A, class AB, class B, and class C amplifier.

The lower limit of output current occurs when the input current or voltage is reduced so low that essentially no output current can flow. This condition is known as **cutoff**. With transistors, cutoff occurs when the input current reduces the forward bias to zero.

Therefore, class A operation is achieved by biasing the transistor midway between its saturation and cut off points. However, care must be exercised to prevent the input signal from overdriving the amplifier. If the input amplitude is so high that it causes the output current to reach the saturation and cutoff points, distortion results. If this occurs, the output signal is distorted as shown in Figure 4-33. Note that both positive and negative-going peaks are clipped off because the saturation and cutoff points are reached. However, this situation can be easily avoided if the proper input signal level is maintained.
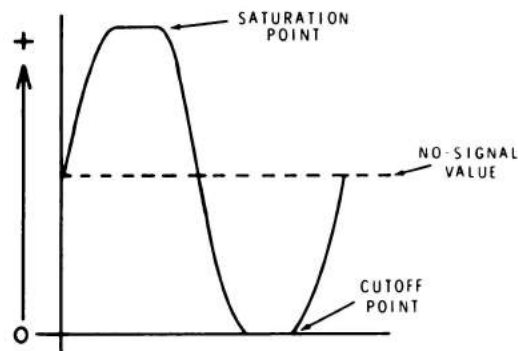
Figure 4-33
Output current distortion in a class A amplifier circuit.

When the amplifier is biased so its output current flows for less than one full cycle of the AC input signal, but for more than half of the AC cycle, the circuit is operating in the **class AB** mode. The output current waveform produced by a class AB amplifier is shown in Figure 4-32. This mode of operation occurs when the transistor is biased near the cutoff point.

You can also bias an amplifier so its output current will flow for only one half of the input AC cycle, as shown in Figure 4-32. Under these conditions, the circuit is operating in the **class B** mode. Class B operation occurs when the circuit is biased at cutoff. Under this condition, the output current can flow only on one input alteration. On the other alternation the circuit is driven further into cut off. Therefore, the class B amplifier amplifies only one half of the input AC signal.

An amplifier can also be biased so its output current flows for less than one half of the AC input cycle, as shown in Figure 4-32. Under this condition, the circuit is operating in the **class C** mode. Class C operation is obtained when the circuit is biased beyond the cutoff point. Therefore, the output current can flow only for a portion of one alternation. The voltage during this alternation must rise above cutoff before current can flow.

A class A amplifier produces an output signal that has the same basic shape and characteristics as the input signal, even though the amplitude of the output signal is higher. Class A circuits produce a minimum amount of distortion and, as a result, they are widely used in applications where a high degree of signal fidelity must be maintained.

Class AB, class B, and class C amplifiers produce a substantial amount of distortion, since these circuits amplify only a portion of the input signal. However, each of these amplifiers have specific applications in electronic equipment. They are often used in conjunction with other circuits or components that compensate for the distortion that they produce. In some applications, they are used intentionally to produce distortion in order to change the characteristics of a signal.

Amplifier Efficiency

Amplifier efficiency is a measure of how effectively an amplifier converts the DC input power (from the power supply) into AC output power. Expressed mathematically, it is:

$$\text{Efficiency in \%} = \frac{P_{out}}{P_{in}} \times 100$$

Let's try an example. What is the efficiency of a class A amplifier whose output is 10 watts while the input is 25 watts?

$$\text{Efficiency in \%} = \frac{P_{out}}{P_{in}} \times 100$$

$$\text{Efficiency in \%} = \frac{10 \text{ W}}{25 \text{ W}} \times 100$$

$$\text{Efficiency in \%} = 0.4 \times 100$$

$$\text{Efficiency} = 40\%$$

Thus, the amplifier's efficiency is 40%. That is, 40% of the DC input power is converted to AC power. Ideally, all the DC input power would be converted to AC output power for 100% efficiency.

While the class A amplifier produces the lowest distortion of any class of amplifier operation, it also has the lowest efficiency. This is due to the bias of the class A amplifier, which allows DC input current to flow whether there is an output signal or not. Therefore, the DC input power remains the same whether the amplifier produces full output or no output. Thus, the class A amplifier's efficiency will vary with the driving or input signal.

The class B amplifier's efficiency is approximately 50%. This is because it is biased at cutoff. Therefore, DC input current flows only when there is an AC input signal. Also, where there is an input signal, the amplifier is only on 50% of the time; the remainder of the time it is cut off. Thus, the class B amplifier is more efficient than the class A amplifier.

The class C amplifier has the highest efficiency of all the classes we have discussed. It is biased beyond cutoff, and the input signal must rise above cutoff to cause output current to flow. The efficiency of class C amplifiers can approach 80%.

Self-Review Questions

12. List the two types of audio amplifiers.

     1. _____

     2. _____

13. If the input to an amplifier is 1 mV and the output is 0.1V, what is the voltage gain in dB? _____
     _____

14. If the input to an attenuator network is 10 mW and the output is 5 mW, what is the attenuation (loss) in dB? _____
     _____

15. Identify the audio amplifier coupling circuits shown in Figure 4-34.

     A. _____
     B. _____
     C. _____
     D. _____
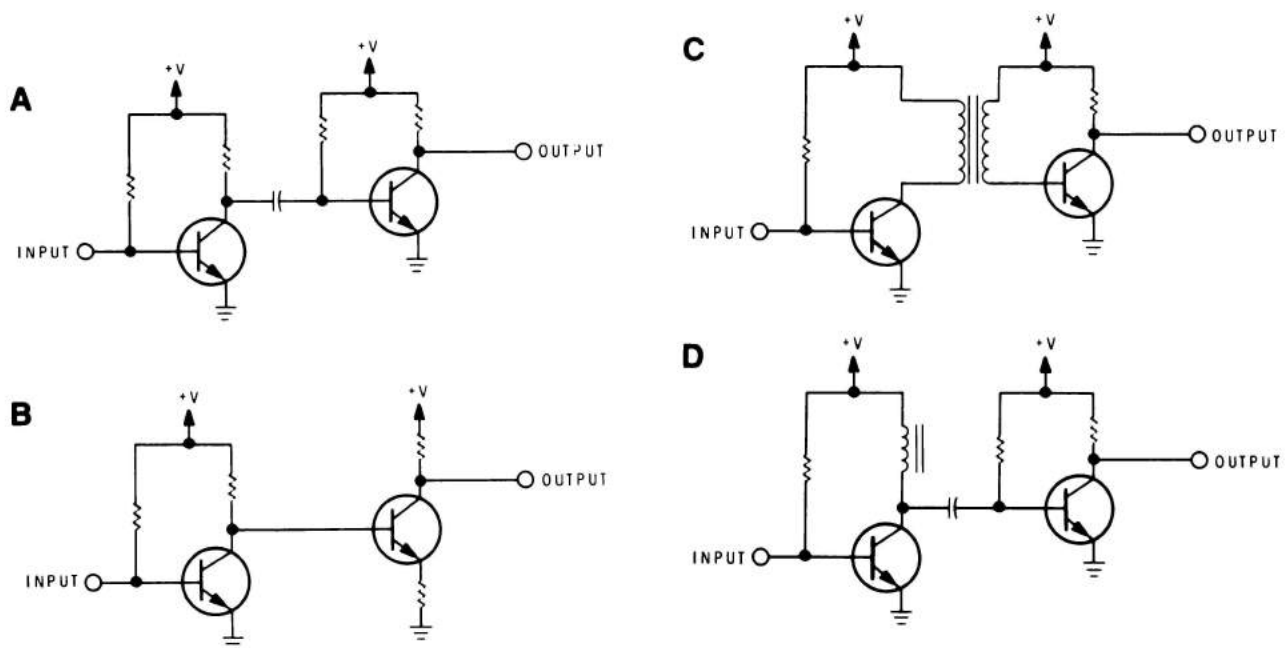
Figure 4-34
Identify these coupling circuits.

16. Draw the output current waveform, for a sine wave input, for the classes of amplifier operation indicated in Figure 4-35.

17. If DC input power to an amplifier is 50 watts and the AC output power is 10 watts, what is the amplifier's efficiency? _____

CLASS A

CUTOFF

CLASS B

CUTOFF

CLASS AB

CUTOFF

CLASS C

CUTOFF
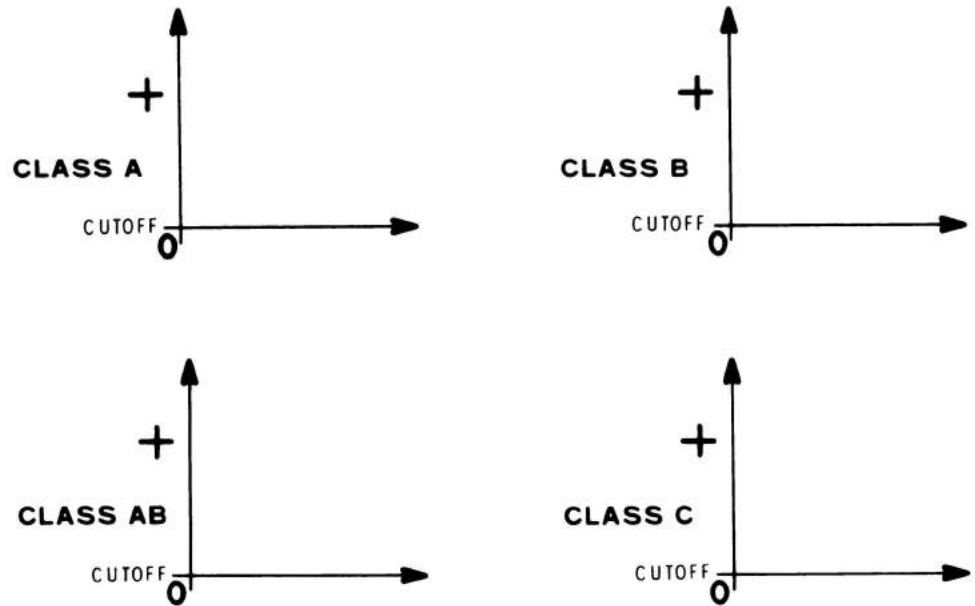
Figure 4-35
Draw the output current waveform for the classes of
amplifier operation indicated.

Self-Review Answers

12. The two types of audio amplifiers are:

   1. Voltage amplifier.
   2. Power amplifier.

13. $dB = 20 \times \log \dfrac{E_2}{E_1}$

   $dB = 20 \times \log \dfrac{0.1V}{1\ mV}$

   $dB = 20 \times \log \dfrac{100\ mV}{1\ mV}$

   $dB = 20 \times \log \quad 100$

   $dB = 20 \times 2$

   $dB = $ Voltage gain is +40 dB

14. $dB = 10 \times \log \dfrac{P_2}{P_1}$

   $dB = 10 \times \log \dfrac{10\ mW}{5\ mW}$

   $dB = 10 \times \log \quad 2$

   $dB = 10 \times 0.3010$

   $dB = 3.01$

   Attenuation is −3.01 dB

15. The audio amplifier coupling circuits shown in Figure 4-34 are:

   A. Capacitive coupling.
   B. Direct coupling.
   C. Transformer coupling.
   D. Impedance coupling.

16. The correct drawings are shown in Figure 4-36.

17. Efficiency in % $= \dfrac{P_{out}}{P_{in}} \times 100$

Efficiency in % $= \dfrac{10W}{50W} \times 100$

Efficiency in % $= 0.2 \times 100$

Efficiency $= 20\%$
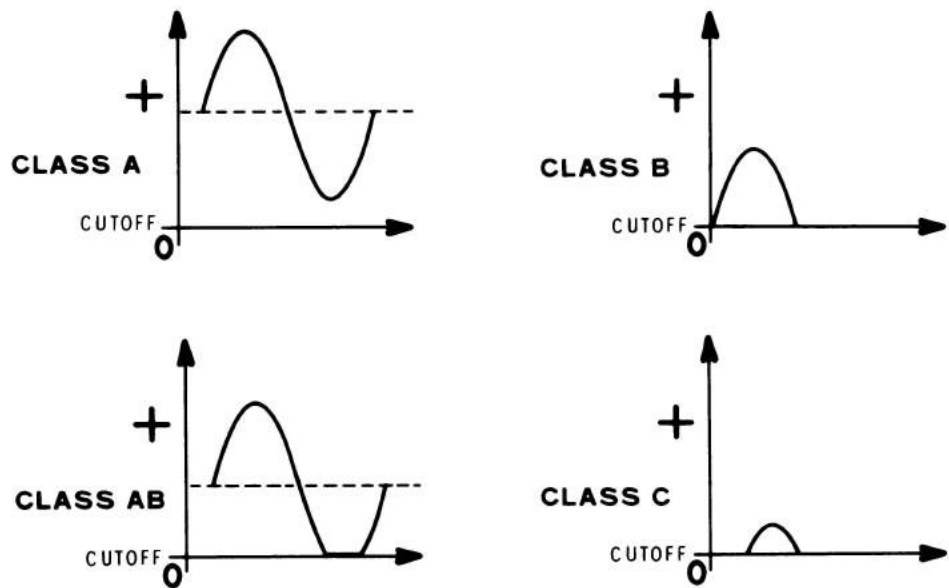
Figure 4-36
Answers to Question 16.

OPERATIONAL AMPLIFIERS

The incredibly fast development of IC technology has made the operational amplifier, or **op amp**, the most widely used and versatile of all electronic circuits. The op amp is basically an extremely high gain DC Amplifier. It was originally developed for use in "Analog" computers where it could perform a variety of functions by merely adding a few external components. In many ways, the op amp is still used in this manner. In this section, you'll discover not only what an op amp is but how it is used in several basic amplifier configurations.

The Basic Op Amp

The schematic symbol for the basic op amp is shown in Figure 4-37. Note that there are two input terminals. One is called the **inverting input** and is indicated by the minus (−) sign. The other input is called the **non-inverting input**, which is indicated by the plus (+) sign.

If a signal is applied to the inverting terminal with respect to ground, the output will be 180° out of phase with respect to the input. On the other hand, a signal applied to the non-inverting input will result in an output that is in phase with respect to the input. If signals are applied to both inputs, the output will be proportional to the difference of the two signals.
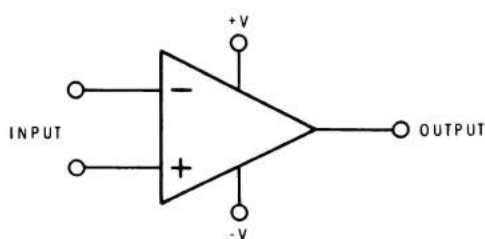
Figure 4-37
Schematic symbol for the basic op amp.

An op amp is basically a DC-coupled multistage linear amplifier. An ideal op amp would have all of the following characteristics:

- Infinite voltage gain

- Infinite input impedance

- Zero output impedance

As you might have guessed, these ideals are not reached in practice. However, op amp designs can come close to these ideals. For example, a typical op amp has a voltage gain of 200,000. Its input impedance is 2 MΩ and output impedance is 75 Ω. While these are not "ideal" characteristics, they are more than adequate for almost all applications.

An ideal op amp would also have zero output when both inputs are at zero. Any deviation from zero is called **offset**. Since a small offset usually occurs in practice, most op amps provide offset adjustment terminals. These and the other op amp terminals are shown in the IC schematic diagram of Figure 4-38. A potentiometer can be connected to the offset terminals to compensate for any offset voltage by adjusting the bias current to the input transistors.
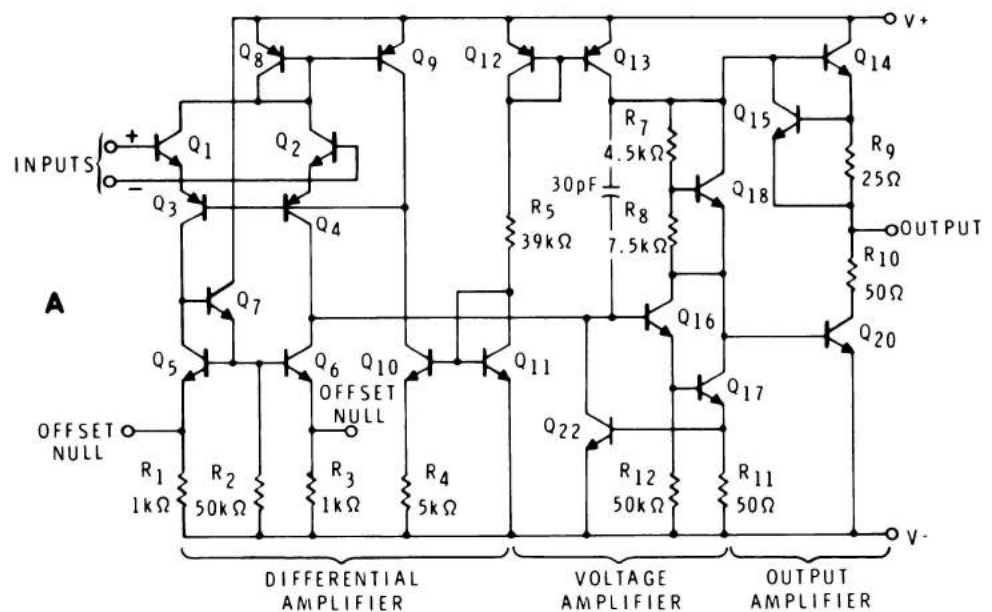
Figure 4-38
Internal circuitry of a typical op amp IC.

The Inverting Amplifier

Just as the transistor and FET have three basic circuit configurations, so too does the op amp. The inverting amplifier configuration is shown in Figure 4-39. In its most basic form, it consists of the op amp and two resistors. The noninverting input is grounded. The input signal (E_{in}) is applied through R_2 to the inverting input. The output signal is also applied back through R_1 to the inverting input. Therefore, the signal at the inverting input is determined not only by E_{in}, but also by E_{out}.
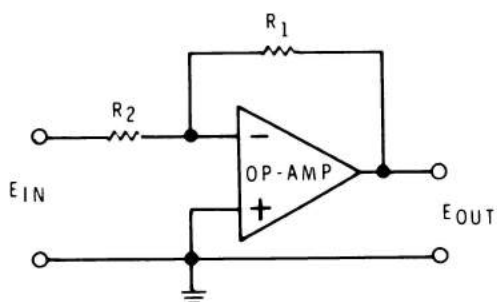
Figure 4-39
Inverting configuration. The power supply connections are not shown.

Resistor R_1 provides a **feedback** path. That is, a fraction of the output voltage is applied or "fed" back to the input. In this case, a feedback voltage is used not only to stabilize the circuit but, more importantly, to reduce the circuit voltage gain to a desired value. Since the output is 180° out of phase with the input, it subtracts from or reduces the signal present at the op amp's inverting input when it is fed back. This in turn decreases the output signal. This means that almost instantaneously, the output voltage, feedback voltage, and the voltage at the inverting input will stabilize at levels determined by the values of R_1 and R_2. Thus, by choosing the correct values of R_1 and R_2, the op amp circuit's voltage gain can be set exactly to the desired value.

The formula for the voltage gain of an inverting amplifier is:

$$A_r = \frac{R_1}{R_2}$$

Therefore, the voltage gain of the circuit shown in Figure 4-40 is:

$$A_r = \frac{R_1}{R_2} = \frac{100 \text{ k}\Omega}{10 \text{ k}\Omega}$$

$$= 10$$
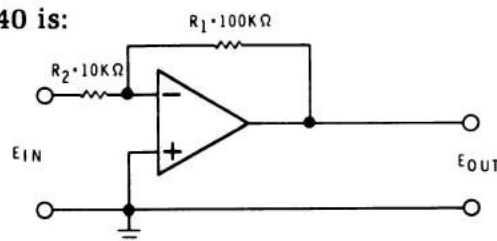
Figure 4-40
The gain of this stage is determined by R_1 and R_2.

The Noninverting Amplifier

The op amp can also be used as a noninverting amplifier. This configuration is shown in Figure 4-41. Since E_{in} is applied to the noninverting input, E_{out} is in phase with E_{in}. Therefore, in order for the feedback voltage to reduce the circuit gain, it must be applied to the inverting input. Thus, the output signal is actually the "difference" between the input signal applied to the noninverting input and the feedback signal applied to the inverting input.

Figure 4-41
The noninverting op amp configuration.
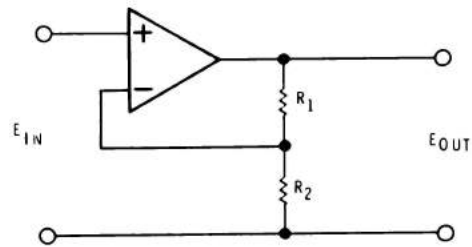
The feedback signal is obtained from the voltage divider network of R_1 and R_2. Therefore, the ratio of R_1 and R_2 determines the amount of feedback voltage and, thus, the circuit's voltage gain. In this case, the circuit's voltage gain is:

$$A_r = 1 + \frac{R_1}{R_2}$$

For example, the gain of the circuit shown in Figure 4-42 is:

$$A_r = 1 + \frac{R_1}{R_2} = 1 + \frac{100 \text{ k}\Omega}{20 \text{ k}\Omega}$$

$$= 1 + 5 = 6$$
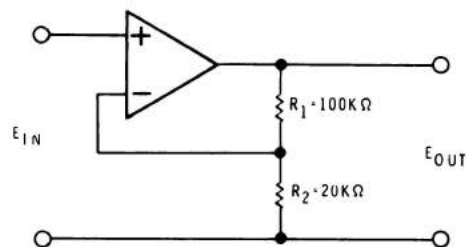
Figure 4-42
The gain of this stage is determined by R_1 and R_2.

The Voltage Follower

A special type of noninverting amplifier, a voltage follower, is shown in Figure 4-43. You will recall that the gain formula for a noninverting amplifier is:

$$A_r \;=\; 1 + \frac{R_1}{R_2}$$

In this case, $R_1 = 0$ and $R_2 = \infty$. Therefore, the voltage follower has a gain of 1. Thus, the output will be an exact replica of the input signal.

The input impedance of this circuit is extremely high, often as high as 400 MΩ, and the output impedance is very low, often less than 1Ω. Thus, the circuit behaves like a "super" emitter follower. And, like the emitter follower, it is used in impedance matching and isolation or buffering applications.

While there are many op amp circuits, almost all of them fall into the three basic circuits: inverting amplifier, noninverting amplifier, and voltage follower. This brief discussion should give you an excellent foundation for any further studies of operational amplifiers.
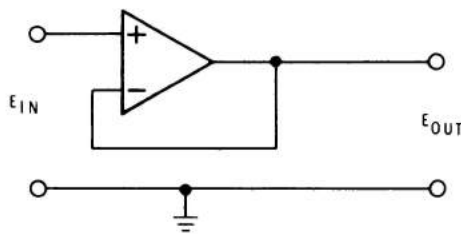
Figure 4-43
The voltage follower.

Self-Review Questions

18. Label the terminals of the op amp shown in Figure 4-44.
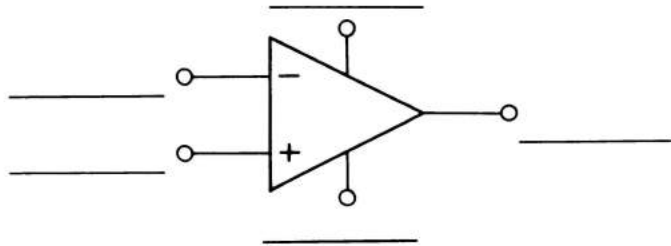
Figure 4-44
Fill in the blanks.

19. List the characteristics of an ideal op amp. _____
    _____
    _____
    _____

20. The circuit shown in Figure 4-45A is a/an _____ amplifier and it has a voltage gain of _____.

21. The circuit shown in Figure 4-45B is a/an _____ amplifier and it has a voltage gain of _____.

22. The circuit shown in Figure 4-45C is a/an _____ _____ amplifier and it has a voltage gain of _____.
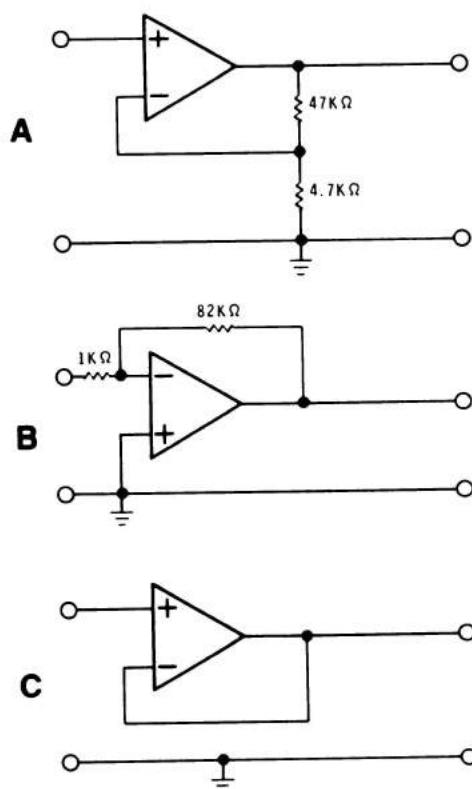
Figure 4-45
Identify these circuits and determine their gains.

Self-Review Answers

18. See Figure 4-46.
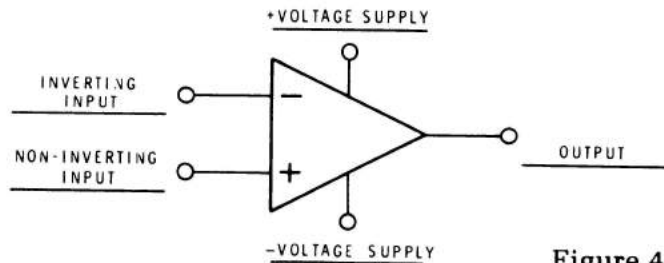
Figure 4-46

19. The ideal op amp has:

   ● Infinite voltage gain.

   ● Infinite input impedance.

   ● Zero output impedance.

   ● Zero OFFSET.

20. The circuit shown in Figure 4-45A is a **noninverting** amplifier and it has a voltage gain of **11**.

$$A_r = 1 + \frac{R_1}{R_2}$$

$$= 1 + \frac{47\ k\Omega}{4.7\ k\Omega}$$

$$= 1 + 10 = 11$$

21. The circuit shown in Figure 4-45B is an **inverting** amplifier and it has a voltage gain of **82**.

$$A_r = \frac{R_1}{R_2}$$

$$= \frac{82\ k\Omega}{1\ k\Omega} = 82$$

22. The circuit shown in Figure 4-45C is a voltage follower and it has a voltage gain of **1**.

RADIO FREQUENCY AMPLIFIERS

In this section we will continue our discussion of amplifiers. More specifically, we will be discussing radio frequency or RF amplifiers. Just as with audio amplifiers. RF amplifiers are divided into two categories: voltage amplifiers and power amplifiers. RF voltage amplifiers are used primarily in receivers, although they are also used in low level stages of transmitters. RF power amplifiers are used in transmitters.

Audio amplifiers are usually designed to amplify the whole AF spectrum, 20 to 20,000 Hz. However, radio frequency amplifiers are designed to amplify a relatively narrow portion of the RF spectrum. (The complete RF spectrum extends from 3 kHz to 300 GHz.) The portion of the spectrum that an amplifier is able to amplify is called its "bandwidth." A CB radiotelephone transmitter, for example, has a 3 kHz bandwidth, while a TV transmitter has a bandwidth of 5 to 6 MHz. Therefore, the transmitter and receiver RF amplifiers need amplify only this relatively small portion of the RF spectrum. In fact, they must reject all other frequencies.

The Tuned Amplifier

Most RF amplifiers are tuned amplifiers. They contain at least one tuned, or resonant, LC circuit. Due to the characteristics of the resonant LC circuit, the tuned amplifier amplifies only a small band of frequencies centered on the resonant frequency. The bandwidth is determined by the Q of the resonant circuit.

A tuned RF amplifier is shown in Figure 4-47. Note that it has tuned circuits on both the input and output, and that they are part of the coupling transformers. The input signal is transformer coupled to the base of the transistor Q_1. The transformer secondary (L_1) and capacitor C_1 form a parallel resonant circuit. Therefore, at the resonant frequency, maximum input voltage is developed across the base of Q_1, while little or no signal is developed at other frequencies. The output signal is developed across the tuned circuit that consists of L_2 and C_2. Again, maximum signal output is developed at the resonant frequency.
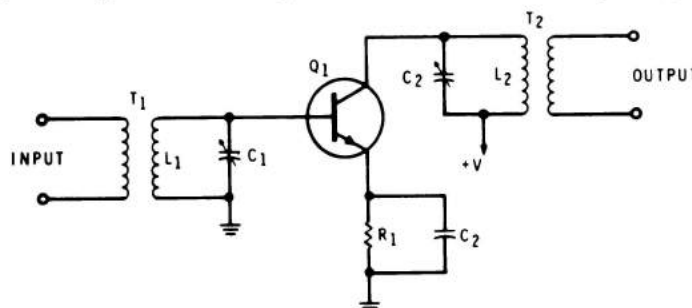
Figure 4-47
A tuned amplifier.

Coupling Circuits

Figure 4-48 shows two types of **transformer coupling**. These are also known as **inductive coupling** since they operate on the principle of mutual inductance. Figure 4-48A shows a double-tuned transformer, both primary and secondary form parallel resonant circuits. Figure 4-48B shows a single tuned transformer, only one side of the transformer is tuned.
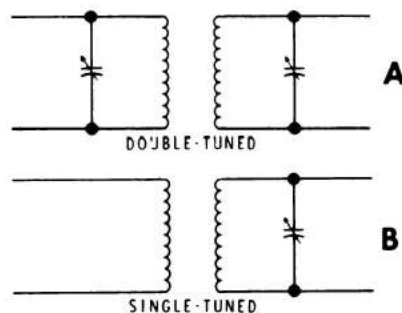
Figure 4-48
Transformer or inductive coupling.

Figure 4-49 shows a variation of transformer coupling known as **link coupling**. This type of coupling is used to connect circuits that are physically separated. As an example, link coupling could be used to couple the output of a transmitter to a separate high power amplifier.
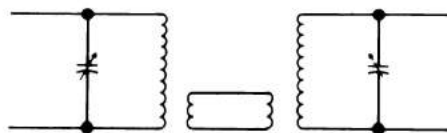
Figure 4-49
Link coupling.

Capacitive coupling is shown in Figure 4-50. The coupling capacitor in Figure 4-50A transfers RF between the two resonant circuits and their respective amplifiers. When the coupling capacitor is varied, it changes the amount of RF that is coupled from one circuit to the other.

Figure 4-50B shows a variation that is used in transistor circuits. The input impedance of the transistor is very low because it is normally forward biased. However, the impedance of the resonant LC circuit is very high. To avoid the loss of power that would come from this impedance mismatch, the coupling capacitor is connected at a lower impedance point on the inductor. This inductor "tap" is adjusted for that point where maximum power transfer occurs.
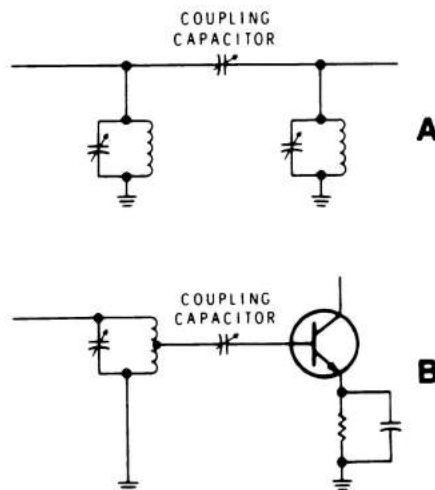
Figure 4-50
Capacitive coupling.

RF Power Amplifiers

Figure 4-51 shows a transistor RF power amplifier biased for class C operation. In this circuit, there is no base bias. However, since it takes between +0.5 and +1 volt to forward bias the emitter-base junction, the transistor is considered to be biased beyond cutoff. Therefore, it operates class C.
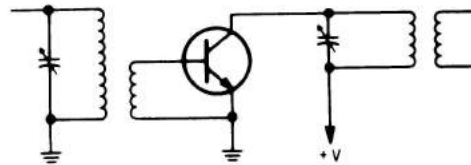
Figure 4-51
RF power amplifier.

FLYWHEEL EFFECT

Although it produces a high amount of distortion, the class C amplifier operates at a high efficiency. Therefore, RF amplifiers use an LC parallel resonant circuit, or "tank" circuit, to compensate for this distortion. The tank circuit has a "flywheel" effect which forms the complete sine wave from the pulse of current that comes from the class C amplifier.

Let's examine the flywheel effect in detail. Figure 4-52A shows a tank circuit with a battery as a current source, and a switch to control the battery current. When the switch is closed, as in Figure 4-52B, the capacitor charges with the polarity shown. Current does not flow through the inductor due to its opposition to current flow during the instant the switch is closed.

When the switch is opened, as shown in Figure 4-52C, the capacitor discharges through the inductor. This discharge current is limited by the inductor's counter emf, which is induced by the expanding magnetic field.

Once the capacitor has completely discharged, as shown in Figure 4-52D, the inductor's magnetic field starts to collapse, inducing an emf in the inductor as shown. Figure 4-52E shows that this emf and the resulting current flow recharges the capacitor in the opposite polarity. The capacitor then discharges in the opposite direction through the inductor, starting the process again. This "flywheel" action would continue indefinitely if there were no losses in the circuit.
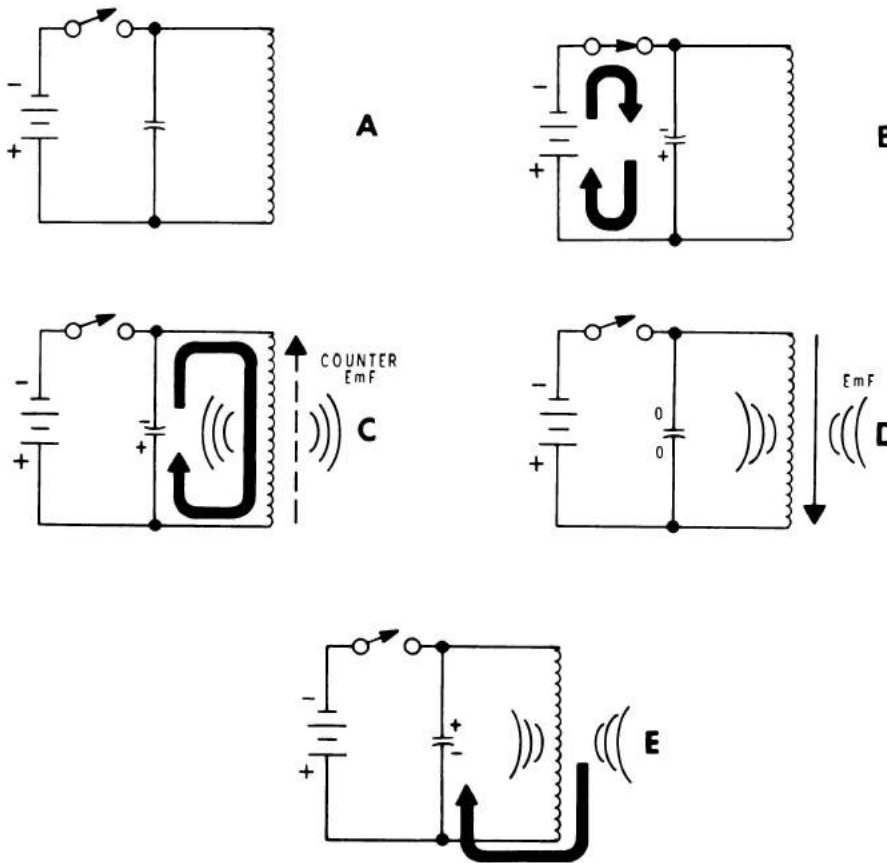
Figure 4-52
The flywheel effect.

However, due to resistance in the conductors and losses in the capacitor and inductor, the AC flywheel-effect current gradually dies out. The waveform that appears across the tank circuit is known as a damped sine wave, and is shown in Figure 4-53. The frequency of the damped sine wave is the resonant frequency of the LC tank circuit.
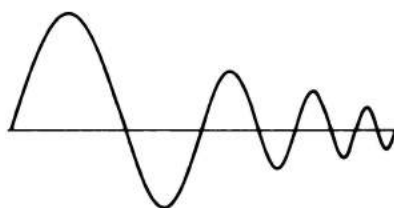
Figure 4-53
Damped sine wave for LC tank circuit.

All we need to do to continue this flywheel effect is to apply a current pulse to the tank each cycle. The class C amplifier does exactly that. Figure 4-54 shows the output current pulses and the resultant output waveform taken across the tank. The final result is a continuous sine wave output, while the amplifier utilizes the higher efficiency of class C operation.
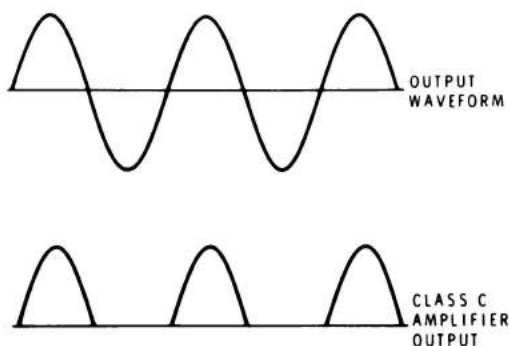
Figure 4-54
The flywheel effect completes the sine wave output signal
for the class C RF amplifier.

Tuning an RF Amplifier

The flywheel effect of the output tank circuit does indeed complete the sine wave; but the tank circuit must also be tuned to exactly the right frequency for maximum output. A typical RF amplifier is shown in Figure 4-55. C_1 and L_1 form the resonant tank circuit, and M_1 is a meter used to indicate the collector current of Q_1.

When a signal is applied to the input, Q_1 amplifies it and the output voltage is developed across the impedance of C_1 and L_1 in parallel. When C_1 and L_1 are in resonance, their impedance is maximum. Therefore maximum voltage is developed across the tank and coupled to the output.

To tune this amplifier, you would adjust C_1, so that the tank circuit resonates at the input signal's frequency. You can do this by monitoring the collector current of Q_1, or the output voltage, while adjusting C_1. When the tank circuit is adjusted to resonance at the input frequency, the output voltage will be maximum and the collector current will be minimum (due to the high impedance of the tank circuit at resonance).
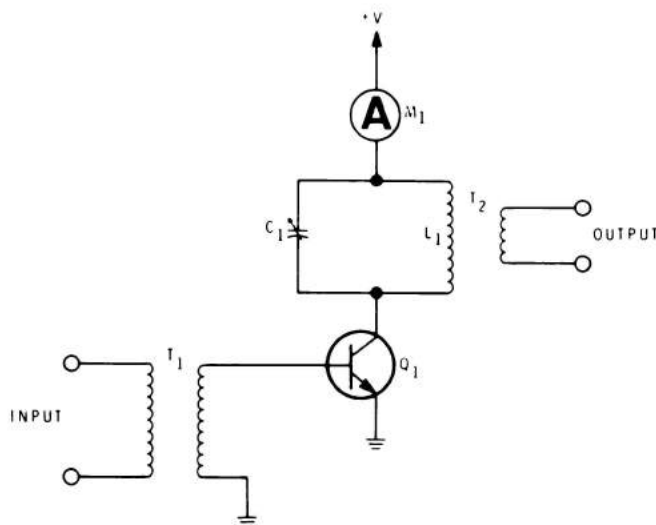
Figure 4-55
RF amplifier.

Self-Review Questions

23. Identify the RF amplifier coupling circuits shown in Figure 4-56.

   A. _____

   B. _____

   C. _____

24. An RF amplifier can be operated class C due to the _____
   _____of the LC tank circuit.

25. When you tune an RF amplifier, resonance is indicated by
   maximum _____ and minimum
   _____.
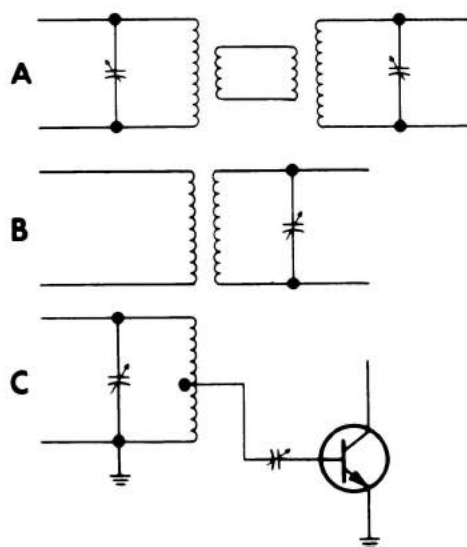
Figure 4-56
Identify these coupling circuits.

Self-Review Answers

23. The RF amplifier coupling circuits shown in Figure 4-56 are:

    A. Link coupling.
    B. Transformer or inductive coupling.
    C. Capacitive coupling.

24. An RF amplifier can be operated class C due to the **flywheel effect** of the LC tank circuit.

25. When you tune an RF amplifier, resonance is indicated by maximum **output voltage** and minimum **collector or plate current.**

OSCILLATORS

An oscillator is a circuit that generates a repetitive AC signal. The frequency of this AC signal may be as low as a few hertz or as high as several gigahertz.

There are many applications for oscillators. They are an extremely important part of every transmitter and receiver. In fact, each modern day transmitter or receiver has at least two oscillators and sometimes more. A TV receiver has at least four oscillators. Oscillators are also required for RF signal generators, audio signal generators, and many other test instruments.

There are almost as many different types of oscillators as there are applications for them. In this section, we will be discussing several of the most commonly used oscillators. However, all oscillators operate on the same basic principles. If you thoroughly understand these basic principles, you should be able to analyze the operation of almost all oscillators.

Oscillator Requirements

Basically, an oscillator is an amplifier that gets its input signal from its own output. That is, a portion of the output signal is fed back to the input to maintain oscillation. This feedback signal must be **regenerative**, or in phase, with the input signal. Compare this to the op amp whose feedback voltage is 180° out of phase. This is called **degenerative** feedback, since it reduces the amplifier gain. On the other hand, regenerative feedback increases amplifier gain and it sustains oscillation.

Figure 4-57 shows a basic block diagram of a common emitter amplifier with regenerative feedback. Since the amplifier has a 180° phase shift the feedback network must also introduce a 180° phase shift so that the feedback signal will be in phase with the input.
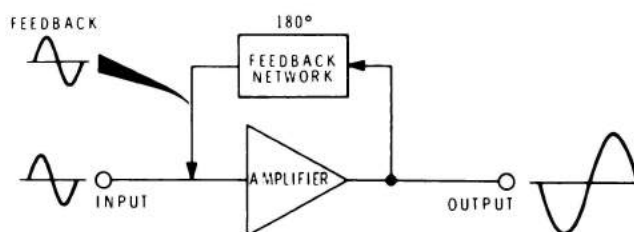
Figure 4-57
Common emitter amplifier with regenerative feedback.

Once the amplifier begins to operate, you can remove the input signal and the circuit will continue to oscillate. The gain of the amplifier replaces energy lost in the circuit, and regenerative feedback sustains oscillation.

However, this circuit could oscillate on any number of frequencies. Even minor noise pulses could change the frequency of this kind of oscillator. Therefore, it is necessary to have a means of setting the oscillator frequency.

To do this, the frequency selectivity of the parallel LC network is useful because it resonates at a specific frequency, determined by the values of inductance and capacitance. Also, the required 180° phase shift is produced across this network. Therefore, an LC tank circuit in the feedback loop, as shown in Figure 4-58, will control the frequency. The tank resonates at its natural frequency and amplifier gain replaces energy lost in the tank.
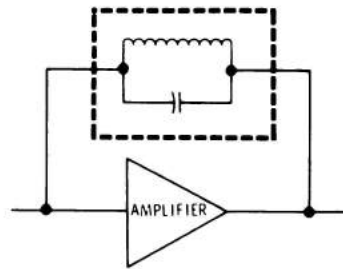
Figure 4-58
An LC network controls the frequency.

Up to this point, the oscillators were amplifiers with input signals applied to start the circuit. In actual practice, oscillators must start on their own. This is a natural phenomenon. When a circuit is first turned on, energy levels do not instantly reach maximum, but gradually approach it. This produces noise pulses that are phase shifted and fed back to the input, as shown in Figure 4-59. The amplifier steps up these pulses, which are again supplied to the input. This action continues and oscillation is underway. Therefore, the oscillator is naturally self-starting.

Let's summarize the requirements for an oscillator:

1. An amplifier is necessary to replace circuit losses.

2. Frequency determining components are needed to set the oscillation frequency.

3. Regenerative feedback is required to sustain oscillations.

4. The oscillator must be self-starting.
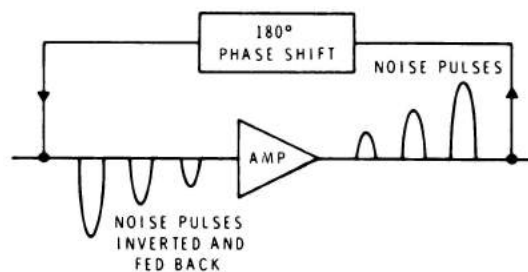
Figure 4-59
Oscillators are self-starting.

The Armstrong Oscillator

The "Armstrong," or "tickler coil" feedback, oscillator is shown in Figure 4-60. This circuit uses mutual inductance or transformer action between L_1 and tickler coil L_2 for regenerative feedback. L_1 and C_1 determine the frequency of operation, and transistor Q1 is the amplifier.

When voltage is applied to the circuit, current flows through Q1 and the tickler coil L_2. This current produces an expanding magnetic field around L_2. This, in turn, induces voltage into L_1, which starts "flywheel effect" oscillations in the tank circuit of L_1-C_1. These oscillations are applied to the base of Q_1 and amplified. The amplified signal is developed across L_2 and, through transformer action, fed back to the resonant tank circuit L_1-C_2; thus sustaining the oscillations.

To maintain the oscillation, the amplifier must supply enough gain to overcome the circuit losses. You can vary the amount of feedback voltage by adjusting the position of the tickler coil with respect to the tank coil. Note that since the transistor introduces a 180° phase shift, the transformer must also supply a 180° phase shift for regenerative feedback to occur.

Notice that the base connection for Q_1 is "tapped" down on L_1. This is done to obtain a low impedance to match the transitor's low base-emitter impedance and therefore obtain maximum power transfer.
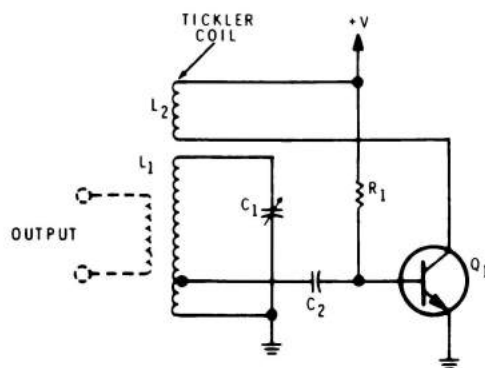
Figure 4-60
Transistor Armstrong oscillator.

Hartley Oscillator

The Hartley oscillator is actually an adaptation of the basic Armstrong oscillator shown in Figure 4-61A. Here, L_2 is the tickler coil. C_2 couples the feedback signal to the tickler coil and prevents the DC voltage from being shorted to ground.

One of the undesirable features of the Armstrong oscillator is that the tickler coil has a tendency to resonate with the distributed capacitance in the circuit. This results in oscillator frequency variations. Figure 4-61B shows you that, by making the tickler coil part of the resonant tank, we can eliminate this undesirable effect. Now, both L_1 and L_2 resonate with C_1. This circuit is known as a Hartley oscillator and you can easily recognize it by the tapped coil in the tank circuit.

The Hartley oscillator operates in a manner that is similar to the Armstrong oscillator. Regenerative feedback is still accomplished by "tickler coil" L_2, which is now part of the resonant tank.
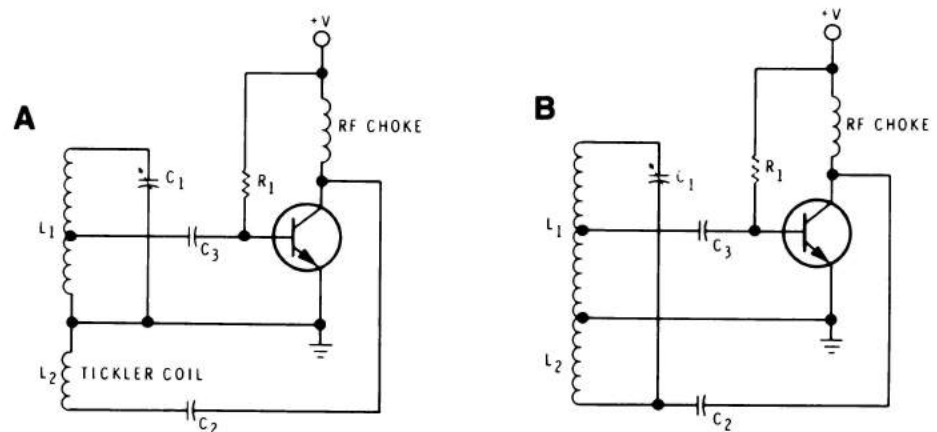
Figure 4-61
How the Hartley oscillator in (B) is derived from the
Armstrong oscillator in (A.)

Figure 4-62 shows two types of transistor Hartley oscillators. You can see that both are Hartley oscillators, by the tapped coil, but the DC current path differs between the two. Figure 4-62A is called a "series-fed Hartley oscillator" because the path for DC current is from ground through L_2, which is part of the tank, and through Q_1 to the power supply.

Figure 4-62B is called a "shunt-fed Hartley oscillator." Its DC current path is from ground through Q_1 to the power supply. Therefore, a series fed oscillator's DC current path is through part of the tank and the transistor, while a shunt-fed oscillator's DC current flows only through the transistor. The disadvantage of the series-fed oscillator is that the DC current heats L_2 causing it to expand and change value. This, in turn, causes the oscillator frequency to change. Therefore, the shunt-fed Hartley has better frequency stability.
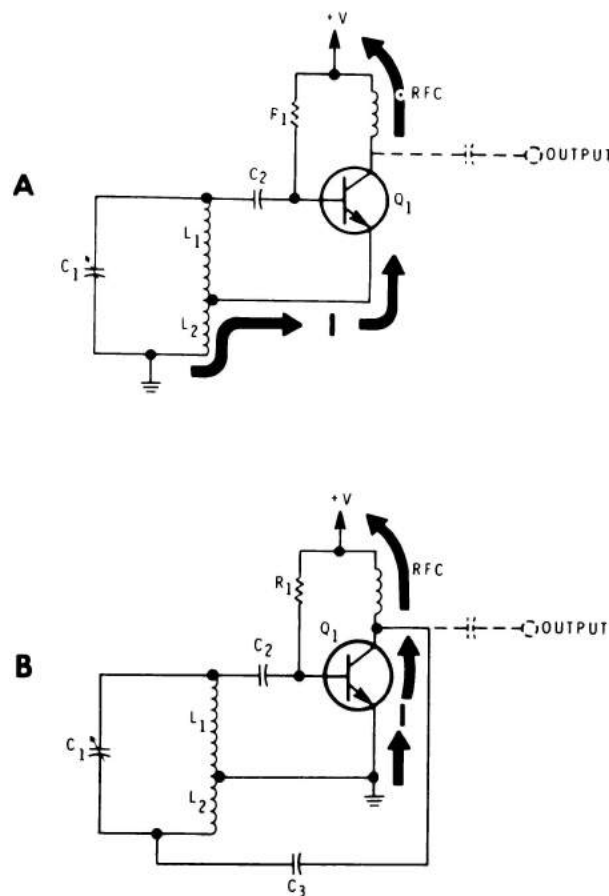
Figure 4-62
(A) Series-fed Hartley oscillator,
(B) shunt-fed Hartley oscillator.

Colpitts Oscillator

You can easily identify the Colpitts oscillator, shown in Figure 4-63, by the two capacitors, C_1 and C_2, in the resonant tank. The operation is very similar to the Hartley, with the exception that the capacitive branch of the tank is tapped rather than the inductive branch. Regenerate feedback voltage is developed across C_2, which sustains the flywheel oscillations of the tank circuit. Therefore, the Colpitts oscillator utilizes capacitive feedback.

The resonant frequency of the Colpitts oscillator, or any LC oscillator, is determined by the usual formula for resonance:

$$f_r = \frac{1}{2\pi \sqrt{LC}}$$

However, C represents the two tuning capacitors, C_1 and C_2, connected in series. For example, if C_1 and C_2 are each 400 pF, then the combined capacitance equals 200 pF. Therefore, the Colpitts oscillator requires larger tuning capacitors for a given resonant frequency than other oscillator circuits. This is actually an advantage. Since the transistor's internal junction capacitance is in parallel with the tank, any changes in the transistor, due to heat or bias variations, causes this capacitance to vary. This, in turn, causes the oscillator frequency to drift. However, because the tuning capacitors are much larger in the Colpitts oscillator, the effect of the junction capacitance is greatly reduced. This is a decided advantage for the Colpitts oscillator.
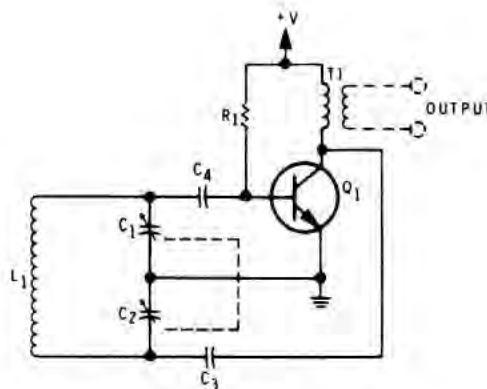
Figure 4-63
Transistor Colpitts oscillator.

Crystal Oscillators

Certain crystalline substances, such as quartz, have an unusual electrical characteristic. If mechanical pressure is applied to the crystal, a DC voltage is generated. Likewise, if a vibrating mechanical pressure is applied, an AC voltage is generated. Conversely, if an AC voltage is applied across the crystal, the crystal undergoes a physical change, resulting in mechanical vibration. This relationship between electrical and mechanical effects is known as the **piezoelectric effect**.

In its natural shape, quartz is a hexagonal prism with pyramids at the ends. Slabs are cut from the "raw" quartz to obtain a usable crystal. There are many ways to cut a crystal, all with different names; such as X cut, NT cut, CT cut, and AT cut. Each cut has a different piezoelectric property. Figure 4-64 shows how the most popular cut, the AT, is obtained from the quartz crystal.
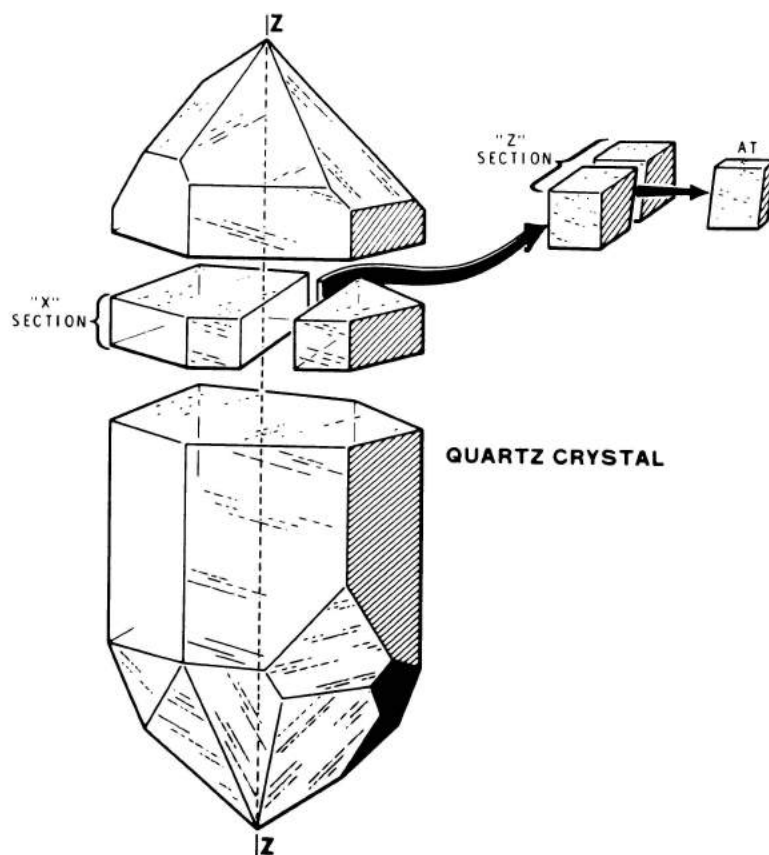
Figure 4-64
How an "AT" cut is obtained from a quartz crystal.

Because of their structure, crystals have a natural frequency of vibration. If the frequency of the applied AC signal matches this natural frequency, the crystal will vibrate a large amount. However, if the frequency of the exciting voltage is slightly different than the crystal's natural frequency, little vibration is produced. The crystal, therefore, is extremely frequency selective, making it desirable for filter circuits. Also, the crystal's mechanical frequency of vibration is extremely constant, which makes it ideal for oscillator circuits.

The natural frequency of a crystal is usually determined by its thickness. As shown in Figure 4-65, the thinner the crystal is, the higher is its natural frequency. Conversely, the thicker a crystal is, the lower is its natural frequency. To obtain a specific frequency, the crystal slab is ground to the required dimensions. Of course there are practical limits on just how thin a crystal can be cut, without it becoming extremely fragile.

Crystals are used in a number of oscillator circuits. Figure 4-66 shows a crystal Colpitts oscillator, which is sometimes called a Pierce oscillator. C_3 provides the path for regenerative feedback, while C_1 and C_2 identify it as a Colpitts oscillator. The crystal replaces the tank circuit's inductor. It also determines the operating frequency and holds it constant.

The major advantage of using crystals is their frequency stability. However, unlike the LC oscillators we have previously discussed, the frequency of oscillation cannot be varied. To change the frequency of a crystal oscillator, you must use a different crystal that is cut to the desired frequency.

In summary: The advantages of crystal oscillators are their extreme precision and frequency stability. The disadvantage is that their frequency cannot be varied.
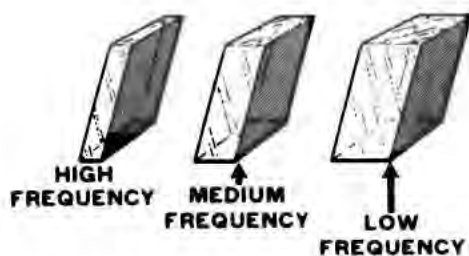
Figure 4-65
Thickness determines the crystal's natural frequency.
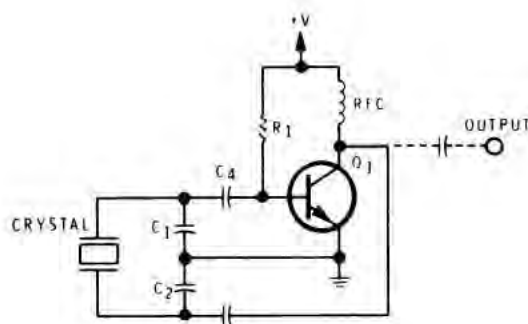
Figure 4-66
Crystal Colpitts oscillator.

Self-Review Questions

26.   What are the requirements for an oscillator? _____

_____

27.   Identify the schematic diagrams shown in Figure 4-67.

    A. _____

    B. _____

    C. _____

    D. _____
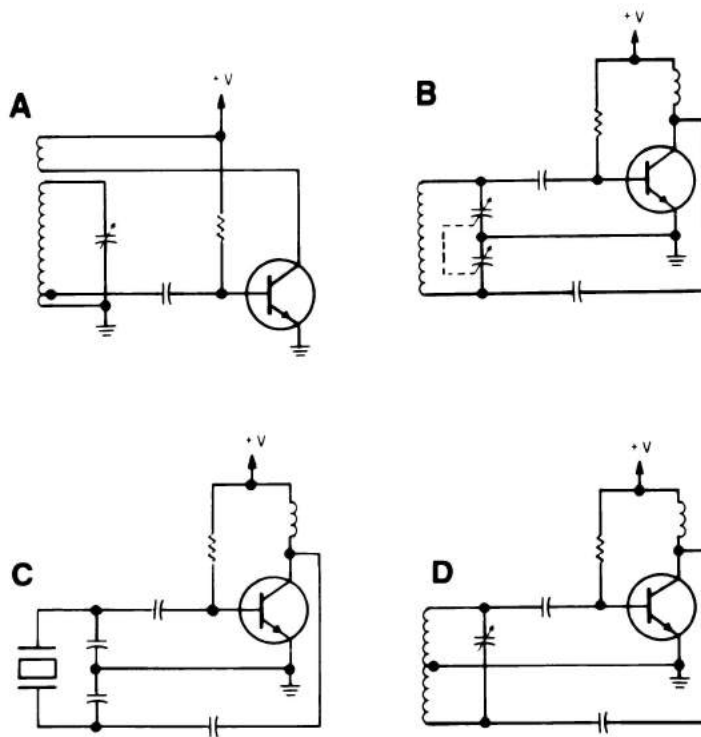
Figure 4-67
Identify these oscillators.

28.   What is the "piezoelectric effect"? _____

_____

29.   What determines the natural frequency of a crystal? _____

_____

30.   What are the advantages of using a crystal oscillator when com-
pared to an LC oscillator? What are the disadvantages? _____

_____

Self-Review Answers

26. The requirements for an oscillator are as follows:

    1. It must have amplifier to replace circuit losses.
    2. It must have frequency determining components to set the oscillation frequency.
    3. It must have regenerative feedback to sustain oscillations.
    4. It must be self-starting.

27. The schematic diagrams shown in Figure 4-67 are:

    A. Armstrong oscillator.
    B. Colpitts oscillator.
    C. Crystal Colpitts oscillator.
    D. Hartley oscillator.

28. The piezoelectric effect is exhibited by quartz crystals. When pressure is applied to the crystal, a voltage is generated. Conversely, when a voltage is applied, the crystal undergoes a physical change, resulting in mechanical vibration.

29. The thickness of a crystal determines its natural frequency. The thinner the crystal is, the higher is its natural frequency.

30. The advantages of crystal oscillators are their extreme precision and frequency stability. The disadvantage is that their frequency cannot be varied.

APPENDIX A

Logarithms

Since pocket calculators and home computers have become an everyday part of our lives, the use of logarithms to simplify problem solving has diminished. However, to properly understand decibels, you need some knowledge of logarithms.

What is a logarithm? In the very simplest terms, it is an exponent. The **common** logarithm is an exponent of the base 10. It is not necessary that 10 be used for a base. Any number could be used; for instance, the system of **natural** logarithms uses 2.71828 for a base. Two could be used, or eight, or any other number. In this discussion, we will limit our study to the **common** logarithms which use the base 10. In your study of electronics, you have already done considerable work with "powers of 10."

When we express a number in powers of 10, its exponent is the logarithm of the number. For example:

$1 = 10^0$, or 0 is the logarithm of 1.
$10 = 10^1$, or 1 is the logarithm of 10.
$100 = 10^2$, or 2 is the logarithm of 100.
$1000 = 10^3$, or 3 is the logarithm of 1000.

Characteristic and Mantissa

Every number has a logarithm. Numbers such as 100 or 1,000 have whole number logarithms since they are exact multiples of the base 10. However, the number 32, for instance, is between 10 and 100. Its logarithm must be between 1 (the log of 10) and 2 (the log of 100). Actually, the log of 32 is 1.5.

With logarithms the whole number is called the **characteristic**, and the fractional part is the **mantissa**. In our example of 1.5 the logarithm consists of a characteristic of 1, and a mantissa of 0.5.

Finding a Log of a Number

The first step in finding the log of a number is to determine the characteristic. How the number fits in the scale of 1, 10, 100, 1000, and so on determines its characteristic. For example, the log of 1 is zero, thus any number less than 1 will have a negative characteristic.

Since the log of 1 is zero and the log of 10 is 1, the log of any number between 1 and 10 will have a characteristic of zero. For instance, 5 has a log of 0.6990. The log of 100 is 2; therefore, the log of any number between 10 and 100 has a characteristic of 1. For instance, 50 has a log of 1.6990. The log of a number from 100 up to 1000 will have a characteristic of 2. For instance, the log of 500 is 2.6990.

In logarithms of 5, 50, and 500, only the characteristic changed. The mantissa, the part of the logarithm following the decimal point, is the same for each number. The mantissa is the same for any given sequence of digits regardless of decimal point placement.

How do we find the mantissa? The easiest way to find both characteristic and mantissa is with a scientific calculator. However, if you do not have such a calculator, you can look up the mantissa in Table 1. Here is the way to find a logarithm of a number using the table.

Find the logarithm of 354.

Since 354 is between 100 and 1000, you know that the characteristic is 2. Turn to Table 1, and go down the extreme left hand column of numbers until you locate 35. Next, move across the page until you intersect the column marked 4. Column numbers are at both the top and bottom of the page. The mantissa of 354 is 5490. Combining this with the characteristic gives 2.5490 which is the logarithm of 354.

Try another. What is the logarithm of 6?

When you look in the column of numbers you see that there is no number less than 10. However, this doesn't matter. Remember that the mantissa of 5 was the same as the mantissa of 50 or 500. Thus the mantissa of 6 is the same as the mantissa of 600. So, go down the numbers column to 60 then across the 0 column and read the mantissa of 7782. Since 6 is between 0 and 10 the characteristic is 0; therefore, the logarithm of 6 is 0.7782.

You have seen that any number may be expressed as a power of 10 and that the exponent involved is called the common logarithm of that number. The logarithm has two parts. The characteristic precedes the decimal point and is used to indicate the location of the decimal point in the original number. The characteristic is determined by inspection. The second part of the logarithm, the mantissa, follows the decimal point and indicates the numerical value of the original number. The mantissa is found from a table.

TABLE 1

	LOGARITHMS OF NUMBERS									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1895	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Courtesy of U.S. Coast and Geodetic Survey

Table 1 (continued)

	LOGARITHMS OF NUMBERS.									
	0	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8209	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	843	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9454	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996
	0	1	2	3	4	5	6	7	8	9

Unit 5

DIGITAL ELECTRONICS

CONTENTS

INTRODUCTION

The purpose of this unit on digital electronics is to give you an overview of the subject and to introduce you to its basic concepts. You will learn the basic digital logic functions; elementary Boolean Algebra; and apply these techniques in several experiments. However; this unit is merely an extended introduction to digital electronics. To fully understand digital techniques; we highly recommend Heath Individual Learning Program EE-3201, Digital Techniques.

Examine the Unit Objectives listed in the next section to see what you will learn in this unit.

UNIT OBJECTIVES

When you have completed this unit you should be able to:

1. State why the binary number system is used in digital electronics.

2. Convert between the decimal, binary, and binary-coded-decimal (BCD) number systems.

3. Identify each of the six basic logic gate symbols, their truth tables, and their logic expressions.

4. Write the truth table and logic expressions for each of the six basic logic elements.

5. Define the term "flip-flop" and name three basic types.

6. Identify the logic diagrams and symbols of the three basic types of flip-flops.

7. Answer questions about and discuss the basic operation of the three types of flip-flops.

8. Name the two most widely used types of sequential logic circuits.

9. Identify binary up, binary down, and BCD counters from their logic diagrams.

10. Answer questions about and discuss the basic operation of counters and shift registers.

11. Define the terms "monostable" and "astable."

12. Determine the output pulse duration of a 555 monostable multivibrator.

13. Determine the output pulse duration of a 555 astable multivibrator.

NUMBER SYSTEMS

Many different number systems are being used throughout the world each day. But the first system that you always think of is the decimal system, the one we all learned in school and still use daily. It has ten digits — 0 through 9.

You also use the "duodecimal" system each day, which uses the symbols (or digits) 1 through 12. Since the day is divided into two 12-hour periods, and the year is divided into 12-month periods, you use this system to record the passage of time.

There are also many other number systems in use. In this section, you will learn about the ones that are used in digital electronics and computers. But first, we'll review the basic principles of the decimal system.

The Decimal System

Most of us are very familiar with the decimal number system, where the digits 0 through 9 are combined in a certain way to indicate a specific quantity. For example, the number 1984 is made up of 4 digits. The right-hand digit (the least significant digit, or LSD) stands for a number times ten raised to the power of zero, or 10^0. You will recall that 10^0, or any number raised to the power of zero, is equal to 1 (See "Appendix A," "Scientific Notation," in Unit 1). The next digit represents a number times 10^1, and this continues until the left-hand, (the most significant digit or MSD) is reached. Therefore, you can express this number as:

$$1984 = (1 \times 10^3) + (9 \times 10^2) + (8 \times 10^1) + (4 \times 10^0) =$$
$$1000 + 900 + 80 + 4$$

Thus, the decimal number system is known as a positional, or weighted, system. Each digit position in a number carries a particular weight in determining the magnitude of that number.

If you tried to design an electronic circuit that would respond to these decimal numbers, you would need a different voltage level to represent each one of the ten digits. The circuit would have to be capable of both generating and detecting each of these ten levels. This would be difficult to do and would require very complex circuitry.

However, consider a system where you could represent all of the numbers you need (only two in this case) by only the absence or the presence of a single voltage; usually, a simple "on or off" state. This can be easily accomplished with a much smaller number of electronic components: a simple switch is either on or off; a transistor is either conducting or cut off; an LED is either on or off. This number system, which has only two characters or digits, is called the "binary number system."

The Binary Number System

The two digits that are used in the binary system are 0 and 1. And since there are only two digits, the binary system is based on powers of two. For example, the binary number 1101 can be broken down in the following manner:

$$1101 = (1 \times 2^3) + (1 \times 2^2) + (0 \times 2^1) + (1 \times 2^0)$$

This number can then be written in decimal notation as:

$$8 + 4 + 0 + 1 = 13$$

In a binary number, each digit is called a **bit**. This is actually a contraction of the words binary digit.

As you can see, it takes a lot more binary digits or "bits" to represent a given quantity than it does to represent the same quantity using the decimal system. However, it is much more convenient to use the binary system in electronic circuits, since there are a large number of components that have two distinct levels of operation.

Number System Conversions

Since electronic circuits use the binary number system exclusively, you must be capable of converting binary numbers to decimal and vice versa.

To convert a binary number into its decimal equivalent, you add together the weights of the positions in the number where binary 1's occur. The weights of the number positions are shown below.

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
128	64	32	16	8	4	2	1

As an example, convert the binary number 1010 into its decimal equivalent. You can add the weights of the positions where 1's occur as shown below.

Binary Number	1		0		1		0		
Position Weights	(8)		(4)		(2)		(1)		
Decimal Equivalent	8	+	0	+	2	+	0	=	10

Next, convert the binary number 101101 to its decimal equivalent.

Binary Number	1	0	1	1	0	1	
Position Weights	(32)	(16)	(8)	(4)	(2)	(1)	
Decimal Equivalent	32 + 0 + 8 + 4 + 0 + 1					= 45	

After you solve a few practice problems you will quickly catch on to this procedure.

To convert a decimal number into binary, you must repeatedly divide the number by 2 and note the remainder. The remainder, which will always be either 1 or 0, forms the equivalent binary number.

As an example, convert the number 175 into its binary equivalent.

	Remainder	
$175 \div 2 = 87$	1 ◄───── Least significant Bit
$87 \div 2 = 43$	1
$43 \div 2 = 21$	1
$21 \div 2 = 10$	1
$10 \div 2 = 5$	0
$5 \div 2 = 2$	1
$2 \div 2 = 1$	0
$1 \div 2 = 0$	1 ◄───── Most significant Bit

Therefore, the decimal number 175 is equal to 10101111 in binary. You can check it by converting the binary version back into decimal form using the procedure we discussed earlier.

Binary Coded Decimal

Even though we know what the binary number system is, it is still difficult to deal with. This is simply because we are unfamiliar with it. And, let's face it; its cumbersome when you have to deal with large numbers. However, the binary system is the natural language of the two-state digital electronic components. Therefore, to simplify the reading and handling of large numbers, special binary codes have been developed.

These special codes convert each decimal digit to a four-bit binary number. This can be easily accomplished because it takes four binary digits to represent the decimal numbers 0 through 9. Therefore, a two-digit decimal number will require 8 bits, a three-digit decimal number — 12 bits, and so on. This code is known as the **binary coded decimal** (BCD) system.

The most commonly used BCD system uses the first ten binary numbers (0-9) and rejects the remaining six (10-15) that the four bits are capable of representing. Therefore, it has ten valid codes and six invalid codes. This code is called the natural BCD code or the 8421 code.

To represent a decimal number in BCD notation, you substitute the appropriate four-bit code for each decimal digit. For example, the number 937 in BCD would be:

Decimal Number —	9	3	7
Binary Code —	1001	0011	0111

The BCD number would be 1001 0011 0111. A space is left between each four-bit group so there will be no danger of confusion between the BCD code and the "pure" binary code.

The beauty of the BCD code is that the ten BCD code combinations are very easily remembered. Once you begin to work with binary numbers on a regular basis, you will find that the BCD numbers will come to you as quickly and automatically as decimal numbers. For that reason by simply glancing at the BCD representation of a decimal number you can make the conversion almost as quickly as if it were already in decimal form.

Self-Review Questions

1. What is the base of the binary number system? _____.

2. Why is the binary number system used in digital electronics? __
   _____
   _____

3. Convert the following binary numbers to decimal.

    A. 11001
    B. 1110010
    C. 1110101
    D. 1101101

4. Convert the following decimal numbers to binary.

    A. 89
    B. 52
    C. 101
    D. 156

5. Convert the following BCD numbers to decimal.

    A. 1001 0011
    B. 1000 0101
    C. 0110 0101 0010
    D. 0100 0000 1001

6. Convert the following decimal numbers to BCD.

    A. 42
    B. 69
    C. 705
    D. 923

Self-Review Answers

1. The base of the binary number system is 2.

2. The binary number system is used in digital electronics because a large number of electronic devices have two distinct operating states, such as "on and off", etc.

3. A —

Binary Number	1	1	0	0	1
Position Weights	(16)	(8)	(4)	(2)	(1)
Decimal Number	16 +	8 +	0 +	0 +	1 = 25

   B — 114
   C — 117
   D — 109

4.   A — 89 ÷ 2 = 44   Remainder   1 (LSB)
          44 ÷ 2 = 22   0
          22 ÷ 2 = 11   0
          11 ÷ 2 = 5   1
           5 ÷ 2 = 2   1
           2 ÷ 2 = 1   0
           1 ÷ 2 = 0   1 (MSB)

   Binary Number = 1011001

   B — 110100
   C — 1100101
   D — 10011100

5.   A — BCD Number   1001   0011
      Decimal Number   9   3   = 93

   B — 85
   C — 652
   D — 409

6.   A — Decimal Number   4   2
         BCD   0100   0010
                 BCD = 0100 0010

   B — 0110 1001
   C — 0111 0000 0101
   D — 1001 0010 0011

LOGIC ELEMENTS

All digital equipment, simple or complex, is constructed from just a few basic circuits. These circuits are called logic elements. A logic element performs some specific logic function on binary data.

There are two basic types of digital logic circuits: decision making and memory. Decision making logic elements monitor binary inputs and produce outputs based on the input states and on the characteristics of the logic element itself. Memory logic elements are used to store binary data. This section discusses the basic decision making logic elements.

The Inverter

The simplest form of logic element is the inverter, or NOT circuit. The inverter is a logic element whose output state is always the opposite of its input state. If the input is a binary 0, the output is a binary 1. If the input is a binary 1, the output is a binary 0. The inverter has an output that is the **complement** of the input. The binary states 1 and 0 are considered to be complementary.

The operation of the inverter, or any logic element, is summarized by a chart known as a **truth table**. This chart shows all of the possible input states and the resulting outputs.

Figure 5-1 is the truth table for an inverter. The input is designated A, while the output is labeled $\overline{A}$, which is pronounced as "A NOT" or "A BAR". The bar over the letter A indicates the complement of A. Note that the truth table shows all possible input combinations and the corresponding output for each. Since the inverter has a single input, there are only two possible input combinations: 0 and 1. The output in each case is the complement or opposite of the input.

INPUT	OUTPUT
A	$\overline{A}$
0	1
1	0

Figure 5-1
Truth table for logic inverter.

The symbols used to represent a logic inverter are shown in Figure 5-2. The triangle portion of the symbol represents the circuit itself, while the circle designates the inversion or complementary nature of the circuit. Either of the two symbols may be used. Note the input and output labeling. Such simplified symbols are used instead of the actual electronic schematic in order to simplify the drawing and application of a logic circuit. It is the logic function and not the circuit that is important. For this reason, we will not be examining the actual circuitry of logic elements. Exactly how the logic operation is performed is not important. It is more important to know what each logic element is, what it does, and how to use it.
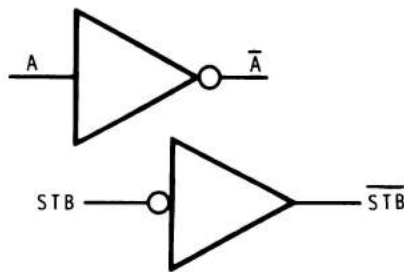
Figure 5-2
Symbols for a logic inverter.

The AND Gate

The AND gate is a logic circuit that has two or more inputs and a single output. The output of this gate is a binary 1 if, and only if, all inputs are binary 1. If any one or more inputs are a binary 0, the output will be binary 0. For example, in a two-input gate, a one input AND a one input results in a one output; hence the name AND gate.

The operation of a two-input AND gate is summarized in the truth table of Figure 5-3. The inputs are designated A and B. The output is designated C. The output for all possible input combinations is indicated in the truth table.

INPUTS		OUTPUT
A	B	C
0	0	0
0	1	0
1	0	0
1	1	1

Figure 5-3
Truth table for an AND gate.

The basic symbol used to represent an AND gate is shown in Figure 5-4. Note that the inputs and outputs are labeled to correspond to the truth table in Figure 5-3. Keep in mind that the AND gate may have any number of logical inputs.

An important point to note about the AND symbol in Figure 5-4 is the equation at the output, $C = A \cdot B$ or $C = AB$. This equation is a form of algebraic expression that is used to designate the logical function being performed. The equation expresses the output C in terms of the input variables A and B. It is read as, "C equals A AND B." Here the AND function is designated by the dot between the two input variables A and B.

Figure 5-4
Logic symbol for an AND gate.

The OR Gate

Another basic logic element is the OR gate. Like the AND gate, it can have two or more inputs and a single output. Its operation is such that the output is a binary 1 if any one or all inputs are a binary 1. The output is a binary 0 only when both inputs are binary 0.

The logical operation of a two input OR gate is expressed by the truth table in Figure 5-5. The truth table designates all four possible input combinations and the corresponding output. Note that the output is a binary 1 when either or both of the inputs are binary 1. That is, the output is 1 if input D OR input E OR both are present.

INPUTS		OUTPUT
D	E	F
0	0	0
0	1	1
1	0	1
1	1	1

Figure 5-5
Truth table for an OR gate.

The logic symbol for an OR gate is shown in Figure 5-6. The inputs are labeled according to the truth table in Figure 5-5. Notice that the output expression for the OR gate is $F = D + E$. The plus sign is used to designate the logical OR function.

Figure 5-6
Logic symbol for an OR gate.

The NAND Gate

The term NAND is a contraction of the expression NOT-AND. Therefore, a NAND gate is an AND gate followed by an inverter. Figure 5-7A shows the basic diagram of a NAND gate. Note the algebraic output expression for the AND gate and the inverter. The entire AND output expression is inverted and indicated by the bar over it.

Figure 5-7B shows the standard symbol used for a NAND gate. It is similar to the AND symbol but a circle has been added to the output to represent the inversion that takes place.

The logical operation of the NAND gate is easy to deduce from the circuit in Figure 5-7. Its operation is indicated by the truth table in Figure 5-8. The NAND output is simply the complement of the AND output.

A   **B**
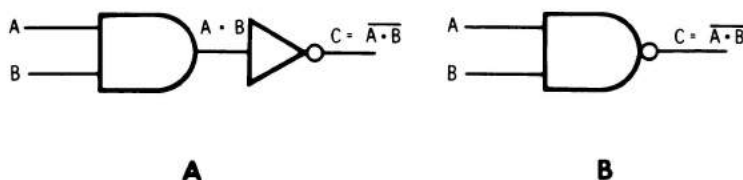
Figure 5-7
A NAND gate.

INPUTS		OUTPUT	
A	B	AND $A \cdot B$	NAND $\overline{A \cdot B} = C$
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

Figure 5-8
Truth table for a NAND gate.

The NOR Gate

Like the term NAND, NOR is a contraction for the expression NOT-OR. Therefore, the NOR gate is essentially a circuit that combines the logic functions of an OR gate and an inverter.

Figure 5-9A is a logic representation of a NOR gate. Figure 5-9B shows the standard symbol used to represent a NOR gate. Note that the output expression is the inverted OR function.

The logical operation of the NOR gate is illustrated by the truth table of Figure 5-10. The NOR output is simply the complement of the OR function. Like any other logic gate, both NAND and NOR gates may have two or more inputs, as required by the application.

A **B**
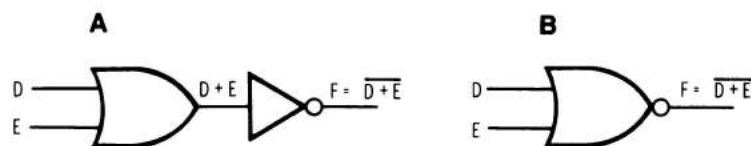
Figure 5-9
A NOR gate.

INPUTS		OUTPUT	
D	E	OR D+E	NOR $\overline{D+E}$=F
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

Figure 5-10
Truth table for a NOR gate.

The Exclusive OR Gate

The standard OR gate is generally referred to as an inclusive OR. The OR circuit produces a binary 1 output if any one or more of its inputs are binary 1. However, the exclusive OR produces a binary 1 output only if one, and only one, of the inputs is a binary 1. The truth table shown in Figure 5-11 compares the output of the standard inclusive OR and the exclusive OR.

A special symbol is used to designate the exclusive OR function in logic expressions. Just as the plus sign represents OR and the dot represents the AND function, the symbol $\oplus$ represents the exclusive OR function. The exclusive OR of inputs A and B is expressed as:

$$C = A \oplus B$$

The symbol for an exclusive OR gate is shown in Figure 5-12.

INPUTS		OUTPUTS	
A	B	STANDARD OR	EXCLUSIVE OR
0	0	0	0
0	1	1	1
1	0	1	1
1	1	1	0

Figure 5-11
Truth table for both inclusive and exclusive OR gates.

Figure 5-12
The Exclusive OR symbol.

Self-Review Questions

7. Identify the logic symbols shown in Figure 5-13.

    A. _____   D. _____
    B. _____   E. _____
    C. _____   F. _____
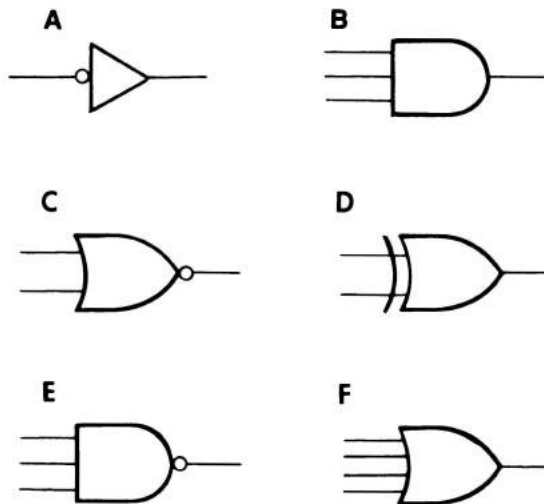
Figure 5-13
Identify these symbols.

8. If the input to a logic inverter is labeled RSA, the output will be _____.

9. Write the logic expression for the circuit of Figure 5-14.

    _____.
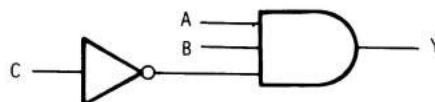
Figure 5-14
What is the logic expression for this circuit?

10. Write the truth table for a 3-input AND gate.

11. Write the truth table for a 4-input OR gate.

12. Write the truth table and draw the logic diagram for the expression $Y = \overline{A \oplus B}$.

13. Which of the following logic expressions matches the circuit shown in Figure 5-15?

$$
\begin{array}{lll}
\text{A.} & X = & \overline{E + \overline{F} + \overline{G}} \\
\text{B.} & X = & \overline{E + F + G} \\
\text{C.} & X = & \overline{E \cdot \overline{F} \cdot \overline{G}} \\
\text{D.} & X = & \overline{\overline{E \cdot F \cdot G}}
\end{array}
$$
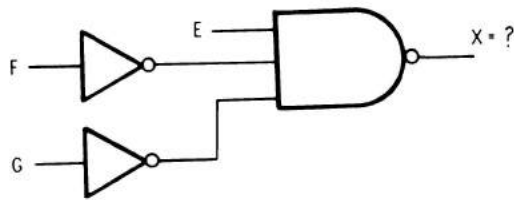
Figure 5-15
What is the logic expression for this circuit?

14. Which of the following truth table outputs indicates the operation of the circuits shown in Figure 5-15.

INPUTS			OUTPUT X = ?			
E	F	G	A	B	C	D
0	0	0	1	1	0	0
0	0	1	1	1	0	1
0	1	0	1	1	0	1
0	1	1	1	1	0	1
1	0	0	1	0	1	1
1	0	1	1	1	0	1
1	1	0	1	1	0	1
1	1	1	0	1	0	1

15. What logic function is indicated by the truth table shown below?

INPUT		OUTPUT
A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

Self-Review Answers

7. The following logic symbols are shown in Figure 5-13:

   A. Inverter or NOT gate.
   B. AND gate.
   C. NOR gate.
   D. Exclusive OR gate.
   E. NAND gate.
   F. OR gate.

8. If the input to a logic inverter is labeled RSA, the output will be $\overline{RSA}$.

9. The logic expression for the circuit of Figure 5-14 is:

$$Y = A \cdot B \cdot \overline{C}$$

10. The truth table for a 3-input AND gate is:

INPUTS			OUTPUT
A	B	C	D
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

11. The truth table for a 4-input OR gate is:

INPUTS				OUTPUT
A	B	C	D	E
0	0	0	0	0
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	1
1	0	1	0	1
1	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	1

12. The logic diagram and truth table for $Y = \overline{A \oplus B}$ is shown in Figure 5-16.
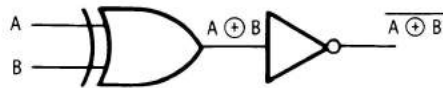
INPUTS		OUTPUT
A	B	Y
0	0	1
0	1	0
1	0	0
1	1	1

Figure 5-16

13. C — The logic expression for Figure 5-15 is:

$$X = \overline{E \cdot \overline{\overline{F}} \cdot \overline{G}}$$

14. B — Output B indicates the operation of the circuit shown in Figure 5-15.

15. The truth table indicates the NOR logic function.

FLIP-FLOPS

The output signals of the logic elements we just discussed depend only on the instantaneous value of the input signals. These elements are known as "combinatorial logic elements." Their operation does not depend on the history of previous states, because they have no memory function.

Elements whose operation depend on their history are called "sequential elements." The most important element of this class is the flip-flop. In this section, we will discuss the types of flip-flops that are used most often in sequential logic and digital memory circuits.

What is a Flip-Flop?

A flip-flop is a digital logic circuit whose basic function is memory, or storage. A flip-flop is capable of storing a single bit of binary data. It can assume either of two stable states, one representing a binary 1 and the other a binary 0. If the flip-flop is put into one of its two stable states, it will remain there as long as power is applied or until it is changed.

The Latch Flip-Flop

There are three basic types of flip-flops: the latch, the D type, and the JK. Let's start with the simplest, the latch or set-reset flip-flop, the simplest form of binary storage element. The symbol shown in Figure 5-17 is used to represent this type of flip-flop.
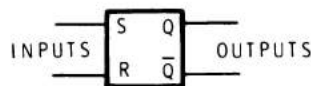
Figure 5-17
The latch flip-flop.

The flip-flop has two inputs, S and R, and two outputs Q and $\overline{Q}$. Applying the appropriate logic signal to either the S or R input will put the latch into one state or the other. The S input is used to **set** the flip-flop. When a flip-flop is set, it is said to be storing a binary 1. The R input is used to **reset** the flip-flop. A reset flip-flop is said to be storing a binary 0.

The latch has two outputs, labeled Q and $\overline{Q}$. These are called the **normal** and **complement** outputs respectively. As in other logic circuits, you can use any letter or alphanumeric combination to designate logic symbols. For example, the designation FF2, meaning flip-flop number 2, could be used as shown in Figure 5-18.
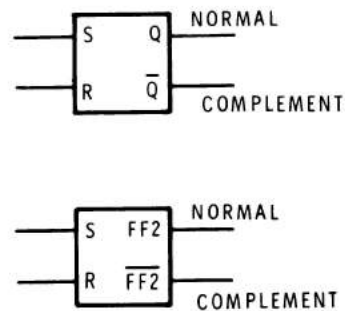
Figure 5-18
Alphanumerics can be used to label the normal and
complement outputs.

To tell what state the flip-flop is in, you look at the normal output. The logic level there tells you which bit, 0 or 1, is being stored. At the same time, the complement output has the state opposite that of the normal output.

A latch can easily be constructed with two NAND gates, as shown in Figure 5-19. Here, the two gates are wired back-to-back so that the output of one feeds the input to the other.
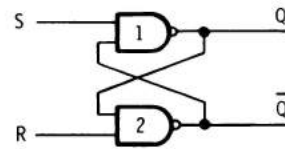
Figure 5-19
A NAND gate latch.

If the S and R inputs are both binary 1, the normal condition for the latch, the circuit is storing a bit put there earlier. For example, if the flip-flop is set, the normal (or Q) output from gate 1 will be at binary 1. This output is fed back to the upper input of gate 2. The lower input to gate 2 is also a binary 1, so its output ($\overline{Q}$) is low or binary 0.

The output from gate 2 is fed back to the lower input of gate 1. This input holds the Q output high (binary 1). You can see why this circuit is called a latch. Because of the feedback arrangement, the flip-flop is latched into this state. It will stay this way until you change it. The way you change it is by applying a low level to either the set or reset inputs.

If a binary zero or low level is applied to the R input of the latch, it will "flip" to the reset condition. The low level on the R input will force the output of gate 2 high. This will cause both inputs to gate 1 to be high, so its output will be low or binary 0. This indicates that the flip-flop is in the reset state. Any further low levels applied to the reset input will have no effect.

If a low level is applied to the S input, the latch will "flop" to the set condition. You can verify this by following the logic levels through the circuit of Figure 5-19. Therefore, the latch is like a toggle switch, it is either in one position or the other. And, once the change-over is made, repeating the action has no further effect.

However, with both S and R inputs low, the Q and $\overline{Q}$ outputs will **both** be high. No longer are the outputs complementary. Therefore, we really don't know what state the latch is in. It is in an ambiguous state and is neither set nor reset. This condition is one of the pecularities of a latch. When you are using it, you have to be careful to avoid simultaneous low inputs on the S and R terminals.

The operation of a latch can be summarized by the truth table of Figure 5-20, which accounts for all possible input and output states. Note that when both S and R inputs are binary 1, the output state of the flip-flop is designated X, where X can be either a 1 or a 0 as determined by previous input conditions. This is known as the "store" condition.

INPUTS		OUTPUTS		
S	R	Q	$\overline{Q}$	STATE
0	1	1	0	SET
1	0	0	1	RESET
1	1	X	$\overline{X}$	SET OR RESET "STORE" CONDITION
0	0	1	1	AMBIGUOUS CONDITION

Figure 5-20
A latch truth table.

D Type Flip-Flop

The symbol for the D type flip-flop is shown in Figure 5-21. Like any other flip-flop, it has two outputs that are used to determine its contents, that is, the outputs indicate what bit is stored there. For example, if the $\overline{Q}$ output is high, the Q or normal output is the complement or a binary 0. The Q output tells you the state of the flip-flop directly. Since it is a binary 0 in this example, the flip-flop is reset. The point here is that you read the outputs of a D flip-flop just as you would a latch or, for that matter, any other flip-flop.
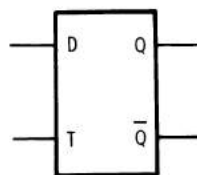
Figure 5-21
The D type flip-flop.

Now look at the inputs. Just as on the latch, there are two. However, these work differently. The D input is where you apply the data or bit to be stored. Of course, it can be either a binary 1 or a binary 0. The T input line controls the flip-flop. It is used to determine whether the input data is recognized or ignored. If the T input line is high (binary 1), the data on the D input line is stored in the flip-flop. If the T line is low (binary 0), the D input line is not recognized. The bit stored previously is retained. The D line can essentially do anything and it will be ignored if T is low.

You can get a better idea of how the D flip-flop works by taking a look at its circuitry. The logic diagram of one type of D flip-flop is shown in Figure 5-22.
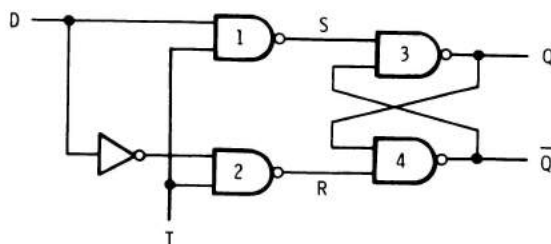
Figure 5-22
A D type flip-flop logic diagram.

In this circuit, gates 3 and 4 form a latch where the bit is stored. Gates 1 and 2 are enabling gates that pass or inhibit the input. The inverter makes sure that the S and R inputs to the latch are always complementary to avoid any possibility of the ambiguous state occurring.

Suppose a binary 1 is applied to the D input. If the T input is low (binary 0), the outputs of NAND gates 1 and 2 will be held at binary 1 regardless of the D input. This is the normal state for the inputs of a latch. Therefore, the latch will be undisturbed. Thus, a low on the T input prevents the flip-flop's state from changing and effectively "disconnects" the D input.

When the T input goes high, the D input determines the outputs of gates 1 and 2 and, therefore, the state of the latch. If the D input is high, tne inputs to gate 1 will both be high and the output will go low. The top input to gate 2 will be low due to the inverter. Therefore, the S input of the latch will be low and R input, high. The latch will go to the set condition. Thus, a binary 1 on the D input, along with a binary 1 on the T input, will set the flip-flop. If the T input now goes low, the D input will be disabled and the latch will remain in the set condition.

What happens when T input is high and the D input is low? In this case, both inputs to gate 2 will be high and its output will go low. The output of gate 1 will remain high. Therefore, the latch will reset.

The operation of a D type flip-flop is completely described by the truth table of Figure 5-23. Note that when T is binary 1, the Q output is the same as the D input. When T is binary 0, the Q output can be either binary 0 or 1, depending upon a previous input. This is indicated by the X state in the table. Note that a D flip-flop does not have an ambiguous state.

INPUTS		OUTPUTS	
D	T	Q	$\overline{Q}$
0	0	X	$\overline{X}$
0	1	0	1
1	0	X	$\overline{X}$
1	1	1	0

Figure 5-23
A D flip-flop truth table.

JK Flip-Flops

The JK flip-flop is the most versatile type of binary storage element in common use. It can perform all of the functions of the RS and D type flip-flops, plus several other things. Naturally, it is more complex and expensive than the other types, so for that reason it isn't always used where simpler and less expensive circuits will do.

A JK flip-flop is really two flip-flops in one. It usually consists of two latches, one feeding the other, with appropriate input gating on each. This is shown in Figure 5-24.
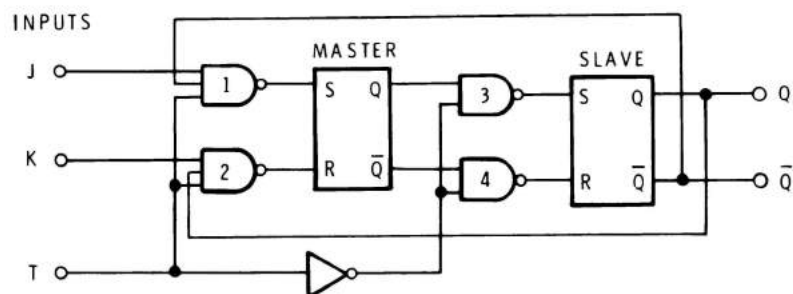
Figure 5-24
A JK flip-flop.

This arrangement is called a master-slave JK flip-flop. The master flip-flop is the input circuit. Logic signals applied to the JK flip-flop set or reset this master latch. However, the slave flip-flop is the latch from which the outputs are taken. The slave latch gets its input from the master latch. Both latches are controlled by a clock pulse*. Since there are **two** places to store bits in a JK flip-flop, there can be times when both master and slave latches are identical or times when they are complementary. But only the slave latch is responsible for indicating the state of the JK flip-flop.

The input of the master latch is controlled by gates 1 and 2. Gates 3 and 4 control the transfer of the master latch state to the slave latch. Input T, which is the clock signal, controls the input gating circuits just as in the D flip-flop. The inverter keeps the clock to the master and slave input gates complementary. The clock pulse controls the JK flip-flop while the J and K inputs determine exactly how it will be controlled. You will also see the T input referred to as CP or CK, designating clock pulse or clock.

Now let's see exactly how the J, K, and T (clock) inputs affect the flip-flop. Consider the time when the clock input is low. Gates 1 and 2 will be "inhibited" and their outputs will remain high. Therefore, the J and K inputs cannot control the state of the master latch when the T input is low.

The slave latch will have the same state as the master latch when the clock input is low. This is because the output of the inverter in the clock line is binary 1, causing gates 3 and 4 to be enabled during this time. Therefore, the state of the master latch is transferred to the slave latch. For example, if binary 1 is stored in the master latch, both inputs to gate 3 will be high, resulting in a low output. Gate 4 will have a low input from the master latch and, thus, a high output. You will recall that a low on the S input and a high on the R input of a latch causes it to set or to store a binary 1. Thus, the binary 1 has been transferred from the master latch to the slave latch.

Now if the clock T goes high, gates 1 and 2 will be enabled. The output of the inverter will inhibit gates 3 and 4. Therefore, the master latch cannot further change the slave latch. However, with gates 1 and 2 now enabled, the J and K inputs can affect the state of the master latch.

---

*A periodic signal that causes logic circuits to be stepped from one state to the next. These signals will be described in detail later in "Clocks and One Shots".

If both J and K inputs are low, the outputs of gates 1 and 2 will be held high, so no change takes place in the master latch.

If the J and K inputs are both high, then the state of the master latch will be determined by the Q and $\overline{Q}$ outputs, which are fed back to gates 1 and 2. For example, if the slave latch is set, the master latch will then be reset. If the slave is reset, the master will be set. The reason for this is the way the outputs are criss-crossed back to gates 1 and 2. Thus, with the J, K, and T inputs high, the state of the master latch will be determined by the Q and $\overline{Q}$ outputs.

The J and K inputs are similar to the set and reset inputs of a latch. If J is 1 and K is 0, the master latch will be set. If J is 0 and K is 1, the master latch will be reset. Remember, the T input line must be high for this to happen. When the T input goes low, gates 1 and 2 will be inhibited, while gates 3 and 4 will be enabled. Therefore, the contents of the master latch will be transferred to the slave latch the instant the T input goes low.

The operation of a JK flip-flop is summed up in the truth table of Figure 5-25. Note that only the normal (Q) output condition is shown. However, it is given twice, once prior to a clock pulse (t) and then after one clock pulse (t + 1). Output state X can represent either set (1) or reset (0); in any case, it represents the state that was previously stored in the flip-flop.

INPUTS		OUTPUTS	
J	K	Q(t)	Q(t+1)
0	0	X	X
0	1	X	0
1	0	X	1
1	1	X	$\overline{X}$

Figure 5-25
A JK flip-flop truth table.

Now, consider each of its input conditions. First, if both J and K inputs are at binary 0, input gates 1 and 2 will be disabled at all times and the JK flip-flop will retain its previous condition, an inhibit or store mode. If K is high (1) and J is low (0), when the clock (T) pulse goes high the master latch will be reset. When the clock pulse goes low, this reset condition will be transferred to the slave latch. Therefore, the flip-flop will be reset and the normal output (Q) will be a binary 0.

If J is high (1) and K is low (0), when the clock pulse goes high the master latch will be set. When the clock pulse goes low, the set condition will be transferred to the slave latch. The flip-flop is now set and the Q output will be a binary 1.

If both J and K are high, the feedback signals from the Q and $\overline{Q}$ outputs will affect the flip-flop. For example, if the flip-flop is set, a clock pulse will cause the master latch to reset. This occurs because the Q and $\overline{Q}$ signals are criss-crossed (see Figure 5-24). When the clock pulse goes low, the reset condition will be transferred to the slave latch. Now, if the J and K inputs remain high, the flip-flop will continue to complement itself with each clock pulse. This is known as the "toggle" mode of JK flip-flop operation.

The waveforms for the toggle mode are shown in Figure 5-26. Notice that the flip-flop only changes state when the T input goes from high to low; this corresponds to the transfer of data from the master to the slave latch. Also note that the output frequency is one-half the input frequency. Therefore, in the toggle mode, the JK flip-flop can be used as a two-to-one frequency divider. Several JK flip-flops can be cascaded to permit division by any factor of 2 (2, 4, 8, 16, 32, etc.)
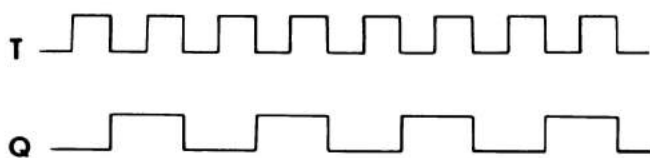
Figure 5-26
Waveforms for a JK flip-flop in the toggle mode.

This completes the description of basic, JK flip-flop operation. The symbol used to represent it is shown in Figure 5-27.
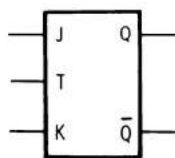
Figure 5-27
The JK flip-flop symbol.

Self-Review Questions

16. What is a flip-flop? _____

_____

_____

17. Name the three types of flip-flops.

1. _____
2. _____
3. _____

18. What type of flip-flop is shown in Figure 5-28?
_____.
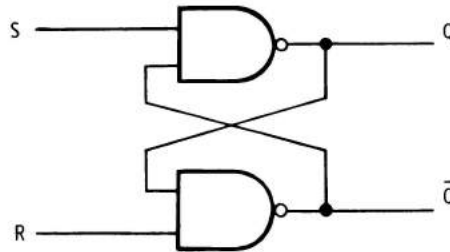
Figure 5-28

19. In Figure 5-28, if both inputs are low, what state will the flip-flop be in? What logic levels will appear at the Q and $\overline{Q}$ outputs?

_____

_____

20. What is the normal output of a flip-flop if it is set?

_____

21. The complement output of a flip-flop is low. What is the bit value stored? _____

22. What type of flip-flop is shown in Figure 5-29?
_____
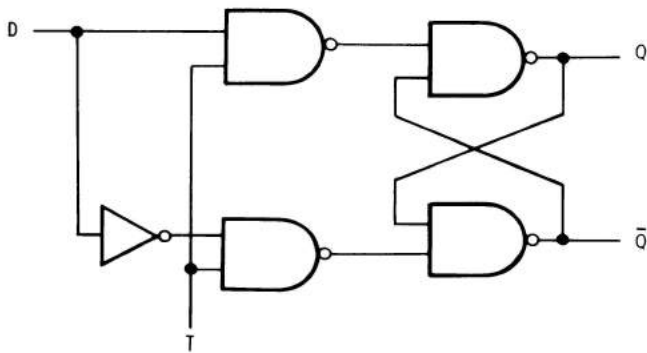
Figure 5-29

INPUTS		OUTPUTS	
D	T	Q	$\bar{Q}$
0	0		
0	1		
1	0		
1	1		

Figure 5-30

23. In Figure 5-29, if the T input is high and the D input is low, what will be the state of the complement output? What will be the state of the normal output?

_____
_____
_____

24. Complete the truth table shown in Figure 5-30 for the D type flip-flop.

25. On a JK flip-flop, the J input is high and the K input is low. What is the state of the flip-flop after one clock pulse occurs on the T input?

_____

26. The state of the JK flip-flop changes when the clock signal on the T input switches from _____.
                                   high to low/low to high

27. If both J and K inputs are low and the Q output is 0, what is the state of the flip-flop after three clock pulses? _____

28. If both J and K inputs are high and the $\overline{Q}$ output is 0, what is the state of the flip-flop after five clock pulses? _____

_____

29. In a JK flip-flop, both J and K inputs are at binary 1. The T input is a 50 kHz square wave. What is the Q output? _____

Self-Review Answers

16. A flip-flop is a digital logic circuit whose basic function is memory or storage.

17. The three types of flip-flops are the:

    1. Latch.
    2. D type.
    3. JK.

18. A NAND gate latch is shown in Figure 5-28.

19. If both inputs to a NAND latch are low, it will be in an ambiguous output state. Both the Q and $\overline{Q}$ outputs will be high.

20. Normal output of a set flip-flop is a binary 1.

21. If the complement output is low, the normal output is high. Therefore, a binary 1 is stored.

22. A D type flip-flop is shown in Figure 5-29.

23. If the T input is high and the D input is low, the complement output will be high and the normal output will be low.

24. See Figure 5-31.

INPUTS		OUTPUTS	
D	T	Q	$\overline{Q}$
0	0	X	$\overline{X}$
0	1	0	1
1	0	X	$\overline{X}$
1	1	1	0

Figure 5-31

25. With the J input high and K input low, after one clock pulse the flip-flop will be set.

26. The state of the JK flip-flop changes when the clock signal on the T input switches from **high to low**.

27. If both the J and K inputs are low, the flip-flop will be in the inhibit or store mode. Therefore, the flip-flop will not change state regardless of the number of clock pulses. The Q output will remain at 0.

28. If both the J and K inputs are high, the flip-flop will toggle. The $\bar{Q}$ output after five clock pulses will, therefore, be a binary 1.

29. With both the J and K inputs high, the flip-flop will toggle and divide the clock frequency by 2. The Q output will be a 25 kHz square wave.

COUNTERS AND SHIFT REGISTERS

A binary counter is a sequential logic circuit made up of flip-flops, that is used to count the number of binary pulses applied to it. The pulses or logic level transitions to be counted are applied to the counter input. These pulses cause the flip-flops in the counter to change state in such a way that the binary number stored in the flip-flops is representative of the number of input pulses that have occurred. By observing the flip-flop outputs, you can determine how many pulses were applied to the input.

There are several different types of counters used in digital circuits. The most commonly used is the binary counter. This circuit counts in the standard pure binary code. BCD counters, which count in the standard 8421 BCD code, are also widely used. In addition, both up and down counters are available.

The shift register is another widely used type of sequential logic circuit. Like a counter, it is made up of binary storage elements, usually flip-flops. These storage elements are cascaded in such a way that the bits stored there can be moved or shifted from one element to another adjacent element.

All of the storage registers in a shift register are actuated simultaneously by a single input clock, or shift, pulse. When a shift pulse is applied, the data stored in the shift register is moved one position in one of two directions. The shift register is basically a storage medium, where one or more binary words may be stored. However, its ability to move the data one bit at a time from one storage element to another makes the shift register valuable in performing a wide variety of logic operations. The following pages will give you detailed descriptions of shift registers and binary counters.

Binary Counters

A binary counter is a sequential logic circuit that uses the standard pure binary code. Such a counter is made up by cascading JK flip-flops, as shown in Figure 5-32. The normal output of one flip-flop is connected to the toggle (T) input of the next flip-flop. The JK inputs on each flip-flop are open or high. The input pulses to be counted are applied to the toggle input of the A flip-flop.
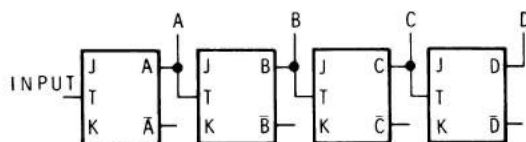
Figure 5-32
A 4-bit binary counter.

To see how this binary counter operates, remember that a JK flip-flop toggles, or changes state, each time a trailing edge transition occurs on its T input. The flip-flops will change state when the normal output of the previous flip-flop switches from binary 1 to binary 0. If we assume that the counter is initially reset, the normal outputs of all the flip-flops will be binary 0. When the first input pulse occurs, the A flip-flop will become set. The binary number stored in the flip-flops indicates the number of input pulses that have occurred. To read the number stored in the counter, you simply observe the normal outputs of the flip-flops. The A flip-flop is the least significant bit of the word. Therefore, the four-bit number stored in the counter is designated DCBA. After this first input pulse, the counter state is 0001. This indicates that one input pulse has occurred.

When the second input pulse occurs, the A flip-flop toggles and, this time, becomes reset. As it resets, its normal output switches from binary 1 to binary 0. This causes the B flip-flop to become set. Observing the new output state, you see that it is 0010, or the binary equivalent of the decimal number 2. Two input pulses have occurred.

When the third input pulse occurs the A flip-flop will again set. The normal output switches from binary 0 to binary 1. This transition is ignored by the T input of the B flip-flop. The number stored in the counter at this time then is 0011, or the number 3, indicating that three input pulses have occurred.

When the fourth input pulse occurs, the A flip-flop is reset. Its normal output switches from binary 1 to binary 0, thereby toggling the B flip-flop. This causes the B flip-flop to reset. As it does, its normal output switches from binary 1 to binary 0, causing the C flip-flop to become set. The number now in the counter is 0100, or a decimal 4. This process continues as the input pulses occur. The count sequence is the standard 4-bit binary code, as indicated in Figure 5-33.

Figure 5-33
Count sequence for a 4-bit binary counter.

D	C	B	A
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0
1	0	0	1
1	0	1	0
1	0	1	1
1	1	0	0
1	1	0	1
1	1	1	0
1	1	1	1

RECYCLE

An important point to consider is the action of the circuit when the number stored in the counter is 1111. This is the maximum value of a four bit number and the maximum count capacity of the circuit. When the next input pulse is applied, all flip-flops will change state. As the A flip-flop resets, the B flip-flop resets. As the B flip-flop resets it, in turn, resets the C flip-flop. As the C flip-flop resets, it toggles the D flip-flop, which is also reset to zero. The result is that the contents of the counter becomes 0000. As you can see from Figure 5-33, when the maximum content of the counter is reached, it simply recycles and starts its count again.

The complete operation of the four-bit binary counter is illustrated by the input and output waveforms in Figure 5-34. The upper waveform is a series of input pulses to be counted. Here they are shown as a periodic binary waveform; but, of course, it is not necessary for the input signal to be of a constant frequency or have equally spaced input pulses. The waveforms also show the normal output of each flip-flop.
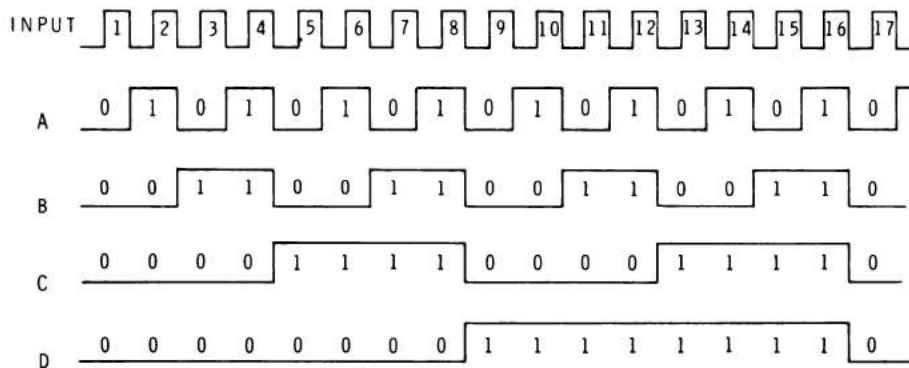
Figure 5-34
Input and output waveforms of a 4-bit binary counter.

In observing the waveforms in Figure 5-34, you should note several important things. First, all the flip-flops toggle (change state) on the trailing edge or the binary 1 to binary 0 transition of the previous flip-flop. With this in mind, you can readily trace the output of the first (A) flip-flop by simply observing when the trailing edges of the input occur.

The output of the B flip-flop is a function of its input, which is the output of the A flip-flop. Note that its state change occurs on the trailing edge of the A output. The same is true of the C and D flip-flops. The binary code after each input pulse is indicated on the waveforms. Of course, this corresponds to the binary count sequence in Figure 5-33.

Another important fact that is clear from the waveforms in Figure 5-34 is that the binary counter is also a frequency divider. The output of each flip-flop is one half the frequency of its input. If the input is a 100 kHz square wave, the outputs of the flip-flops are:

A — 50 kHz
B — 25 kHz
C — 12.5 kHz
D — 6.25 kHz

The output of a pure binary counter is always some sub-multiple of two. The four-bit counter divides the input by 16, (100 kHz ÷ 16 = 6.25 kHz).

DOWN COUNTER

The binary counter just described is referred to as an up-counter. Each time that an input pulse occurs, the binary number in the counter is increased by one. We say that the input pulses **increment** the counter. It is also possible to produce a down counter where the input pulses cause the binary number in the counter to decrease by one. In this case, the input pulses are said to **decrement** the counter.

The four-bit binary down counter is shown in Figure 5-35. It is practically identical to the up counter described earlier. The only difference is that the complement output rather than the normal output of each flip-flop is connected to the toggle input of the next flip-flop in sequence. This causes the count sequence to be the exact reverse of the up counter. The count sequence is illustrated in Figure 5-36. The waveforms associated with this counter are shown in Figure 5-37.
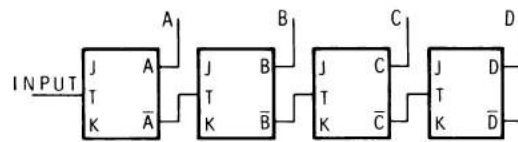
Figure 5-35
A 4-bit binary down counter.

D	C	B	A	
1	1	1	1	
1	1	1	0	
1	1	0	1	
1	1	0	0	
1	0	1	1	
1	0	1	0	
1	0	0	1	
1	0	0	0	RECYCLE
0	1	1	1	
0	1	1	0	
0	1	0	1	
0	1	0	0	
0	0	1	1	
0	0	1	0	
0	0	0	1	
0	0	0	0	

Figure 5-36
Count sequence for 4-bit down counter.
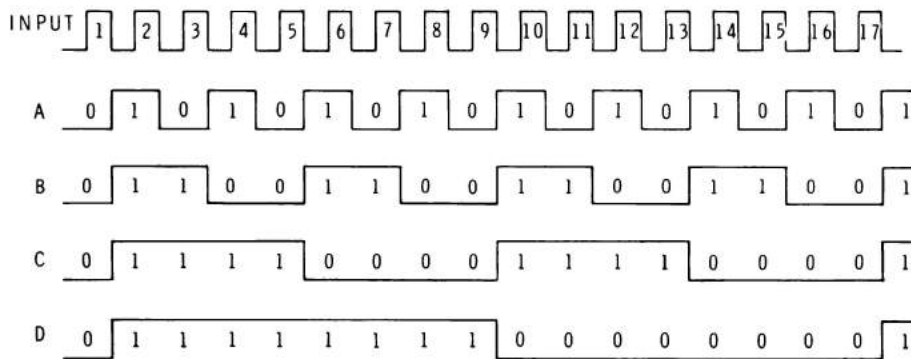
Figure 5-37
Input and output waveforms of a 4-bit binary down
counter.

In analyzing the operation of this counter, keep in mind that we still determine the contents of the counter by observing the normal outputs of the flip-flops as we did with the up-counter. Assuming the counter is initially reset and its contents are 0000, the application of an input pulse will cause all flip-flops to become set. With the A flip-flop reset, its complement output is high. When the first input pulse is applied, the A flip-flop will set. As it does, its complement output will switch from binary 1 to binary 0, thereby toggling the B flip-flop. The B flip-flop becomes set and its complement output also switches from binary 1 to binary 0. This causes the C flip-flop to set. In the same way, the complement output of the C flip-flop switches from high to low, thereby setting the D flip-flop. The counter recycles from 0000 to 1111.

When the next input pulse arrives, the A flip-flop will again be complemented. It will reset. As it resets, the complement output will switch from binary 0 to binary 1. The B flip-flop ignores this transition. No further state changes take place. The content of the counter then is 1110. As you can see, this input pulse causes the counter to be decremented from 15 to 14.

Applying another input pulse again complements the A flip-flop. It now sets. As it sets, its complement output switches from binary 1 to binary 0. This causes the B flip-flop to reset. As it resets, the complement output switches from binary 0 to binary 1. The C flip-flop ignores this transition. The new counter contents then is 1101, or 13. The input pulse again caused the counter to be decremented by one. By using the table in Figure 5-36 and the waveforms in Figure 5-37, you can trace the complete operation of the 4-bit binary down counter.

UP-DOWN COUNTER

The up-counting and down counting capabilities can be combined within a single counter as illustrated in Figure 5-38. AND and OR gates are used to couple the flip-flops. The normal output of each flip-flop is applied to gate 1. The complement output of each flip-flop is connected to gate 2. These gates determine whether the normal or complement signals toggle the next flip-flop in sequence. The count control line determines whether the counter counts up or down.
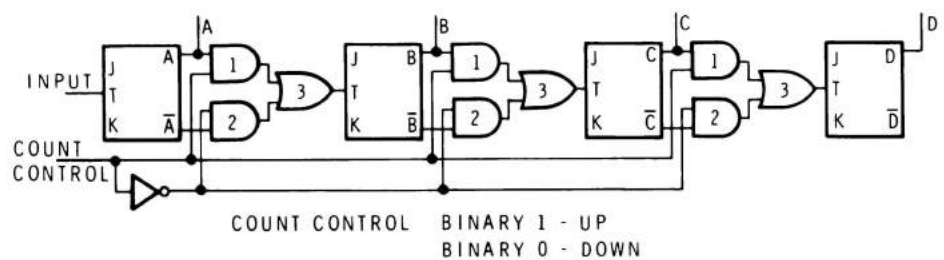
Figure 5-38
A binary up/down counter.

If the count control input is binary 1, all gate 1's are enabled. The normal output of each flip-flop then is coupled through gates 1 and 3 to the T input of the next flip-flop. The counter therefore counts up. During this time, all gate 2's are inhibited.

By making the count control line binary 0, all gate 2's are enabled. The complement output of each flip-flop is coupled through gates 2 and 3 to the next flip-flop in sequence. With this arrangement, the counter counts down.

BCD Counters

A BCD counter is a sequential circuit that counts by tens. It has ten discrete states, which represent the decimal numbers 0 through 9. Because of its ten-state nature, a BCD counter is also sometimes referred to as a decade counter.

The most commonly used BCD counter counts in the standard 8421 binary code. The table in Figure 5-39 shows the count sequence. Note that a four-bit number is required to represent the ten states 0 through 9. These ten four-bit codes are the first ten of the standard pure binary code. As count pulses are applied to the binary counter, the counter will be incremented as indicated in the table. Upon the application of a tenth input pulse, the counter will recycle from the 1001 (9) state to the 0000 state.

D	C	B	A
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0
1	0	0	1

RECYCLE

Figure 5-39
Count sequence of
8421 BCD counter.

An 8421 BCD counter constructed with JK flip-flops is shown in Figure 5-40. This counter will generate the BCD code given in the Table of Figure 5-39. Note that the counter consists of four flip-flops like the four-bit pure binary counter discussed earlier. The output of one flip-flop drives the T input to the next in sequence. Unlike the binary counter discussed earlier, however, this circuit has several modifications which permit it to count in the standard 8421 BCD sequence. The differences consist of a feedback path from the complement output of the "D" flip-flop back to the J input of the "B" flip-flop. Also, a two input AND gate monitors the output states of flip-flops B and C and generates a control signal that is used to operate the J input to the D flip-flop. These circuit modifications, in effect, trick the standard four-bit counter and cause it to recycle every ten input pulses.
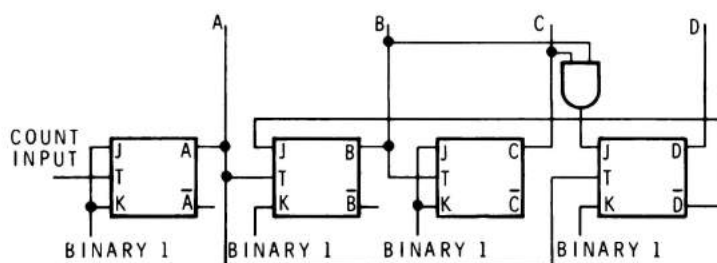
Figure 5-40
An 8421 BCD counter.

The waveforms shown in Figure 5-41 illustrate the operation of the 8421 BCD counter. The count sequence is identical to that of the standard four-bit pure binary counter discussed earlier for the first eight input pulses. The operations that occur during the 9th and 10th pulses are unique to the BCD counter.
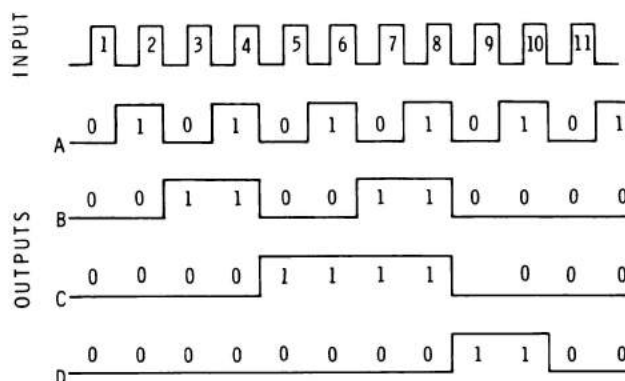
Figure 5-41
Waveforms of the 8421 BCD counter.

Assume that the counter in Figure 5-40 is initially reset. The outputs of flip-flops B and C will be binary 0 at this time. This makes the output of the AND gate low and causes the J input of the D flip-flop to be held low. The D flip-flop, cannot be set by the toggle input from the A flip-flop until the J input goes high. Note also that the complement output of the D flip-flop, which is binary 1 during the reset state, is applied to the J input of the B flip-flop. This enables the B flip-flop, permitting it to toggle when the A flip-flop changes state.

If count pulses are now applied, the states of the flip-flops will change as indicated in Figure 5-41. The count ripples though the first three flip-flops in sequence as in the standard 4-bit binary counter. However, consider the action of the counter upon the application of the 8th input pulse. With flip-flops A, B, and C set and D reset, the B and C outputs are high, thereby enabling the AND gate and the J input to the D flip-flop. This means that, upon the application of the next count input, all flip-flops will change state. The A, B, and C flip-flops will be reset while the D flip-flop is set. The counter state changes from 0111 to 1000 when the trailing edge of the 8th input pulse occurs.

In this new state, the B and C outputs are low, therefore causing the J input to the D flip-flop to again be binary 0. With the J input 0 and the K input binary 1 and the D flip-flop set, the conditions are right for this flip-flop to be reset when the T input switches from binary 1 to binary 0. In additon, the complement output of the D flip-flop is low at this time, thereby keeping the J input to the B flip-flop low. The B flip-flop is reset at this time and, therefore, the occurance of a clock pulse at the T input will not affect the B flip-flop.

When the 9th input pulse occurs, the A flip-flop sets. No other state changes occur at this time. The binary number in the counter is now 1001. The transition of the A flip-flop switching from binary 0 to binary 1 is ignored by the T input of the D flip-flop.

When the 10th input pulse occurs, the A flip-flop will toggle and reset. The B flip-flop will not be affected at this time since its J input is low. No state change occurs in the C flip-flop since the B flip-flop remains reset. The changing of the state of the A flip-flop, however, does cause the D flip-fop to reset. With its J input binary 0 and K input binary 1, this flip-flop will reset when the A flip-flop changes state. This 10th input pulse therefore causes all flip-flops to become reset. As you can see by the waveforms in Figure 5-41, the counter recycles from the 1001 (9) state to the 0000 state on the 10th input pulse.

Like any counter, the BCD counter can also be used as a frequency divider. Since the BCD counter has ten discrete states, it will divide the input frequency by ten. The output of the most significant bit flip-flop in the BCD counter will be one tenth of the input frequency. From Figure 5-40, you can see that only a single output pulse occurs at the D output for every ten input pulses. While the D output does not have a 50 percent duty cycle, the frequency of the signal is nevertheless one tenth of the input frequency.

Shift Registers

The illustration in Figure 5-42 shows you how a shift register operates. Here, the shift register consists of four binary storage elements, such as flip-flops. The binary number 1011 is currently stored in the shift register. Another binary word, 0110, is generated externally. As shift pulses are applied, the number stored in the register will be shifted out and lost while the external number would be shifted into the register and retained.

```
A   0 1 1 0 [1|0|1|1]          INITIAL
                               CONDITION
B     0 1 1 [0|1|0|1] 1        AFTER 1ST
                               SHIFT PULSE
C       0 1 [1|0|1|0] 1 1      AFTER 2ND
                               SHIFT PULSE
D         0 [1|1|0|1] 0 1 1    AFTER 3RD
                               SHIFT PULSE
E           [0|1|1|0] 1 0 1 1  AFTER 4TH
                               SHIFT PULSE
```
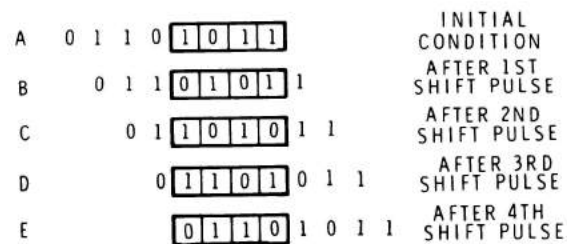
Figure 5-42
Operation of a shift register.

The initial conditions for this shift register are illustrated in Figure 5-42A. After one clock pulse, the number stored in the register initially is shifted one bit position to the right. The right-most bit is shifted out and lost. At the same time, the first bit of the externally generated number is shifted into the left-most position of the shift register, this is illustrated in Figure 5-42B. The remaining three illustations in C, D, and E show the results after the application of additional shift pulses. After four shift pulses have occurred, the number originally stored in the register has been completely shifted out and lost. The number appearing at the input on the left has been shifted into the register and now resides there.

An important point to note is that the data is shifted one bit position for each input clock, or shift, pulse. Clock pulses have full control over the shift register operation. In this shift register, the data was shifted to the right. However, in other shift registers, it is also possible to shift data to the left. The direction of the shift is determined by the application. Most shift registers are of the shift-right type.

The shift register is one of the most versatile of all sequential logic circuits. It is basically a storage element used for storing binary data. A single shift register made up of many storage elements can be used as a memory for storing many words of binary data.

Shift registers can also be used to perform arithmetic operations; shifting the data stored in a shift register to the right or to the left a number of bit positions is equivalent to multiplying or dividing that number by a specific factor. They can also be used to generate a sequence of control pulses for a logic circuit. And in some applications, shift registers can be used for counting and frequency dividing.

Shift registers are usually implemented with JK flip-flops. Type D flip-flops can also be used, but shift registers implemented with JK flip-flops are far more versatile. A typical shift register constructed with JK flip-flops is shown in Figure 5-43. The input data and its complement are applied to the JK inputs of the input (A) flip-flop. From there, the other flip-flops are cascaded with the outputs of one connected to the JK inputs of the next. Note that the clock (T) input lines to all flip-flops are connected together. The clock or shift pulses are applied to this line. Data applied to the input will be shifted to the right through the flip-flops. Each clock, or shift, pulse will cause the data at the input and that stored in the flip-flops to be shifted one bit position to the right.
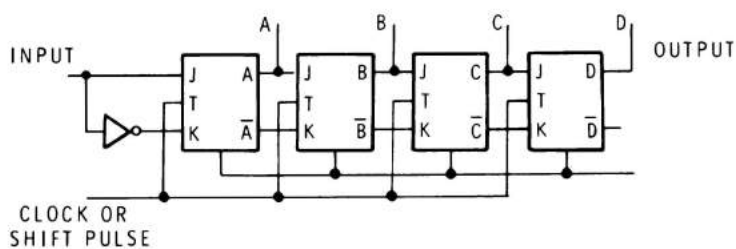
Figure 5-43
A 4-bit shift register made with JK flip-flops.

The waveforms in Figure 5-44 illustrate how a data word is loaded into the shift register of Figure 5-43. As the waveforms show, the binary number 0101 occurs in synchronization with the input clock, or shift, pulses. In observing the waveforms in Figure 5-44, keep in mind that time moves from left to right. This means that the clock pulses on the right occur after those on the left. In the same way, the state of the input shown on the left occurs prior to the states to the right. With this in mind, let's see how the circuit operates.
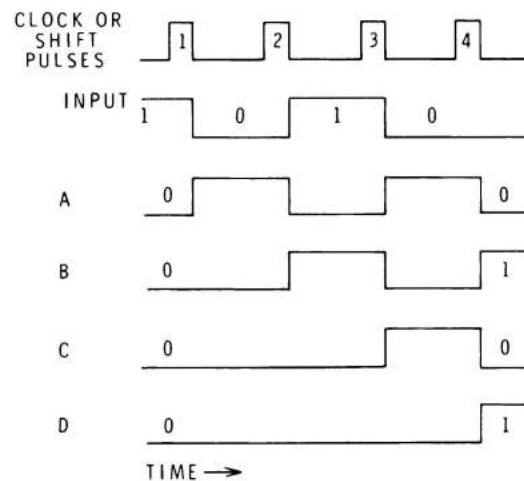
Figure 5-44
Waveforms illustrating how the serial binary number
0101 is loaded into a shift register.

Note that the shift register is originally reset. The A, B, C, and D outputs of the flip-flops, therefore, are binary 0 as indicated in the waveforms. Prior to the application of the number 1 shift pulse, the input state is binary 1. This represents the first bit of the binary word to be entered. On the trailing edge of the first clock pulse, the binary one will be loaded into the A flip-flop. The JK inputs of the A flip-flop are such that when the clock pulse occurs the flip-flop will become set. This first shift pulse is also applied to all other flip-flops. The state stored in the A flip-flop will be transferred to the B flip-flop. The states stored in the B and C flip-flops will be transferred to the C and D flip-flops respectively. Since all flip-flop states are initially zero, no state changes in the B, C, or D flip-flops will take place when the first clock pulse occurs.

After the first clock pulse, the A flip-flop is set while the B, C, and D flip-flops are still reset. The first clock pulse also causes the input word to change. The clock, or shift, pulses are generally common to all other circuits in the system; therefore, any data available in the system will generally be synchronized to the clock.

The input to the A flip-flop is now binary 0. When the trailing edge of the second clock pulse occurs, this binary 0 will be written into the A flip-flop. The A flip-flop, which was set by the first clock pulse, causes the JK inputs to the B flip-flop to be such that it will become set when the second clock pulse occurs. As you can see by the waveforms, when the second clock pulse occurs, the A flip-flop will reset while the B flip-flop will set. The 0 state previously stored in the B flip-flop will be transferred to the C flip-flop, and the C flip-flop state will be shifted to the D flip-flop. At this point the first two bits of the data word have been loaded into the shift register.

The input is now binary 1, representing the third bit of the input word. When the third clock pulse occurs, the A flip-flop will set. The zero previously stored in the A flip-flop will be transferred to the B flip-flop. The binary 1 stored in the B flip-flop will now be shifted into the C flip-flop. The D flip-flop remains reset.

The input to the A flip-flop is now binary 0. When the trailing edge of the fourth clock pulse occurs, the A flip-flop will reset. The binary 1 stored there previously will be transferred to the B flip-flop. The 0 stored in the B flip-flop will be shifted into the C flip-flop. The binary 1 in the C flip-flop now moves to the D flip-flop. As you can see, after four clock pulses have occurred, the complete four-bit binary word 0101 is now shifted into the register, as indicated by the states shown in the waveforms. A glance at the flip-flop output waveforms will show the initial binary 1 bit moving to the right with the occurrrence of each shift pulse.

While we have illustrated the operation of the shift registers with only four bits. Naturally, as many flip-flops as needed can be cascaded to form longer shift registers. Most shift registers are made up to store a single binary word. In most modern digital systems, shift registers have a number of bits that is some multiple of four.

Self Review Questions

30. Name the two most widely used types of sequential logic circuits.

    _____ _____

31. The circuit of Figure 5-45 is a _____-bit _____
    _____ counter.
    up/down
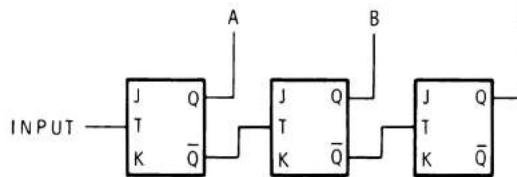
Figure 5-45
Identify this circuit.

32. The counter of Figure 5-45 contains the number 010. Six input pulses occur. The new counter state is _____.

33. A binary counter made up of five JK flip-flops will divide an input frequency by _____.

34. A BCD counter contains the number 1000. Six input pulses occur. The new counter state is _____.

35. A BCD counter divides its input signal frequency by _____.

36. A BCD counter is cascaded with a three-flip-flop binary counter. The overall frequency division ratio is _____.

37. What is a shift register? _____
_____
_____

38. An 8-bit shift register contains the number 10000110. The number 11011011 is applied to the input. After five shift pulses, what is the number in the shift register? (Assume shift right operation.)
_____

39. How many shift pulses are required to load a 16-bit word into a 16-flip-flop shift register? _____

Self Review Answers

30. The two most widely used types of sequential logic circuits are the **counter** and the **shift register**.

31. The circuit of Figure 5-45 is a **3-bit binary down counter**.

32. The new counter state is 000. The counter starts at 010 or decimal 2. After five input pulses, the counter will be at 111, or decimal 7. Therefore, the sixth input pulse will cause the counter to reset to 000.

33. A binary counter made up of five JK flip-flops will divide an input frequency by 32.

34. The new counter state is 0100. The counter starts at 1000, or decimal 8. After one input pulse, the counter will be at 1001, the maximum count of a BCD counter. After the second input pulse, the counter will reset to 0000. Now the counter will continue its count for the four remaining pulses to 0100, or decimal 4.

35. A BCD counter divides its input signal frequency by **10**.

36. The overall frequency division ratio is 80. The BCD counter divides by 10. The three flip-flop binary counter divides by 8. The total division is the product of the two, or $10 \times 8 = 80$.

37. A shift register is basically a storage element used for storing binary data.

38. The number in the shift register after five shift pulses is 11011100.
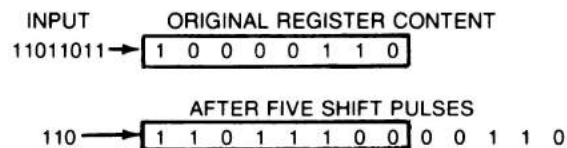
39. Sixteen shift pulses are required to load a 16-bit word into a 16-flip-flop shift registor.

CLOCKS AND ONE SHOTS

Most sequential logic circuits are driven by a clock, a periodic signal that causes logic circuits to be stepped from one state to the next through their normal operating states. The clock signal is generated by a clock oscillator circuit, which generates rectangular output pulses with a specific frequency, duty cycle, and amplitude.

Practically all digital clock oscillator circuits use some form of **astable** circuit to generate these periodic pulse waveforms. Such a circuit switches repeatedly between its two **un**stable states, hence the name astable.

Another circuit that is widely used, in addition to the clock, to implement sequential logic operations is known as the "one shot multivibrator," or "one shot." This circuit produces a fixed duration output pulse each time it receives an input trigger pulse. Its normal state is "off", until it receives a trigger pulse. Therefore, it is called a **monostable** circuit, since it has one stable state. The duration of its output pulse is usually controlled by external components.

This section discusses the very widely used 555 timer integrated circuit. It can be used either as a clock oscillator or as a one shot.

The 555 Timer

The 555 timer is a low cost linear IC which has dozens of different functions. It can act as a monostable circuit or as an astable circuit, and it has many other applications as well. However, this discussion will be limited to its astable and monostable modes of operation. For a more detailed discussion of the 555 and other IC timers, see the Heath "Electronics Technology Series" course number EE-103, "IC Timers."

Many different manufacturers produce the 555 Timer. Different versions have numbers like SE555, CA555, SN72555, and MC14555. However, you will notice that all contain the basic 555 number. Dual 555 timers are also available on a single chip and they carry "556" numbers.

A simplified diagram of the 555 circuit is shown in Figure 5-46. Notice that it contains two comparators, a flip-flop, an output stage, and a discharge transistor (Q_1). With the proper external components, several different functions can be implemented.
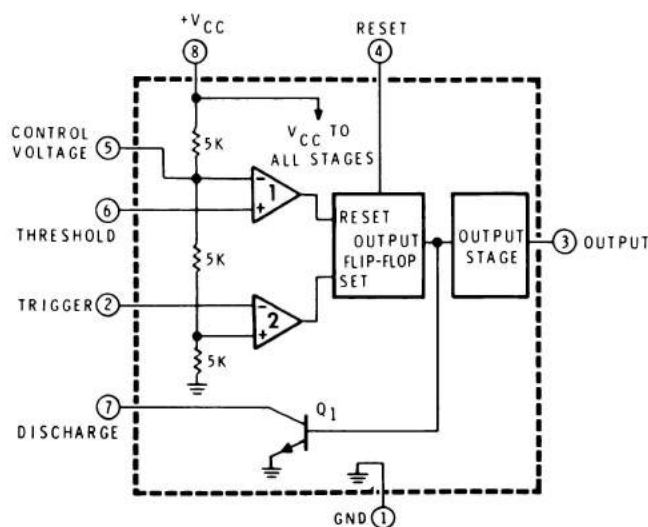
Figure 5-46
The 555 timer circuit.

If you examine the comparators in more detail, you will see that they are actually op amps with no feedback path. Therefore, they have an extremely high voltage gain. In fact, a few microvolts input will cause the output to drive to the full supply voltage or to zero volts, depending upon the input polarity.

In this circuit, a voltage divider consisting of three 5 Ω resistors develops a reference voltage at one input of each comparator. The reference voltage at the — input of comparator 1 is 2/3 of V_{cc}. The other input to the comparator comes from an external circuit via pin 6. When the voltage at pin 6 rises above the reference voltage, the output of comparator 1 swings positive. This, in turn, resets the flip-flop.

The reference voltage at the + input of comparator 2 is set by the voltage divider at 1/3 of V_{cc}. The other input to comparator 2 is the trigger input. When the trigger input falls below the reference voltage, the output of the comparator swings positive. This sets the flip-flop.

The output of the flip-flop will always be at one of two levels. When the flip-flop is reset, its output goes to a positive voltage which we will call +V. When set, its output falls to a very low voltage which we will call 0 volts.

The output of the flip-flop is amplified and inverted by the output stage. A load can be connected between the output terminal (pin 3) and either $+V_{cc}$ or ground. When the load is connected to $+V_{cc}$, a heavy current flows through the load when the output terminal is at 0 volts. Little current flows when the output is at +V. However, if the load is connected to ground, maximum current flows when the output is at +V and little current flows when the output is at 0 volts.

Notice that the output of the flip-flop is also applied to the base of Q_1. When the flip-flop is reset, this voltage is positive and Q_1 acts as a very low impedance between pin 7 and ground. On the other hand, when the flip-flop is set, the base of Q_1 is held at 0 volts. Thus, Q_1 acts as a high impedance between pin 7 and ground.

The 555 One Shot

Figure 5-47A shows the 555 timer being used as a one shot or monostable. This circuit produces one positive pulse output for each negative pulse at the trigger input. The duration of the output pulse can be precisely controlled by the value of external components C_1 and R_A.
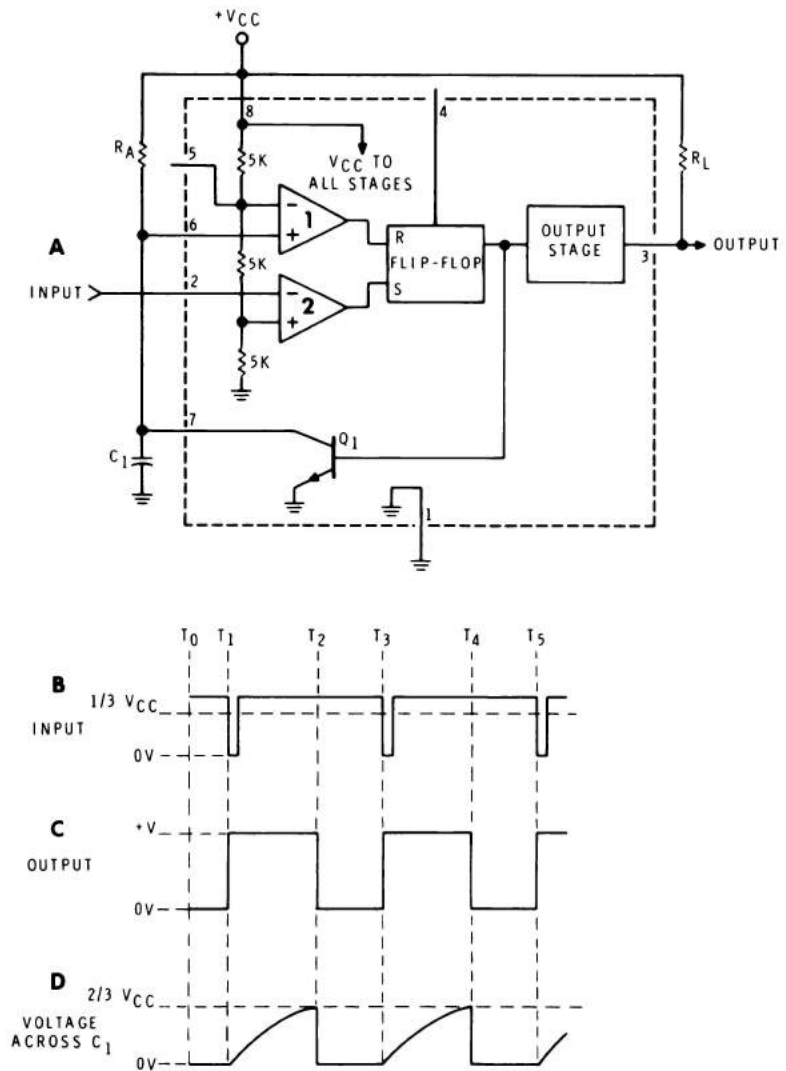
Figure 5-47
The 555 timer as a monostable circuit.

Figure 5-47B shows the input pulses. Between pulses, the input voltage is held above the trigger voltage of comparator 2. The flip-flop is reset and its output is at $+V$. This output is inverted by the output stage so pin 3 is at 0 volts. Thus, a heavy current flows through R_L. The output of the flip-flop $(+V)$ is also applied to the base of Q_1, causing the transistor to conduct. Q_1 acts as a short across C_1. These conditions are shown at time T_0. Notice that the output voltage (Figure 5-47C) and the capacitor voltage (Figure 5-47D) are at 0 volts at this time.
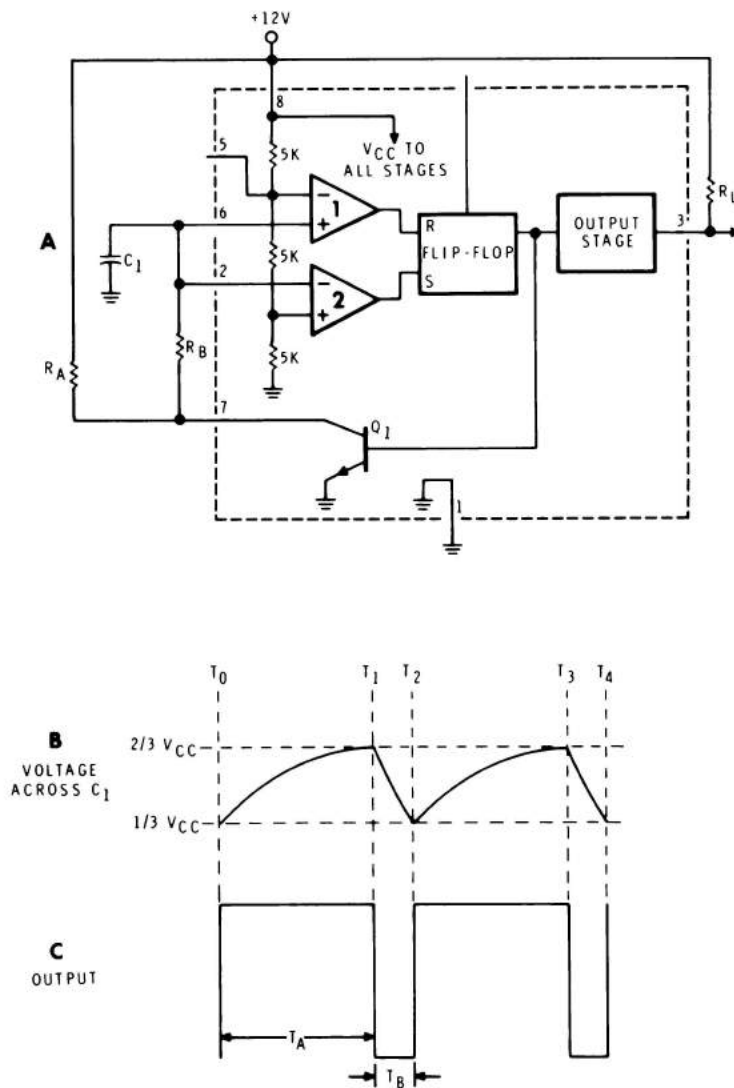
Figure 5-48
The 555 timer as an astable circuit.

At time T_1, a negative input pulse occurs. This forces the voltage at the $-$ input of comparator 2 below the 1/3 V_{cc} reference. The comparator switches states, setting the flip-flop. The output of the flip-flop falls to 0 volts. This voltage is applied to Q_1, cutting the transistor off. This removes the short from around C_1 and the capacitor begins to charge through R_4 toward $+V_{cc}$.

The output of the flip-flop is also applied to the output stage where it is inverted. Thus, the output at pin 3 swings to $+V$. The output will remain in this state until the flip-flop is reset.

In the circuit shown, you can only reset the flip-flop by switching the state of comparator 1. Between times T_1 and T_2, the voltage at the $+$ input of comparator 1 is below the 2/3 V_{cc} reference. Capacitor C_1 is charging toward this level and reaches it at time T_2. This switches the output of the comparator, resetting the flip-flop. The output of the flip-flop turns on Q_1 again, allowing C_1 to quickly discharge. Also, the output of the flip-flop is inverted and the voltage at pin 3 falls back to 0 volts.

The output is a positive pulse whose leading edge is determined by the input pulse. The duration of the pulse is determined by the time required for C_1 to charge to 2/3 of V_{cc}. This, in turn, is determined by the R_4C_1 time constant. C_1 can charge to 2/3 V_{cc} in just over one time constant. Thus the pulse duration is approximately:

$$PD = 1.1\ R_4C_1$$

The charge rate of C_1 and the threshold voltage of comparator 2 are both directly proportional to $+V_{cc}$. Thus, the pulse duration remains virtually constant regardless of the value of $+V_{cc}$.

If C_1 is a 0.01 μF capacitor and R_1 is a 1 megohm resistor, the pulse duration is

$$PD = 1.1\ (1 \times 10^6\ \Omega)\ (1 \times 10^{-8}F)$$
$$PD = 1.1\ (1 \times 10^{-2})$$
$$PD = 1.1 \times 10^{-2}\ \text{seconds or 11 milliseconds.}$$

The 555 Clock

Another application of the 555 timer is shown in Figure 5-48A. This is an astable circuit. It free runs at a frequency determined by C_1, R_A, and R_B. Figure 5-48B and C show the voltage across C_1 and the output voltage. Between times T_0 and T_1, the flip-flop is set and its output is 0 volts. This holds Q_1 cut off and holds the output (pin 3) at +V.

With Q_1 cut off, C_1 begins to charge toward $+V_{cc}$ through R_B and R_A. At time T_1, the voltage across the capacitor reaches 2/3 of $+V_{cc}$. This causes the flip-flop to reset. The output of the flip-flop goes to +V and the output at pin 3 drops to 0 volts. Q_1 conducts allowing C_1 to discharge through R_B. As C_1 discharges, the voltage across C_1 decreases. At time T_2, the voltage has decreased to the trigger level of comparator 2. This sets the flip-flop again, cutting off Q_1. The capacitor begins to charge once more and the entire cycle is repeated.

As you can see, the capacitor charges and discharges between 2/3 of $+V_{cc}$ and 1/3 of $+V_{cc}$. C_1 charges through both R_A and R_B. Approximately 0.7 time constants are required for C_1 to charge. Thus, the duration of the positive output pulse (T_A) is approximately

$$T_A = 0.7\ C_1\ (R_A + R_B)$$

Also, the duration of the negative going pulse is determined by the $C_1 R_B$ time constant. Consequently,

$$T_B = 0.7\ C_1 R_B$$

The total period of one cycle is

$$T = T_A + T_B$$

And, since frequency is the reciprocal of time,

$$f = \frac{1}{T} = \frac{1}{T_A + T_B} = \frac{1.43}{C_1\ (R_A + 2\ R_B)}$$

For example, if C_1 is a 0.01 μF capacitor, R_A is a 1 MΩ resistor, and R_B is a 100 kΩ resistor, the positive output pulse duration is:

$$T_A = 0.7\ C_1\ (R_A + R_B)$$

$$= 0.7\ (0.01\ \mu F)\ (1\ M\Omega + 100\ k\Omega)$$

$$= 0.7\ (1 \times 10^{-8}\ F)\ (1.1 \times 10^6\ \Omega)$$

$$= 7.7 \times 10^{-3}\ \text{seconds or 7.7 milliseconds}$$

The duration of the negative going output pulse is:

$$T_B = 0.7\ C_1\ R_B$$

$$= 0.7\ (0.01\ \mu F)\ (100\ k\Omega)$$

$$= 0.7\ (1 \times 10^{-8}\ F)\ (1 \times 10^5\ \Omega)$$

$$= 7 \times 10^{-4}\ \text{seconds or 0.7 milliseconds}$$

The total period of one cycle is then:

$$T = T_A + T_B$$

$$T = 7.7\ \text{milliseconds} + 0.7\ \text{milliseconds}$$

$$T = 8.4\ \text{milliseconds}$$

And the frequency is:

$$f = \frac{1}{T}$$

$$f = \frac{1}{8.4\ \text{milliseconds}}$$

$$f = 119\ \text{Hz}$$

Self-Review Questions

40. What is an astable circuit? _____

   _____

   _____

41. What is a monostable circuit? _____

   _____

   _____

42. In a 555 monostable circuit, if R_A is 2 kΩ and C_1 is 1 μF, what is the output pulse duration?

43. In a 555 astable circuit, if R_A is 100 kΩ, R_B is 500 kΩ and C_1 is 0.1 μF, what is the output frequency?

Self-Review Answers

40. An astable is a circuit that has two unstable states. It switches between these two states continuously, thus generating a periodic pulse waveform.

41. A monostable circuit has one stable state. When it is triggered, it produces a fixed duration output pulse and then returns to its stable state.

42. $PD = 1.1\ R_A C_1$

   $PD = 1.1\ (2\ k\Omega)\ (1\ \mu F)$

   $PD = 1.1\ (2 \times 10^3\ \Omega)\ (1 \times 10^{-6}\ F)$

   $PD = 2.2 \times 10^{-3}$ seconds

   $PD = 2.2$ milliseconds

43.

   $$f = \frac{1.43}{C_1\ (R_A + 2\ R_B)}$$

   $$f = \frac{1.43}{0.1\ \mu F\ [\,100\ k\Omega + 2\ (500\ k\Omega)\,]}$$

   $$f = \frac{1.43}{(0.1 + 10^{-6}\ F)\ (1.1 \times 10_6\ \Omega)}$$

   $$f = \frac{1.43}{0.11}$$

   $$f = 13\ Hz$$

Unit 6

DIGITAL COMPUTERS

CONTENTS

INTRODUCTION

Digital circuits were originally developed to provide a means of implementing digital computers. As new circuits and techniques were developed, computer performance was improved. But the greatest impact on digital computers has been the development of integrated circuits, or "IC's," which made them more powerful and greatly reduced their size and cost. Over the years, digital computers have continued to decrease in price. Their size and power consumption have also decreased significantly. At the same time, their performance and sophistication have increased, making them practical for a wider range of applications.

Recent technological advances in semiconductor techniques have created a unique digital product. Large scale integration of digital circuits have permitted the semiconductor manufacturers to put an entire digital computer on a single chip of silicon. These computers are known as microprocessors. We normally think of digital integrated circuits as being the gates and flip-flops used to implement a computer. Now, the computer itself is a single, low-cost integrated circuit. But the power of this device is significant, and for many applications it can replace hundreds of small scale and medium scale integrated circuits. This significant development will further broaden the applications for digital computers. Best of all, it will increase the sophistication and capabilities of the electronic equipment that uses them.

While it is impossible to cover all aspects of this exciting field in this unit, it will introduce you to the digital computer and its related techniques. The primary emphasis will be on the microprocessor and its ability to replace standard hard-wired digital logic systems.

The "Unit Objectives" are listed next. They state exactly what you are expected to learn in this unit. Review this list now and refer to it as you are completing this unit to be sure you meet each objective.

UNIT OBJECTIVES

When you have completed this unit you will be able to:

1. Define digital computer, data, instructions, program, software, and peripheral unit.

2. Name the four types of small computers.

3. List the four major sections of a digital computer and define each.

4. Define computer instruction, address, central processing unit, accumulator, and interrupt.

5. List the seven steps of computer programming.

6. Define machine language programming, algorithm, flow chart, coding, and loop.

7. Analyze a simple program, when given the computer instruction set.

8. Define subroutine, assembler, compiler, cross-assembler, cross-compiler, and utility program.

9. State the primary use and applications of the microprocessor.

10. List the benefits of using a microprocessor over a hard-wired logic system.

11. Write a simple microprocessor program when given the problem and computer instruction set.

WHAT IS A DIGITAL COMPUTER?

A **digital computer** is an electronic machine that automatically processes data by the use of digital techniques. **Data** refers to any information such as numbers, letters, words, or even complete sentences and paragraphs. Processing is a general term referring to a variety of ways in which the data can be manipulated. The computer processes the data by performing arithmetic operations on it, editing and sorting it, or evaluating its characteristics and making decisions based upon it. In addition to being able to manipulate data in a variety of ways, the computer contains an extensive memory where data is stored. The key characteristic of a digital computer is its ability to process data automatically without operator intervention.

The manner in which the data is manipulated is determined by a set of instructions, or "instruction set," contained within the machine. These instructions form a **program** that tells the computer exactly how to handle the data. The instructions are executed sequentially to carry out the desired manipulations. Most computers are general purpose, in that the instructions can be assembled into an almost infinite variety of application programs.

Each computer has a specific instruction set. These instructions are put into the proper sequence so they will perform the required calculation or operation. The process of writing the desired sequence of instructions is called programming.

How Computers are Classified

There are many different types of digital computers and a variety of ways in which they can be classified. One method of classifying computers is by size and computing power. At one end of the spectrum are large-scale computers with extensive memory and high-speed calculating capabilities. These machines can process huge volumes of data in a short period of time and in any desired manner. At the other end of the spectrum are the small-scale, low-cost digital computers, such as the microprocessor — whose application and computing power is more limited.

Computers are also classified by function or application. The most commonly known digital computer is the electronic data processor that is used by most business industry, and government organizations to maintain records, perform accounting functions, maintain an inventory, and provide a wide variety of other data processing functions. Then there are the scientific and engineering computers that are used primarily as mathematical problem solvers. They greatly speed up and simplify the calculations of complex and difficult scientific and engineering problems.

Another way to classify digital computers is general purpose or special purpose. General purpose machines are designed to be as flexible as possible. This means that they can be programmed for virtually any application. Special purpose computers, on the other hand, are generally dedicated to a specific application. They are designed to carry out only a single function. General purpose computers with a fixed program become special purpose computers.

Most digital computers are of the general purpose type, and most have versatile instruction sets so they can be programmed to perform almost any operation. With the proper program, a general purpose computer can perform business data processing functions, scientific and mathematical calculations, or industrial control functions.

The most widely used computers are the small-scale machines. These include the minicomputer, the microcomputer, the programmable calculator, and the microprocessor. While all of these small-scale machines together account for less than 10% of the total computer dollar investment, they represent more than 95% of the unit volume of computers. Small-scale computer systems are very low priced. Today, you can purchase a complete computer system for less than the price of a new automobile. Microprocessors and programmable calculators are even less expensive. There are many thousands of small computers in use today. Your own personal contact with a digital computer will no doubt be through some type of small-scale computer.

MINICOMPUTERS

The minicomputer is the largest of the four types of small computers. This is a general purpose digital computer, usually constructed of bipolar logic circuits, and supported with software and peripheral units. **Software** refers to the programs supplied with the computer that make it easy to use. **Peripheral units** are the input-output devices that allow an operator to communicate with the computer. Typical peripheral units are typewriters, card readers and printers. Minicomputers are similar to the larger digital computers, but their memory capacity, speed, and applications are more limited.

You can purchase a complete but minimum minicomputer for less than $1,000. This does not include peripheral equipment. However, such machines are often purchased to be built into a larger piece of equipment or a system for use as a controller. The users of such computers are referred to as original equipment manufacturers (OEM). A complete stand-alone minicomputer with sufficient memory, peripheral devices, and software to be used for general purpose computing may cost less than $5,000.

MICROCOMPUTERS

A microcomputer is similar in many respects to a minicomputer in that it is a general purpose machine that can be programmed to perform a wide variety of functions. However, the microcomputer is normally smaller and more restricted in its application. Its speed and memory capacity is less than a minicomputer. As a result, microcomputers are substantially less expensive than minicomputers. Microcomputers are more often used in dedicated, single function applications. Software and peripheral support is minimum. Most microcomputers are implemented with MOS LSI circuitry.

PROGRAMMABLE CALCULATOR

A programmable calculator can be classified as a special purpose microcomputer. These machines are similar in many respects to hand-held and desk-top electronic calculators. The programmable calculator has an input keyboard for entering data and a decimal display for reading out the results of calculations.

In a standard calculator, an operator enters the numbers to be manipulated and the functions to be performed by depressing keys on the keyboard in the proper sequence. The solution to the problems then appear on the display. A programmable calculator can also be used in this way, but it contains a memory and control unit that is used to automate the problem solving process. The data to be operated upon and the functions to be performed are entered via the keyboard and stored in the memory in the proper sequence. When it is enabled, the programmable calculator will then automatically solve the problem stored in its memory without operator control.

Programmable calculators offer the advantage of improved speed and convenience over standard calculators when the same problem must be computed several times with different data. Long problems requiring complex data and many mathematical operations are also best solved by a programmable calculator, as they relieve the operator from the tedious work and greatly minimize errors. Another advantage of the programmable calculator over other types of digital computers is its ability to communicate directly with the operator through the keyboard and decimal readout display.

MICROPROCESSORS

A microprocessor is the smallest and least expensive type of digital computer that still retains all of the basic features and characteristics of a computer. It can be implemented with standard digital integrated circuits or it is available as a single large scale integrated (LSI) circuit. While the capabilities of a microprocessor are limited when compared with a microcomputer or minicomputer, this device is still a very powerful unit. It extends the applications of computer techniques to many areas where minicomputers and microcomputers are not economically feasible.

Microprocessors are generally designed to perform a dedicated function. These devices are built into electronic equipment that will be used for some specific application. Some typical dedicated applications include traffic light controllers, electronic scales and cash registers, and electronic games. In addition, engineers are finding that low cost microprocessors can be used to replace standard hard-wired digital logic. Design time and cost can be significantly reduced in the design of a digital system when microprocessors are used.

A microprocessor can be used economically if the design is equivalent to thirty or more standard integrated circuit packages. Such hard wired logic designs are replaced by a microprocessor with a stored program. The program stored in a read only memory permits the microprocessor to carry out the same functions as a hard-wired logic controller. Microprocessors can also be used as the main component of a minicomputer or microcomputer.

Self Review Questions

1. What is a digital computer? _____
   _____
   _____

2. What is data? _____
   _____
   _____

3. The way the computer manipulates data is determined by a set of
   _____ .

4. A list of computer instructions for solving a particular problem is
   called a _____ .

5. What is software? _____
   _____
   _____

6. List several types of computer peripheral units. _____
   _____
   _____
   _____

7. Name the four types of small computers. _____
   _____
   _____
   _____

Self-Review Answers

1. A digital computer is an electronic machine that uses digital techniques to automatically process data.

2. Data is any information such as numbers, letters, words, or even complete sentences and paragraphs.

3. The way the computer manipulates data is determined by a set of **instructions.**

4. A list of computer instructions for solving a particular problem is called a **program**.

5. Software refers to the programs supplied with a computer.

6. Peripheral units are typewriters, card readers, printers, and other input-output devices.

7. The four types of small computers are:

> minicomputers
> microcomputers
> programmable calculators
> microprocessors

DIGITAL COMPUTER ORGANIZATION AND OPERATION

All digital computers are made up of four basic units: the memory, the control unit, the arithmetic logic unit (ALU), and the input-output (I/O) unit. These major sections and their relationship to one another are illustrated in Figure 6-1. An understanding of digital computer operation starts with a knowledge of how these sections operate and how they affect one another.
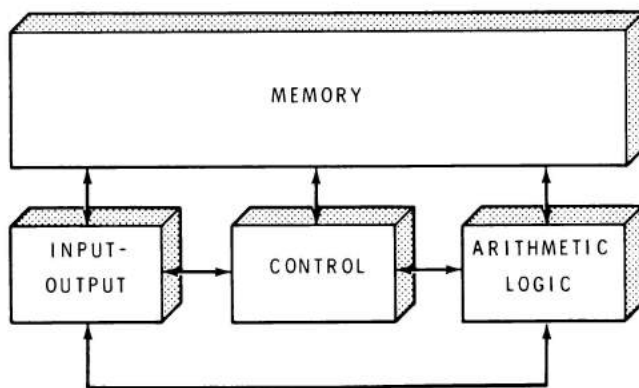
Figure 6-1
General block diagram of a digital computer.

Memory

The heart of any digital computer is its memory, where the program and data are stored. As indicated earlier, the program is a series of instructions that are stored and executed in sequence to carry out some specific function. The instructions cause the computer to manipulate the data in some way.

Computer memories are organized as a large group of storage locations for fixed length binary words. A computer instruction is nothing more than a binary word whose bit pattern defines a specific function to be performed. The data to be processed by the computer is also a binary word. A computer memory is an accumulation of storage registers for these instruction and data words. Most computers have memories capable of storing many thousands of words.

Digital computers typically have a fixed word size. A 32-bit word is common for many large computers. Minicomputers usually have a 16-bit word. Microprocessors widely use an 8-bit word. Memory sizes range from approximately several hundred words to several hundred thousand words of storage. A typical minicomputer may provide 4096 16-bit words. A microprocessor may use 1024 words of 8-bit memory. The number of words in memory is generally some power of two.

Each memory location appears to be like a storage register. Data can be loaded into the register and retained. The word can also be read out of memory for use in performing some operation.

Each memory word is given a numbered location called an **address**. The address is a binary word used to locate a particular word in memory. The normal procedure is to store the instruction words in sequential memory locations. The instruction word generally contains an address which refers to the location of some data word to be used in carrying out the operation specified. The instructions stored in the sequential memory locations are executed one at a time until the desired function is performed.

Most modern digital computers use semiconductor memory, MOS LSI circuits where data is stored in latch flip-flops or as the charge on a capacitor. Semiconductor memories are small, fast, and inexpensive. Many computers, however, still use magnetic core memories. In these memories, binary data is stored in tiny donut-shaped magnetic cores. By magnetizing the core in one direction a binary zero is stored. Magnetizing the core in the opposite direction causes it to store a binary one. Electronic circuitry associated with the cores is used to store data into the memory and read it out.

The advantage of core memories over semiconductor memories is their non-volatility. When power is removed from a semiconductor memory, all of the data is lost. Removing the power from a magnetic core memory has no effect on the data contents. Because the cores are permanently magnetized in one direction or the other, all data is retained.

The typical organization of a computer memory is shown in Figure 6-2. It consists of the semiconductor or magnetic core elements that retain the binary data. The memory used in a digital computer is generally referred to as a random access read/write memory. Random access refers to the ability of the computer to directly seek out and access any specific word stored in the computer memory. Read/write refers to the ability of the memory to store data (write) or to retrieve data for use elsewhere (read).

As you can see from Figure 6-2, the access to a specific word in memory is achieved through the memory address register (MAR) and memory address decoder. The memory address register is a flip-flop register into which is placed a multi-bit binary word that designates the location of a desired word in memory. If the address 0001 0011 is stored in the MAR, the content of memory location 19 is referenced. The address word may refer to the location of an instruction or a data word. The size of the address word determines the maximum memory size. For example, if the memory address word is 12-bits in length, the maximum number of words that the computer memory can contain is $2^{12} = 4096$ words (called a 4K memory).
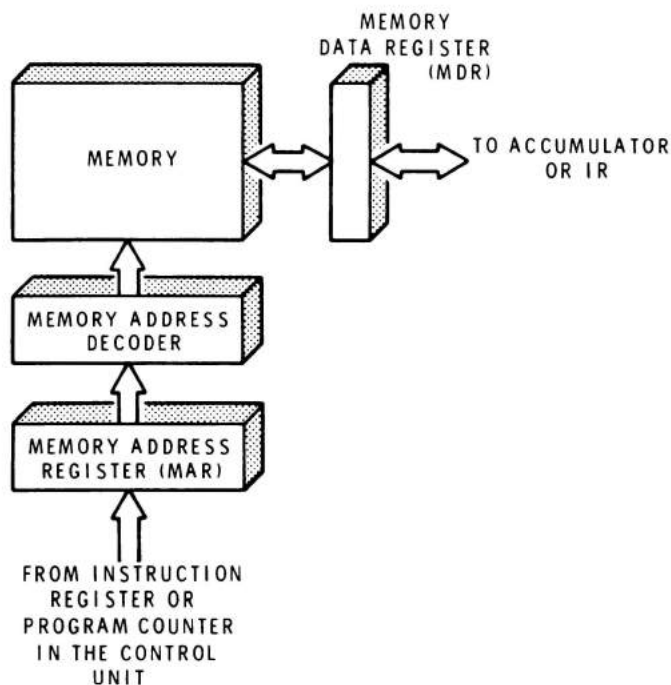
Figure 6-2
Typical computer memory organization.

The output of the memory address register drives the memory address decoder, which recognizes one unique memory address word at a time and enables the appropriate location. In semiconductor memories, the memory address decoder is generally a fixed part of the integrated circuit memory itself. When an address word is loaded into the MAR, the specific location in memory designated by that address is enabled. Data can then be written into or read out of that memory location.

The access to the addressed memory location is made through a memory data register (MDR) or memory buffer register (MBR). This is a flip-flop register into which the data or instruction word is stored on its way into or out of the memory. A word to be stored in memory is first loaded into the MDR and then stored in the addressed memory location. If a read operation is being carried out by the memory, the data stored in the addressed location is first loaded into the MDR. From there it is set to other portions of the computer as needed. Many computers do not use an MDR. Instead, the data or instruction goes to or comes from another register in the computer.

Control Unit

The control unit in a digital computer is a sequential logic circuit. Its purpose is to examine each of the instruction words in memory, one at a time, and generate the control pulses necessary to carry out the function specified by that instruction. The instruction, for example, may call for the addition of two numbers. In this case, the control unit would send pulses to the arithmetic logic unit to carry out the addition of the two numbers. If the instruction calls for the storage data in memory, the control unit would generate the necessary control pulses to carry out that storage operation. As you can see, it is the control unit that is responsible for the automatic operation of the digital computer.

Almost any type of sequential logic circuit can be used to implement the control unit. However, most modern digital computers incorporate a microprogrammed control unit using a preprogrammed circuit known as a read only memory (ROM). Here, special binary words known as microinstructions are stored in the read only memory. When an instruction is analyzed by the control unit, that instruction will cause a certain sequence of microinstruction words in the ROM to be executed. The result is the generation of logic signals that will carry out the operation designated by the instruction. The instruction set for any digital computer is defined by the operation of the control unit.

The exact logic circuitry used in the control unit varies widely from one machine to another. However, the basic elements are shown in Figure 6-3. The control unit consists of an instruction register, a program counter, an instruction decoder, a clock oscillator, and some type of sequential logic circuit used for generating the control pulses.
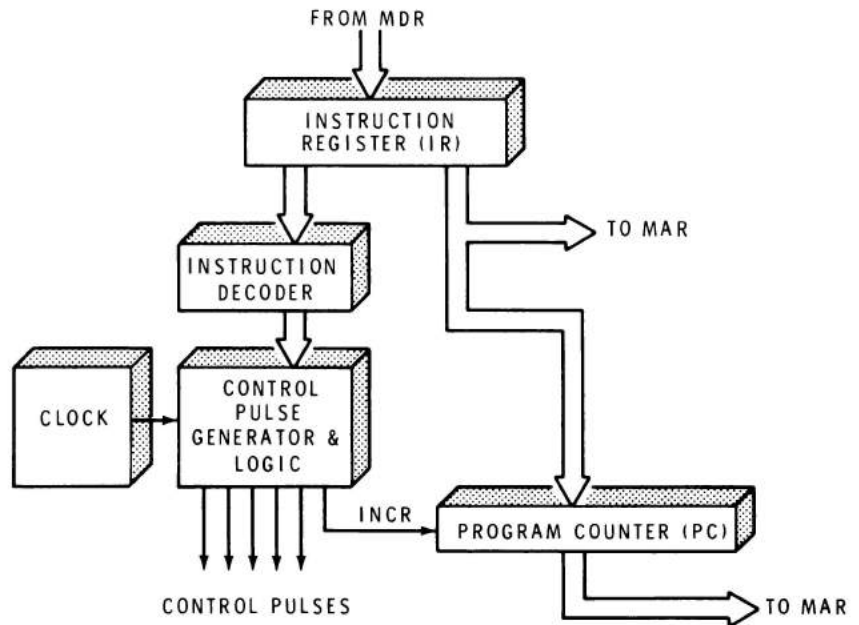
Figure 6-3
Typical control unit organization.

The instruction register is a multi-bit flip-flop register used for storing the instruction word. When an instruction is taken from memory, it passes through the MDR and then into the instruction register. From here, the instruction is decoded by the instruction decoder. This logic circuitry recognizes which instruction is to be performed. It then sends the appropriate logic signals to the control pulse generator. Under the control of the clock oscillator, the control pulse generator then produces the logic signals that will enable the other circuitry in the machine to carry out the specified instruction.

The program counter is simply a binary up counter that keeps track of the sequence of instructions to be executed. The program consists of instructions that are stored in sequential memory locations. To begin a program, the program counter is loaded with the starting address. The starting address is the location of the first instruction in the program to be executed. The first instruction is then read out of memory, interpreted, and carried out. The control circuitry then increments the program counter. The contents of the program counter is then fed to the memory address register that then permits the next instruction in sequence to be addressed. Each time an instruction is executed the program counter is incremented so that the next instruction in sequence is fetched and executed. This process continues until the program is complete.

In Figure 6-3, you will notice a connection between the instruction register and the program counter. There are times when the instruction itself will modify the contents of the program counter. Some instructions specify a jump or branch operation that causes the program to deviate from its normal sequential execution of instructions. The instruction register will contain an address that will be loaded into the program counter to determine the location to which the program jumps.

Arithmetic Logic Unit

The arithmetic logic unit (ALU) is that portion of the digital computer that carries out most of the operations specified by the instructions. It performs mathematical operations, logical operations, and decision-making functions. Most arithmetic logic units can perform addition and subtraction. Multiplication and division operations are generally programmed. The ALU can also perform logic operations such as inversion, AND, OR, and exclusive OR. In addition, the ALU can make decisions. It can compare numbers or test for specific quantities such as zero or negative numbers.

The arithmetic logic unit and control unit are very closely related, so much so that it is sometimes difficult to separate them. Because of this, the ALU and control unit together are often referred to as the central processing unit (CPU). Most microprocessors are single chip LSI CPUs.

The arithmetic-logic unit in a digital computer varies widely from one type of machine to another. Figure 6-4 shows the ALU circuitry associated with a very simple, minimum digital computer. The heart of the arithmetic-logic unit is the accumulator register. It is in this register where most of the computer operations take place. Here, the data is manipulated, computations are carried out, and decisions are made.
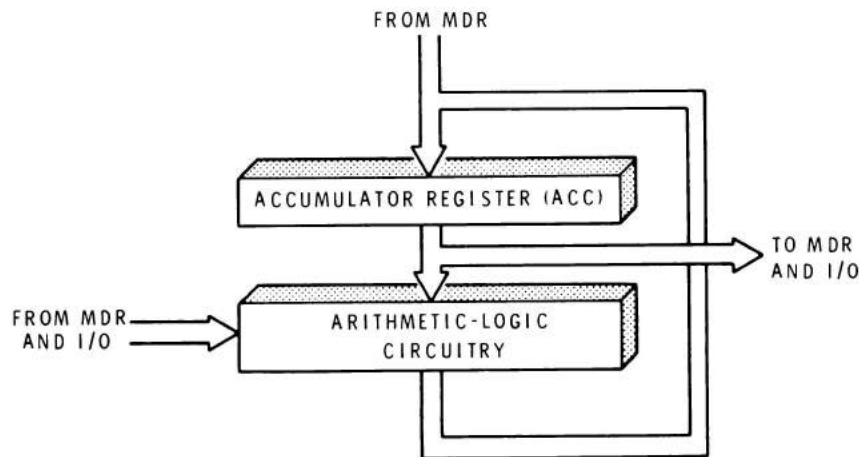
Figure 6-4
Arithmetic logic unit organization.

The accumulator register is a flexible unit that can usually be incremented and decremented. It can also be shifted right or shifted left. Many of the instructions define operations that will be carried out on the data stored in the accumulator register. The size of the accumulator register is generally determined by the basic computer word size, which is the same as the memory word size.

Associated with the accumulator register is the arithmetic logic circuitry. For the most part, this circuitry is a binary adder. With binary adder, both binary addition and subtraction can be accomplished. The arithmetic logic circuitry is also usually capable of carrying out logic operations such as AND, OR, and exclusive OR on the data stored in the accumulator register.

The arithmetic-logic circuitry is capable of adding two binary words. One of the binary words is stored in the accumulator. The other binary word is stored in the memory data register. The sum of these two numbers appears at the output of the arithmetic-logic circuitry and is stored in the accumulator register, replacing the number originally contained there. Most of the other operations with the arithmetic logic circuitry are carried out in this manner. The two words to be manipulated are initally stored in the accumulator and the MDR, with the results of the operation appearing back in the accumulator replacing the original contents.

Input-Output Unit

The input-output (I/O) unit of a computer is that section that interfaces the computer circuitry with the outside world. The term "outside world" refers to everything outside the computer. In order for the computer to communicate with an operator or with peripheral equipment, some means must be provided for entering data into the computer and reading it out. Data and programs to be stored in the memory are usually entered through the input-output unit. The solutions to calculations and control output signals are usually passed to the external equipment through the I/O unit.

The I/O unit is generally under the control of the CPU. Special I/O instructions are used to transfer data into and out of the computer. More sophisticated I/O units can recognize signals from extra peripheral devices called interrupts that can change the operating sequence of the program. Some I/O units permit direct communications between the computer memory and an external peripheral device without interference from the CPU. Such a function is called direct memory access (DMA).

The input/output section of a digital computer is the least clearly defined of all digital computer sections; it can vary from practically no circuitry at all to very complex logic circuitry approaching the magnitude of the remainder of the computer itself. For our explanation of digital computer operation here, we will assume the simplest form of input/output circuitry.

Data transfers between the computer and external peripheral devices take place via the accumulator register. Data to be inputted and stored in memory will be transferred a word at a time into the accumulator and then into the memory through the MDR. Data to be outputted is first transferred from the memory into the MDR, then into the accumulator, and finally to the external peripheral device. These data transfers into and out of the accumulator register take place under the control of the CPU and are referred to as programmed I/O operations. Special input/output instructions cause the proper sequence of operations to take place.

Most digital computers can also perform I/O operations at the request of an **interrupt**, a signal from an external device requesting service. The external device may have data to transmit to the computer or may require the computer to send it data. When an interrupt occurs, the computer completes the execution of its current instruction, and then jumps to another program in memory that services the interrupt. Once the interrupt request has been handled, the computer resumes execution of the main program. Data transfers occurring in the interrupt mode can also take place through the accumulator.

Digital Computer Operation

Now that you are familiar with the basic architecture of a digital computer, you are ready to see how the various sections operate together to execute a program. The units we described previously, when used together, actually form a simple hypothetical digital computer. In this section, it is used to demonstrate how a computer operates. A program for solving a problem is already stored in memory. The computer will execute each instruction until the problem is solved. The complete operation is described and the contents of each register shown as the program is carried out.

Assume that the problem to be solved is a simple mathematical operation that tells us to add two numbers, subtract a third number, store the result, print the answer, and then stop. The numbers that we will work with are 36, 19, and 22. The program calls for adding 36 and 19, subtracting 22, and then storing and printing the answer.

The solution to this simple problem as it is solved step-by-step by the computer is illustrated in Figure 6-5. Here, we show a simplified block diagram of the digital computer, emphasizing the memory and the major registers. The program is stored in memory. The contents of each memory location, either instruction or data, is shown adjacent to the memory address. To solve this problem, the computer sequentially executes the instructions. This is done in a two-step operation. First, the instruction is fetched or read out of memory. Second, the instruction is executed. This fetch-execute cycle is repeated until all of the instructions in the program have been executed.

MEMORY

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	
10	HALT

| 0 | ACC |

| LOAD 6 | MDR |

| LOAD 6 | IR |

| MAR | 0 |

| PC | 0 |
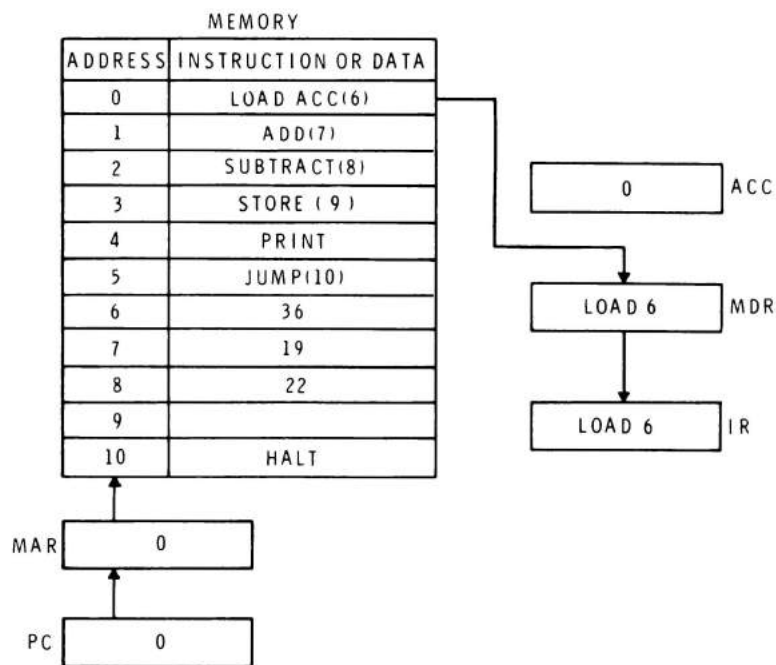
Figure 6-5
Fetch first instruction (LOAD).

In Figure 6-5, the first instruction of the program is fetched. The instruction word is read out of memory and appears in the memory data register (MDR). It is then transferred to the instruction register (IR) where it is interpreted. Note that the memory address register (MAR) contains 0, which is the address of the first instruction. The accumulator register (ACC) is set to 0 prior to the execution of the program.

Figure 6-6 shows the execution of the first instruction, LOAD ACC (6), which tells us to load the accumulator with the data stored in memory location 6. In executing this instruction, the number 36 is transferred to the accumulator. Note how this is done. The address specified by the instruction word (6) is transferred from the instruction register to the memory address register (MAR). This causes the number 36 stored in that location to be transferred to the MDR and then to the accumulator. During this step, the program counter (PC) is incremented by one so the next instruction in sequence will be fetched.
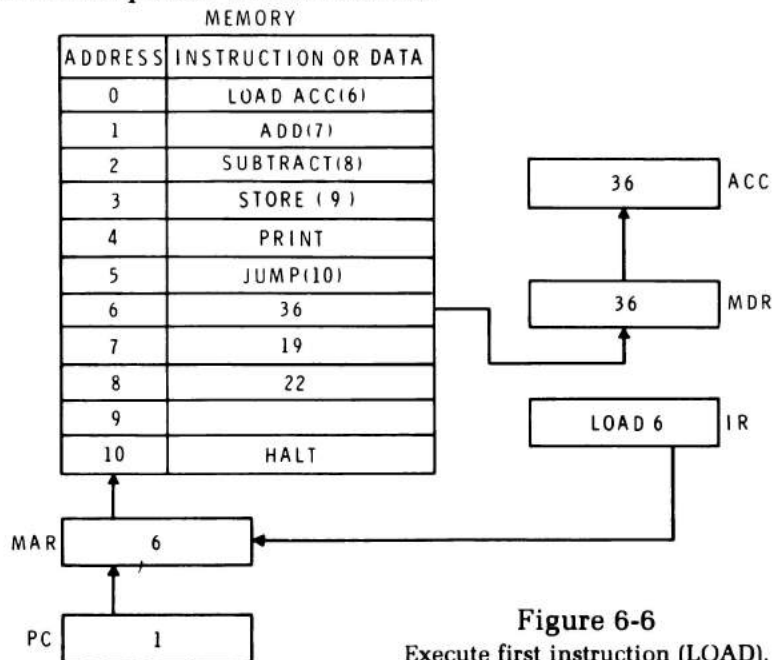
Figure 6-6
Execute first instruction (LOAD).

Figure 6-7 shows the fetch operation for the second instruction. The contents of the program counter is transferred to the MAR so that the ADD(7) instruction is fetched. This instruction passes though the MDR into the instruction register.

The execution of the add instruction is shown in Figure 6-8. This instruction tells us to add the contents of memory location 7 to the contents of the accumulator. The address of the add instruction is transferred to the MAR. This causes the contents of memory location 7, the number 19, to be transferred to the MDR. The contents of the MDR are added to the contents of the accumulator with the sum appearing back in the accumulator. As you can see, the sum of 36 and 19 is 55. Note that the program counter is again incremented so that the next instruction in sequence will be fetched.

MEMORY

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	
10	HALT

36   ACC

ADD 7   MDR

ADD 7   IR

MAR   1

PC   1
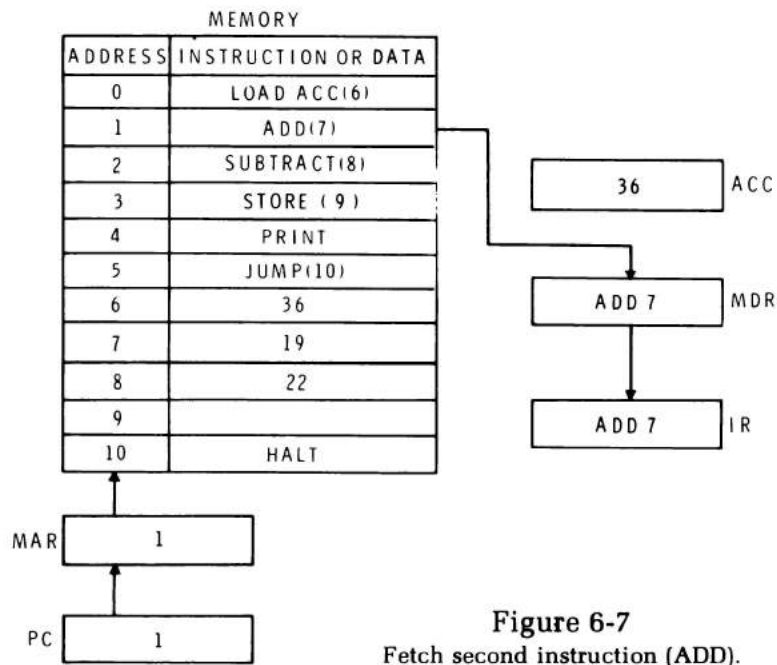
Figure 6-7
Fetch second instruction (ADD).

MEMORY

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	
10	HALT

55   ACC

19   MDR

ADD 7   IR

MAR   7

PC   2
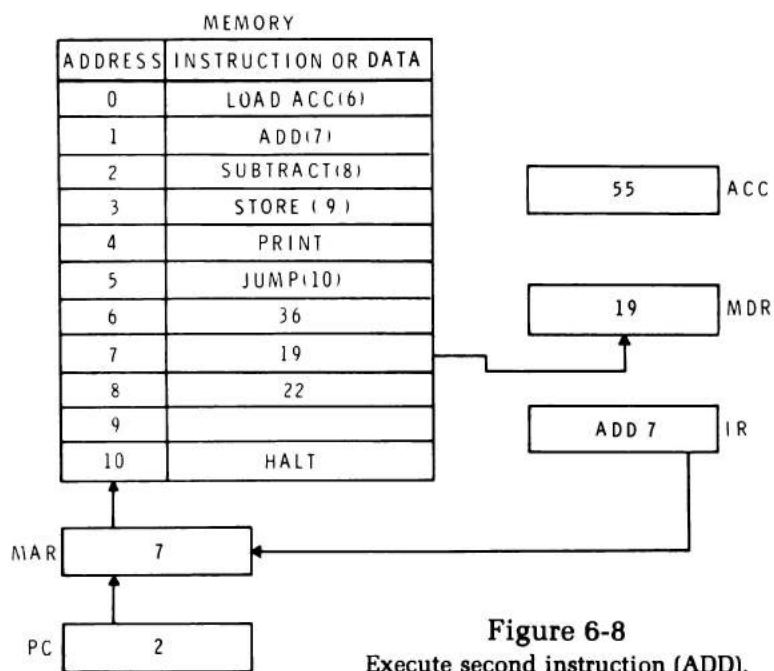
Figure 6-8
Execute second instruction (ADD).

The remaining instructions in the program are fetched and executed in a similar manner. The third instruction, a subtract, causes the memory contents of location 8 to be subtracted from the contents of the accumulator with the resulting remainder appearing in the accumulator. This produces an answer of 33. This fetch-execute sequence is shown in Figures 6-9 and 6-10.

MEMORY

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	
10	HALT

55	ACC

SUB 8	MDR

SUB 8	IR

MAR	2

PC	2
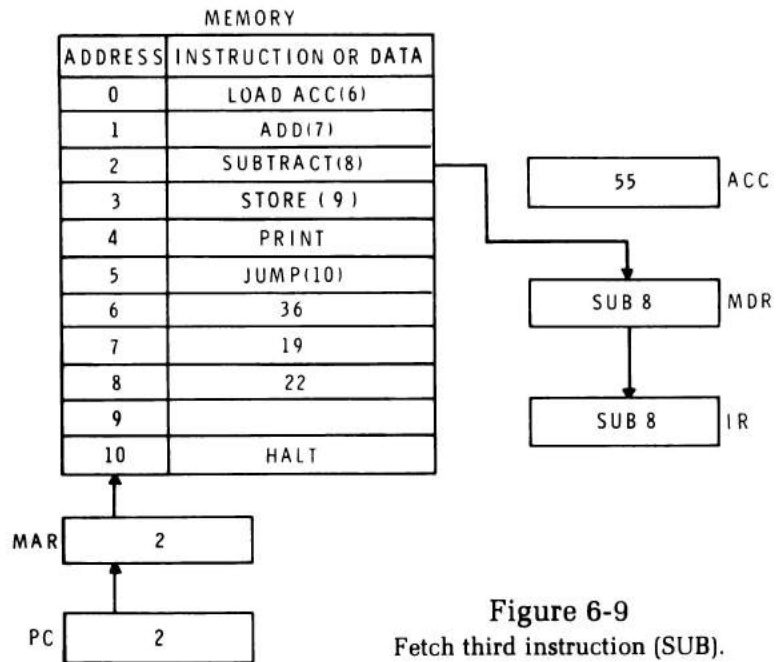
Figure 6-9
Fetch third instruction (SUB).

MEMORY

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	(ANSWER)
10	HALT

33	ACC

22	MDR

SUB 8	IR

MAR	8

PC	3
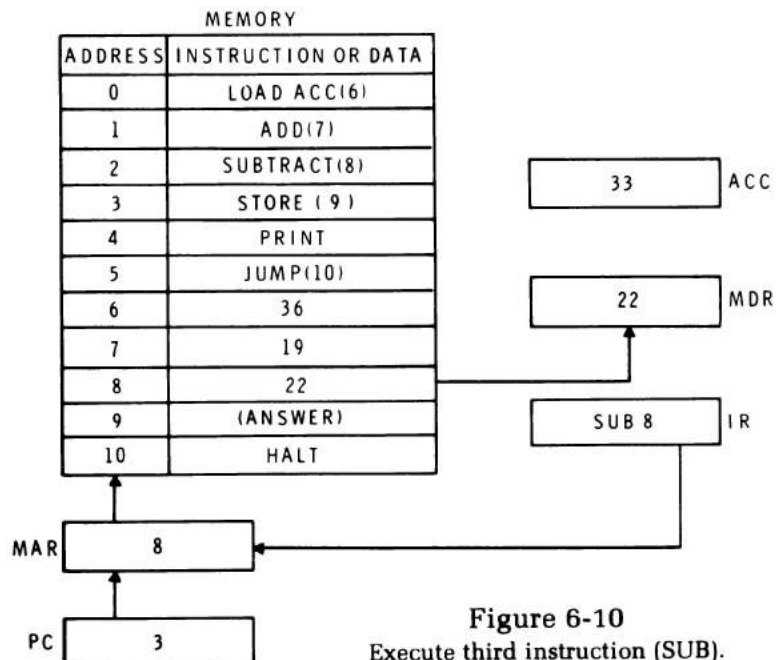
Figure 6-10
Execute third instruction (SUB).

The next instruction in sequence, STORE(9), tells us to store the contents of the accumulator in memory location 9. The number 33 in the accumulator is transferred to the MDR and stored in location 9, as indicated in Figures 6-11 and 6-12.

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	
10	HALT

33	ACC

STORE 9	MDR

STORE 9	IR

MAR | 3

PC | 3
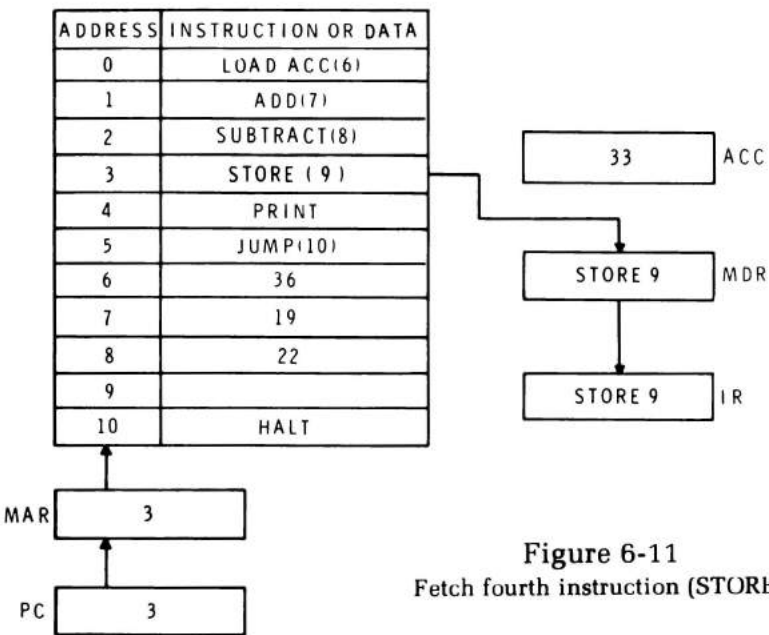
Figure 6-11
Fetch fourth instruction (STORE).

MEMORY

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	33
10	HALT

33	ACC

33	MDR

STORE 9	IR

MAR | 9

PC | 4
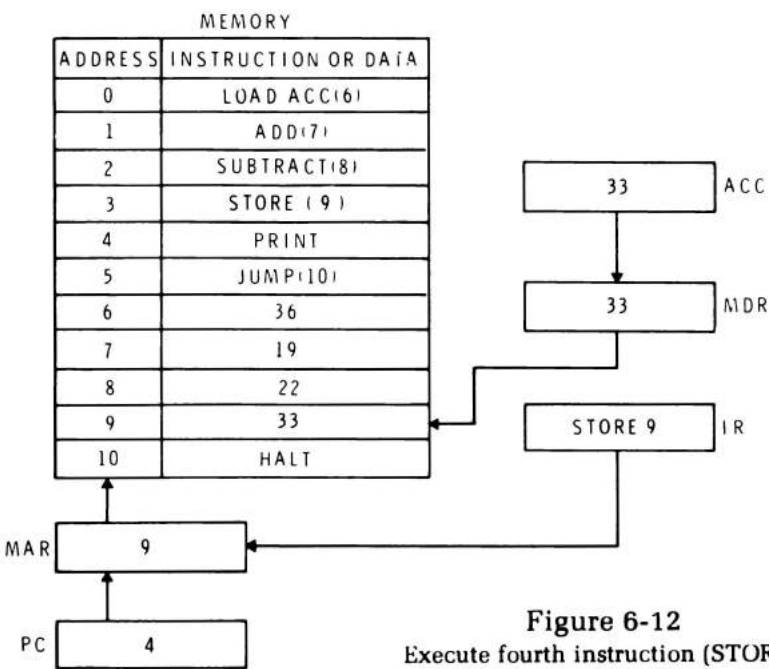
Figure 6-12
Execute fourth instruction (STORE).

The fifth instruction in the program, PRINT, tells us to print the contents of the accumulator on the external printer. The number stored in the accumulator will then be transferred to a printer where it is printed. The fetch-execute cycle for this operation is shown in Figures 6-13 and 6-14.
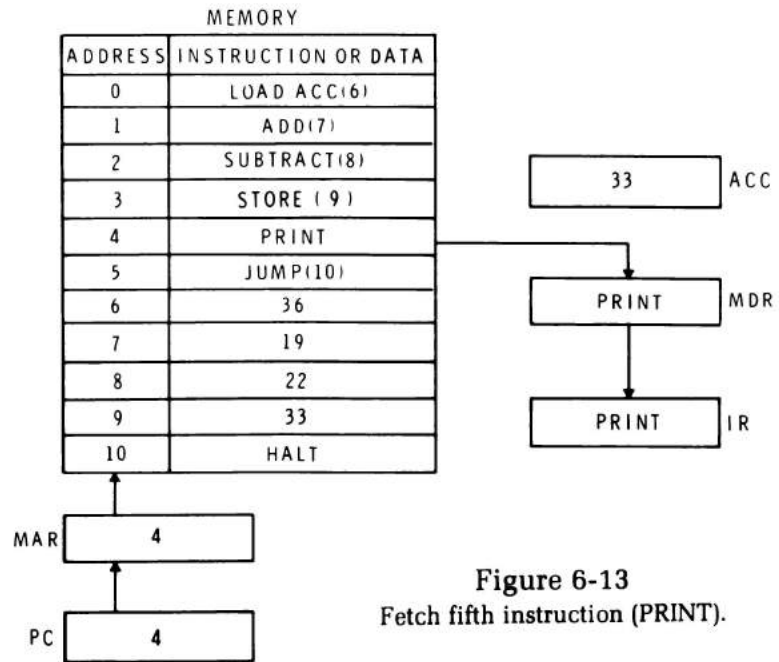
Figure 6-13
Fetch fifth instruction (PRINT).
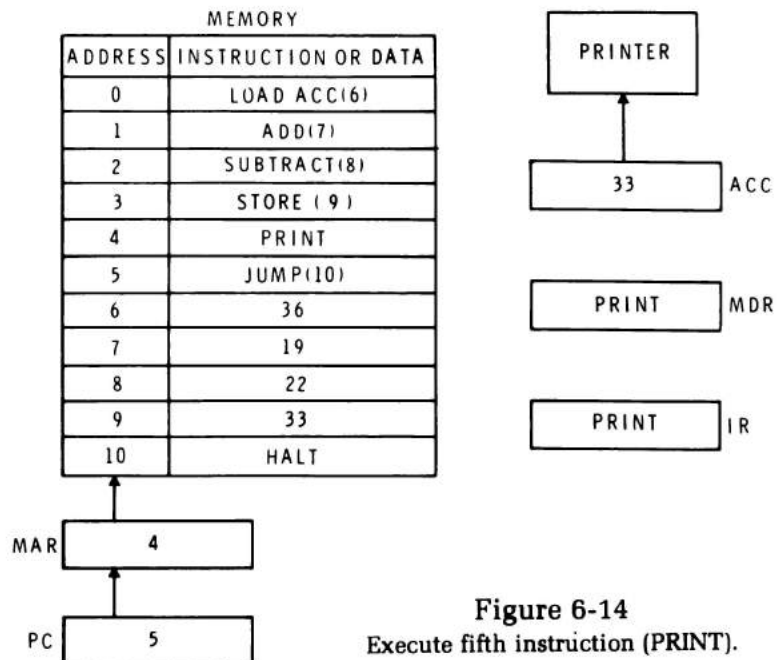
Figure 6-14
Execute fifth instruction (PRINT).

The sixth instruction in sequence is a JUMP(10) instruction that causes the normal sequence of program executions to change. The jump instruction tells us not to execute the contents of the next memory location in sequence. Instead, it tells us to take the next instruction from memory location 10. You can see by referring to Figure 6-15 that the contents of the next memory location in sequence (address 6) contains a data word.
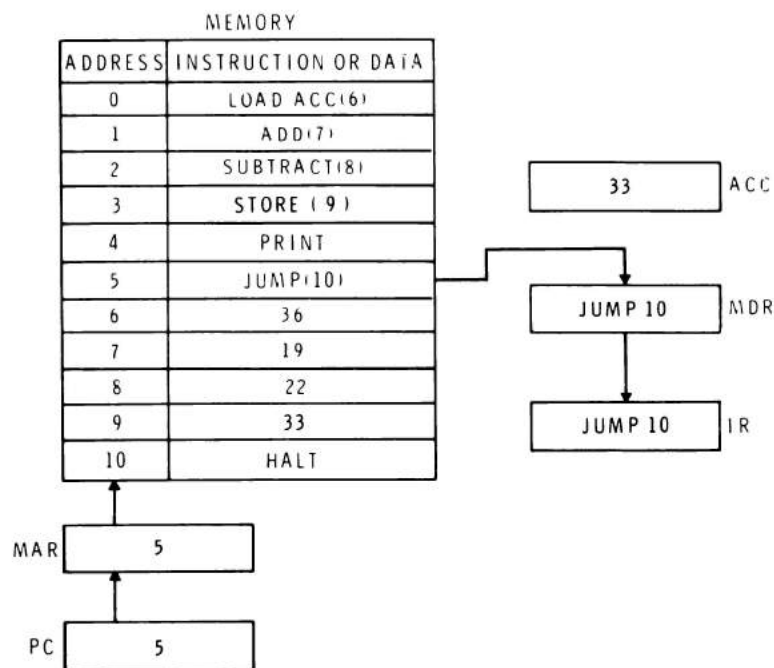
Figure 6-15
Fetch sixth instruction (JUMP).

The computer, being a dumb machine, would simply interpret a data word as an instruction and attempt to execute it. If this ever happens, the result of the computation will be erroneous. The purpose of the jump instruction in our program is to jump over the data words in the program stored in locations 6, 7, 8, and 9. The program is continued in location 10, where a HALT instruction is stored. By executing the jump instruction, the program counter is loaded with the address portion of the jump instruction (10) instead of being incremented as it normally is. See Figure 6-16. This causes the computer to fetch and execute the instruction stored in location 10. This is illustrated in Figures 6-17 and 6-18.

The last instruction in the program is a HALT. This instruction has no effect other than to stop the operation of the machine. Note in Figure 6-18 that the program counter was incremented so that it contains the memory location (11) of the next instruction in sequence to be fetched.

Study the program shown in Figures 6-5 to 6-18. Trace through each of the fetch and execute cycles for each instruction to be sure that you fully understand the operation. All digital computers operate in this same way with minor variations.
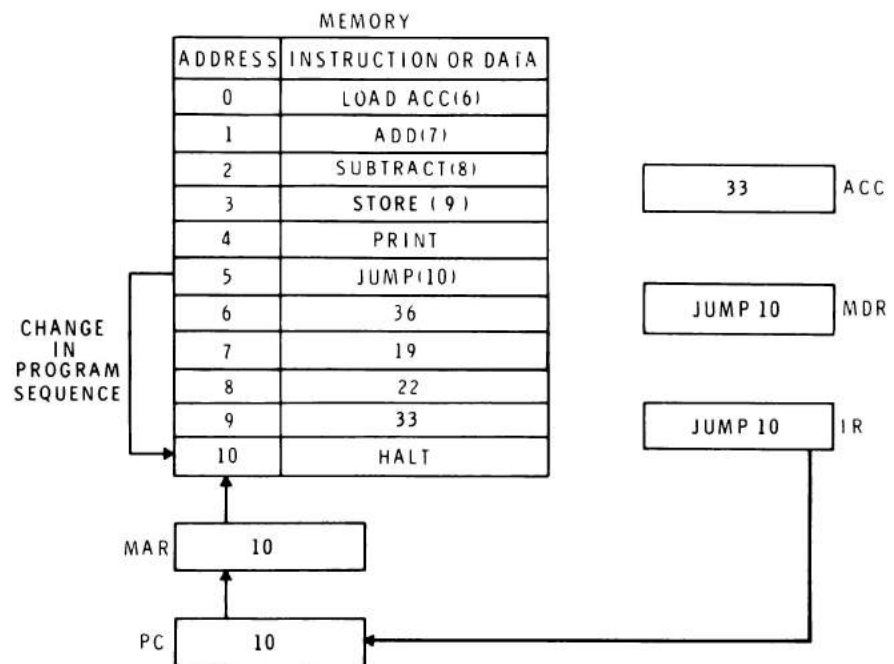
Figure 6-16
Execute sixth instruction (JUMP).

MEMORY

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	33
10	HALT

| 33 | ACC |

| HALT | MDR |

| HALT | IR |

MAR | 10 |

PC | 10 |
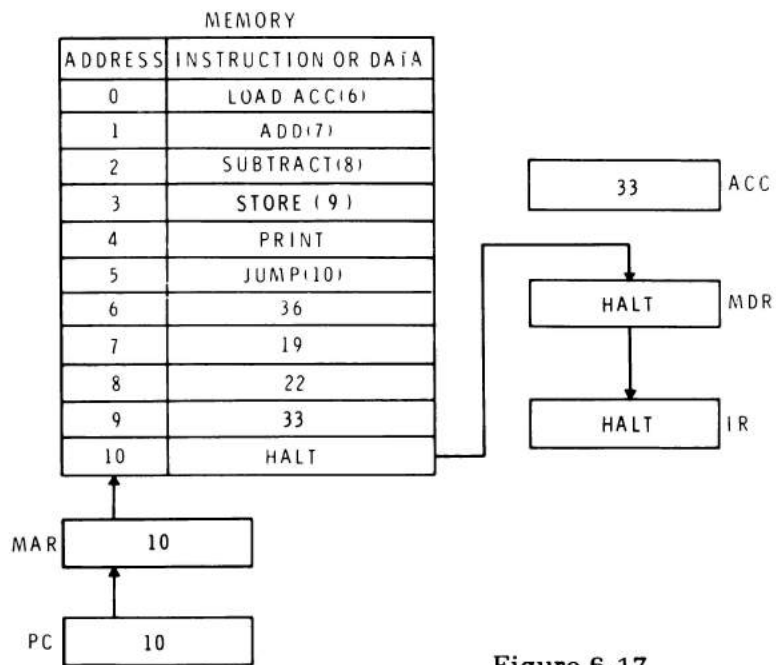
Figure 6-17
Fetch seventh instruction (HALT)

MEMORY

ADDRESS	INSTRUCTION OR DATA
0	LOAD ACC(6)
1	ADD(7)
2	SUBTRACT(8)
3	STORE (9)
4	PRINT
5	JUMP(10)
6	36
7	19
8	22
9	33
10	HALT

| 33 | ACC |

| HALT | MDR |

| HALT | IR |

MAR | 10 |

PC | 11 |
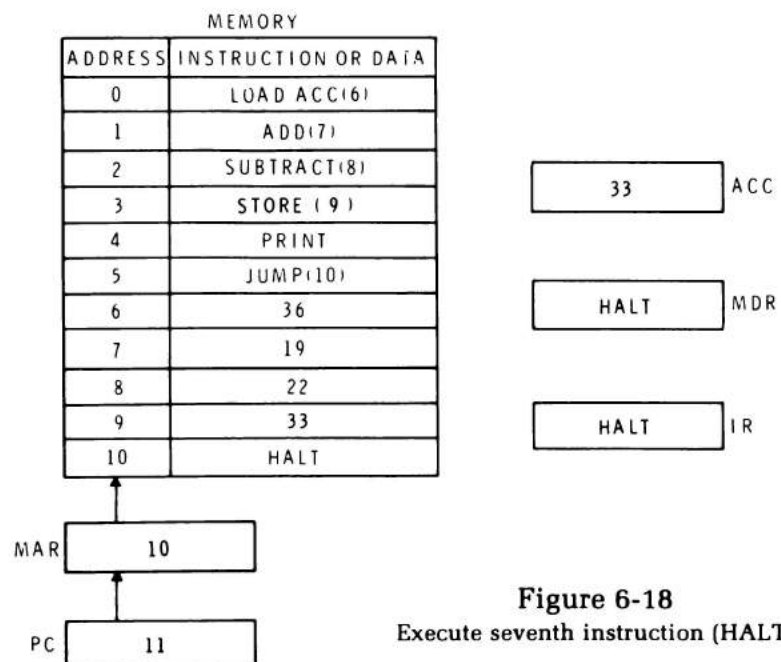
Figure 6-18
Execute seventh instruction (HALT)

Self-Review Questions

8. The four major sections of a digital computer are:

1. _____
2. _____
3. _____
4. _____

9. The program and data are stored in the _____.

10. What is a program? _____
_____
_____

11. A binary word whose bit pattern defines a specific function to be performed by the computer is called a _____.

12. A binary word used to locate a particular word in memory is called an _____.

13. List the four basic sections of the memory.

1. _____
2. _____
3. _____
4. _____

14. What is the control unit of a computer and what is its purpose?

_____

_____

_____

15. What is the arithmetic logic unit?

_____

_____

16. The main computational and data manipulation register in a computer is the _____.

17. The arithmetic logic unit and control unit combined are referred to as the _____ _____.

18. What is the input-output unit of a computer? _____

_____

_____

19. What is an interrupt? _____

_____

_____

20. In carrying out a program, the computer repeats a series of _____ and _____ operations on the instructions in memory.

Self-Review Answers

8. The four major sections of a digital computer are:

   1. Memory
   2. Control Unit.
   3. Arithmetic Logic Unit
   4. Input-Output Unit

9. The program and data are stored in the **memory**.

10. A program is a series of instructions that are stored and executed in sequence to carry out some specific function.

11. A binary word whose bit pattern defines a specific function to be performed by the computer is called a **computer instruction**.

12. A binary word used to locate a particular word in memory is called an **address**.

13. The four basic sections of memory are:

    1. The memory itself
    2. Memory address register (MAR)
    3. Memory address decoder
    4. Memory data register (MDR)

14. The control unit is a sequential logic circuit. Its purpose is to examine each instruction word and generate the control pulses needed to carry out the specific function.

15. The arithmetic logic unit is the portion of the computer that carries out most of the operations specified by the program instructions.

16. The main computational and data manipulation register in a computer is the **accumulator**.

17. The arithmetic logic unit and control unit combined are referred to as the **central processing unit** (CPU).

18. The input-output unit of a computer is the section that interfaces the computer circuitry with the outside world.

19. An interrupt is a signal from an external device requesting service.

20. In carrying out a program, the computer repeats a series of **fetch** and **execute** operations on the instructions in memory.

COMPUTER PROGRAMMING

A digital computer without a program is useless. The logic circuitry making up the computer is incapable of performing any useful end function without a program. It is this characteristic of a digital computer that sets it apart form other types of digital circuitry. And it is this characteristic that makes the digital computer the versatile machine that it is. For this reason, a discussion of digital computers is not complete without information on programming.

The process of using a digital computer is mainly that of programming it. Whether the computer is a simple microprocessor or a large scale system, it must be programmed in order for it to perform some useful service. The application of the computer will define the program. Programming is the process of telling the computer specifically what it must do to satisfy our application.

Programming is a complex and sophisticated art. In many ways, it is almost a field apart from the digital circuitry and the computer hardware itself. There are many different levels of programming and many unique methods that are employed. For that reason, it is impossible to cover them all here. The purpose of this section is to give you an overview of the process of programming a computer. The emphasis will be on programming small scale digital computers such as the microprocessor.

Programming Procedure

There are many different ways to program a digital computer. The simplest and most basic form of programming is **machine language programming**. This is the process of writing programs by using the instruction set of the computer and entering the programs in binary form, one instruction at a time. Programming at this level is difficult, time consuming, and error prone. It also requires an in-depth understanding of the computer organization and operation. Dispite these disadvantages, however, this method of programming is often used for short simple programs. While most computer applications do not use machine language programming, it is desirable to learn programming at this level. It helps to develop a through knowledge of machine operation and generally results in the shortest, most efficient programs. Many microprocessors and minicomputers are programmed in machine language.

To illustrate the concepts of programming in this unit, we will use machine language programming. Other more sophisticated methods of programming will be discussed later.

Programming a digital computer is basically a seven step process. These seven steps are: (1) define the problem; (2) develop a workable solution; (3) flow chart the problem; (4) code the program; (5) enter the program into the computer; (6) debug the program; (7) run the program. We will discuss each of these steps in detail.

The first and perhaps the most important step in programming a digital computer is defining the problem to be solved. The success of the program is directly related to how well you define the operation to be performed. There is no set standard for the problem defining procedure, and you can use any suitable method. The definition can be a written statement of the function to be carried out, or it may take the form of a mathematical equation. In some cases, the problem may be more easily defined by graphical means. For control applications, the problem may be expressed with a truth table. The form in which you place the definition is strictly a function of the application.

Once the problem is analyzed and defined, you can begin thinking of how the computer may solve the problem. Remember that a computer program is a step-by-step sequence of instructions that will lead to the correct results. You should think in terms of solving your problem in some step-by-step sequential manner. What you will be doing in this phase of the programming procedure is developing a algorithm. An **algorithm** is a method or procedure for solving a problem.

An important point to remember is that there is usually more than one way to solve a given problem. In other words, there is more than one algorithm suitable for achieving the goal that you have set. Much of the job of programming is in determining the alternatives and weighing them to select the best suitable approach. The simplest and most direct algorithms are usually the best.
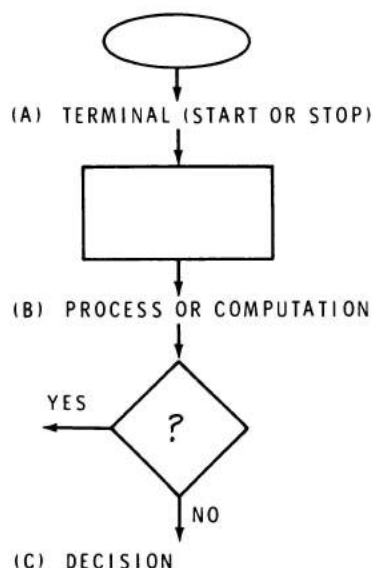
Figure 6-19
Basic flowcharting symbols.

The next step is to flowchart the problem. **A flow chart** is a graphical description of the problem solution. Various symbols are used to designate key steps in the solution of the problem. Figure 6-19 shows the basic flowcharting symbols. An oval defines the starting and finishing points. A rectangular box defines each individual computational step leading to the solution. Each rectangle contains some basic operation or calculation that is to take place. The diamond shaped symbol represents a decision point. It is often necessary to observe the intermediate results in a problem solution and make a decision regarding the next step to be taken. There are generally two exits to the diamond-shaped, decision-making symbol. These represent a yes or no type of decision.

Figure 6-20 shows a simple flow chart for the problem solved earlier. No decision was made in this program.

As you can see, the flow chart is a graphical representation of the basic method used to solve the problem. The flow chart permits you to visualize the algorithm you developed. In many cases, the flowcharting of a problem helps to determine the best approach to solve a problem since it forces you to think in a logical sequence and express the solution in a step-by-step form.

At this point in the programming procedure, your problem is quite well defined and a basic method of solving the problem has been determined. You are now ready to convert your flow chart and algorithm into a machine language program. This process is called coding. **Coding** is the procedure of listing the specific computer instruction sequentially to carry out the algorithm defined by the flow chart. This requires a familiarity with the instruction set of the computer you plan to use.

The next step in the programming procedure is to load the program into the computer memory. Once you have written the program with the computer instructions, you have all of the information necessary to enter that program into the computer memory. If you are dealing with machine language programming, you will convert the instruction words into their binary equivalents and then load them into the computer. If the program is a simple one, it can be loaded by using the binary switches on the front panel of the computer. However, for long, complex programs, this manual loading procedure is difficult and time consuming.

Figure 6-20
Flow chart for the problem 36 + 19 = 33
and print.

Most computers make it easy for the programmer to enter his program. Because of the availability and use of support programs residing within the machine, the program can usually be entered automatically. One of the most common ways of entering data into a comupter is with a keypunch machine. This is a typewriter-like machine that punches a standard computer card with the instructions to be entered. Teletype input/output machines using perforated paper tape are also commonly used for program entry. The instruction designations are typed on the machine and as they are typed, a paper tape is punched.

Once the cards or paper tape are punched, they are then fed into a tape reader or card reader and loaded into the computer memory. A special program residing within the computer memory, called a loader, causes the program to be loaded automatically.

With the program now in the computer memory, you can begin to run it. However, before you use it to obtain your final answer, it is often necessary to run through the program slowly a step at a time to look for programming errors and other problems. This process is called debugging. You test the program to see that it produces the desired results. If the program produces the correct result, it is ready to use. Often, programming mistakes are encountered and it is necessary to modify the program by changing the instruction steps. Often the entire program may be discarded and a new one written, using a different algorithm.

Once the program has been debugged, it is ready for use. With the program stored in memory, your problem can be solved. The computer is started and the desired results are produced.

Writing Programs

Before you can begin coding programs, you must be familiar with the instruction set of the computer you are using. Most digital computer instruction sets are basically alike in that they all perform certain basic functions such as addition, branching, input/output and the like. But each instruction set is different because the logic circuits unique to each computer carry out these operations in different ways. To code the program properly, you must know exactly what each instruction does. You can get this information by studying the instruction set as it is listed and explained in the computer's operation and programming manuals. By studying the instructions set, you will learn how the computer is organized and how it operates. The insight you gain from this will be valuable to you not only in coding the program but also in developing the best solution to a problem with a given machine.

Computer Instructions

A computer instruction is a binary word that is stored in the computer memory and defines a specific operation that the computer is to perform. The instruction word bits indicate the function to be performed and the data which is to be used in that operation.

There are two basic types of computer instructions: memory reference and non-memory reference. A memory reference instruction specifies the location in memory of the data word to be used in the operation indicated by the instruction. A non-memory reference instruction simply designates an operation to be performed. Non-memory reference instructions generally refer to internal housekeeping operation to be performed by the computer and manipulations on data stored in the various registers in the computer.

Figure 6-21 shows typical instruction word formats for an 8-bit microprocessor. The format shown in Figure 6-21A is a memory reference instruction. The instruction is defined by three 8-bit words which are stored in sequential memory locations. The first 8-bit word is the op code (or operations code) which is simply a binary bit pattern specifying some operation. The second and third 8-bit words specify the memory address of the data or operand to be used. The 8-bit op code defines 256 possible operations or functions. It is the op code that designates the operation that is to be performed. The 16-bit address specifies the memory location of the data to be operated upon. The size of the address generally indicates the maximum memory size of the computer. With 16 bits of address information, $2^{16} = 65,536$ words can be directly addressed. We usually say that the maximum memory size is 65K.

The word format in Figure 6-21B is the typical format for non-memory reference instructions. Only a 8-bit op code is used. In this type of instruction, an address is not needed since we do not reference a location in memory where data is stored. Instead, the bits in this field are used to specify various operations that are to take place within the CPU. For example, such an instruction might call for the resetting (clearing) of a register or the transfer of data from one register to another. Certain types of input-output instructions have this format.

Another instruction type is the immediate instruction which is widely used in microprocessors. The format for this instruction is shown in Figure 6-21C. It consists of an 8-bit op code that specifies the operation. The second 8-bit part of this instruction is the data or operand to be used in the operation called for. The immediate instruction is like the memory reference instruction in that it specifies the use of some data word. The data to be used is in the instruction word itself rather than being referenced by an address in the instruction word. Immediate instructions save memory space and shorten instruction fetch and execution times.
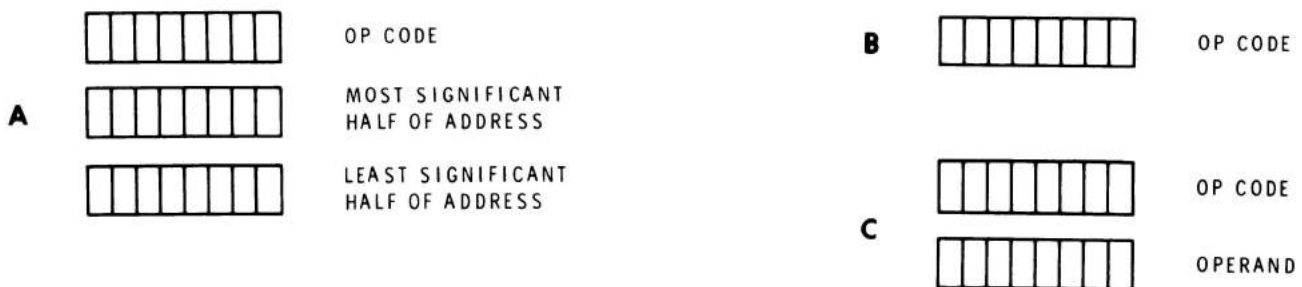
Figure 6-21

Typical computer or microprocessor instruction formats.

(A) memory reference,

(B) nonmenory reference,

(C) immediate.

Another method of classifying computer instruction is to group them according to the type of functions that they perform. These include arithmetic and logic, decision-making, data moving, and control. We will consider each of these in more detail.

An arithmetic instruction defines a specific mathematical operation that is to take place. The most commonly used arithmetic instructions are add and subtract. In larger computers, the multiply and divide functions are also included. Multiply and divide operation in smaller computers such as minicomputers and microprocessors are carried out by special sub-routines. As an example, multiplication can be performed by repeated addition. Division can be programmed by the use of repeated subtractions. Arithmetic instructions are generally of the memory reference type.

Logical instructions specify digital logic operations that are to be performed on computer data. These include the standard logic functions of AND, OR and invert(complement). Many computers include the exclusive OR function. Other logic instructions include shift right and shift left operations. The AND, OR and XOR logical instructions are usually memory reference type. The shifting and inversion instructions are of the non-memory reference type, as they generally refer to operations carried out on data stored in one of the computer's registers.

A decision-making instruction is one that permits the computer to test for a variety of results and, based upon these tests, make a decision regarding the next operation to be performed. It is the decision-making instructions that set the computer apart from the standard calculator and allow the computer to automate its operations. A decision-making instruction generally follows a sequence of other instructions that perform some arithmetic or logical operation. Once the operation is performed, the decision-making instruction tests for specific results. For example, decision-making instructions test for positive or negative numbers, zero, odd or even numbers, or equality. These tests are generally made on the data stored in various registers in the machine. If the test for a specific condition exists, the computer is usually instructed to deviate from its normal sequential execution of instructions. Jump or branch instructions are memory reference instructions that test for certain conditions and then specify a memory location where the next instruction to be executed is located. Skip instructions also change the computing sequence. These instructions test for a specific condition and then, if that condition exists, direct the computer to skip the next instruction in sequence. Skip instructions are non-memory reference types.

A data moving instruction is one that causes data words to be transferred from one location to another in the computer. It is these instructions that are used to take data from memory and load it into one of the operating registers in the computer. Other data moving instructions cause data stored in a register to be stored in a specific memory location. These are memory reference instructions. Other data moving instructions specify the transfer of data words between registers in the machine. These are non-memory reference instructions. The data moving instructions provide a flexible means of transferring data within the machine to prepare it to be processed as required by the application.

A special class of data moving instructions are the input/output instructions. I/O instructions cause data to be transferred into and out of the computer. These non-memory reference instructions often specify one of several input/output channels or a specific peripheral device. Input/output operations can be programmed to take place through the operating registers of the machine, or in some computers, directly between the memory and the peripheral unit.

A control instruction is a non-memory reference instruction that does not involve the use of data. Instead, it designates some operation that is to take place on the circuitry in the computer. Clearing a register, setting or resetting a flip-flop or halting the computer are examples of control instructions.

A Hypothetical Instruction Set

A typical but hypothetical instruction set for a minicomputer or microprocessor is shown in Table I. Only a few of the most commonly used instructions are listed so that you can become acquainted with them quickly. Real instruction sets are far more extensive. Nevertheless, the instruction set in Table I is representative. We will use it to demonstrate the writing and coding of programs.

The instruction set in Table I can apply to the hypothetical computer described earlier or a typical microprocessor. For the instructions listed here, we assume that the computer has an 8-bit word length and 65K of memory. The accumulator and memory data registers are 8 bits in length. The program counter and MAR are 16 bits in length. I/O transfers take place though the accumulator. A single instruction may occupy one, two or three consecutive memory locations depending upon its format as shown in Figure 6-21. Study the instructions in Table I so that you will be familiar with the operation each performs. Note that each instruction is designated by a three-letter mnemonic. The type of instruction is designated by the letters R (memory reference), N (non-memory reference), A (arithmetic-logic), T (data moving or transfer), D (decision) and C (control). Despite the simplicity of this instruction set, it can be used to program virtually any function.

HYPOTHETICAL COMPUTER INSTRUCTION SET

Table I

MNEMONIC	TYPE OF INSTRUCTION	OPERATION PERFORMED
LDA	R, T	Load the data stored in the specified memory location (M) into the accumulator register.
STA	R, T	Store the data in the accumulator register in the specified memory location (M).
ADD	R, A	Add the contents of the specified memory location (M) to the contents of the accumulator and store the sum in the accumulator.
SUB	R, A	Subtract the contents of the specified memory location (M) from the contents of the accumulator and store the remainder in the accumulator.
AND	R, A	Perform a logical AND on the data in the specified memory location (M) and the contents of the accumulator and store the results in the accumulator.
OR	R, A	Perform a logical OR on the data in the specified memory location (M) and the contents of the accumulator and store the results in the accumulator.
JMP	R, D	Unconditionally jump or branch to the specified memory location (M) and execute the instruction stored in that location.

JMZ	R, D	Jump to the specified memory location if the content of the accumulator is zero (reset). Execute the instruction stored in that location. If the accumulator is not zero, continue with the next instruction in normal sequence.
CLA	N, C	Clear or reset the accumulator to zero.
CMP	N, A	Complement the contents of the accumulator.
SHL	N, A	Shift the contents of the accumulator one bit position to the left.
SHR	N, A	Shift the contents of the accumulator one bit position to the right.
INP	N, T	Transfer an 8-bit parallel input word into the accumulator.
OUT	N, T	Transfer the contents of the accumulator to an external device.
HLT	N, C	Halt. Stop computing.
INC	N, C	Increment the contents of the accumulator.
DCR	N, C	Decrement the contents of the accumulator.
SKO	N, D	If the number in the accumulator is odd (LSB = 1), skip the next instruction and execute the following instruction. If the accumulator is even (LSB = 0), simply execute the next instruction in sequence.

Example Programs

The following examples illustrate the use of the instruction set in writing programs. The program description, flowchart and instruction code are given in each example. Study each program, mentally executing the instructions and imagining the outcome. The format of the instruction coding is shown below.

3 ADD (7)

The number on the left is the memory address. The mnemonic specifies the instruction. The number in parenthesis is the address of the operand called for by a memory reference instruction. This line of instruction coding says that memory location 3 contains an add instruction that tells us to add the content of location 7 to the content of the accumulator.

The program shown below is a repeat of the program given in Figures 6-5 to 6-18. The only differences are the memory location numbers of the instructions, the use of mnemonics, and the substitution of the OUT instruction for the PRINT instruction.

```
 0   LDA   (16)
 3   ADD   (17)
 6   SUB   (18)
 9   STA   (19)
12   OUT
13   JMP   (20)
16   36
17   19
18   22
19   ANSWER
20   HLT
```

The difference in memory addresses is the result of the assumption that our computer uses an 8-bit word and that memory reference instructions occupy three sequential memory locations. In the program of Figures 6-5 to 6-18, we assumed that one memory address contained one instruction. In the program above, the LDA (16) occupies memory locations 0, 1 and 2. The op code is in 0, the most significant part of the address (0000 0000) is in location 1, and the least significant part of the address (0001 0000) is in location 2. The ADD, SUB, STA and JMP memory reference instructions each occupy three sequential locations. The OUT and HLT instructions do not reference memory so they occupy only a single location.

The program below illustrates the use of the logical instructions.

This program is designed to implement the logical NOR function. Since the computer instruction set contains only the AND and OR instructions, it is necessary to complement the OR function. Figure 6-22 shows the flow chart, and the program is shown below. The algorithm is $\overline{A + B}$.
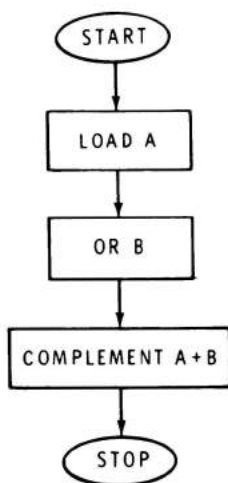
Figure 6-22
Flow chart for $\overline{A + B}$.

0	LDA	(11)
3	OR	(12)
6	CMP	
7	STA	(13)
10	HLT	
11	A	
12	B	
13	ANSWER	

The next program illustrates several important concepts. First, it shows how the computer makes decisions. Second, it demonstrates the use of a program loop. A **loop** is a sequence of instructions that is automatically repeated. The sequence is executed once and a jump instruction causes the program to branch back (loop) to the beginning of the sequence and repeat it again.

The program below is designed to enter two 4-bit BCD numbers and store them in a single 8-bit memory location. The desired memory format is shown in Figure 6-23. The BCD digits are entered, one at a time, into the
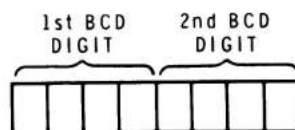
Figure 6-23
Memory format for two BCD digits.
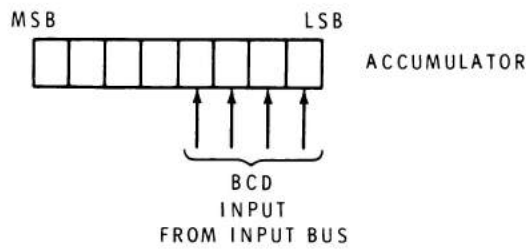
Figure 6-24
Loading the accumulator from the data
bus.

four least significant bit positions of the accumulator as shown in Figure
6-24. The first digit entered must be moved to the four most significant bit
positions of the accumulator before the second digit can be entered. This
is accomplished with a series of shift left instructions.

The flow chart in Figure 6-25 shows how this is accomplished. The
program is given below.

0	CLA	
1	INP	
2	SHL	
3	STA	(25)
6	LDA	(24)
7	DCR	
8	STA	(24)
11	JMZ	(20)
14	LDA	(25)
17	JMP	(2)
20	LDA	(25)
21	INP	
22	STA	(26)
23	HLT	
24	4	
25	Temorary storage	
26	Storage for two BCD digits	
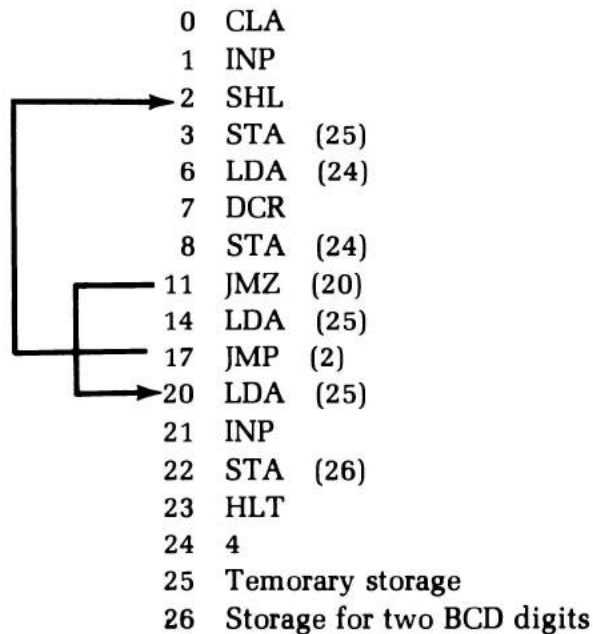
Figure 6-25
Flow chart showing how two BCD words
are stored in one memory location.
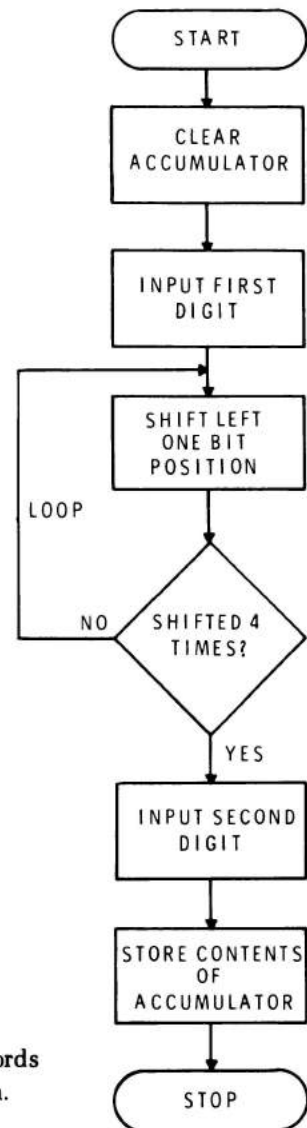
The first instruction (CLA) clears the accumulator. The second instruction (INP) loads the first BCD digit. This digit is then shifted left one bit position. We need to shift it four positions to the left. The sequence of instructions in locations 2, 3, 6, 7, 8, 11, 14 and 17 form a loop and a decision-making test to accomplish this. Stored in memory location 24 is a number that tells us how many times to shift. That number is loaded into the accumulator decremented by one and restored each time a shift occurs. We test that number with a jump on zero instruction (JMZ). When the BCD digit has been shifted four times, the number in location 24 has been reduced to zero. The JMZ instruction detects this condition and branches the program to location 20 where the data word is retrieved from temporary storage and the second BCD digit is inputted.

Now, we will consider the loop and decision-making process in more detail. After the first input digit is loaded, it is shifted left once. We then store it temporarily in location 25. This is to prevent loss of the data while we are in our decision-making loop. Next, the number of desired shifts is loaded into the accumulator. We decrement it by one, indicating that we have shifted left once. Next, we restore this number (now 3) in location 24. This number, which is still in the accumulator, is now tested with the JMZ instruction. At this time the accumulator is not zero so the program does not branch. Instead, the next instruction in sequence is executed. This is a load accumulator instruction that retrieves the data words which we temporarily stored in location 25. Then the JMP instruction is executed. This instruction returns us to location 2 to produce another shift. It is the jump instruction that creates the loop.

The loop is then repeated three more times. On the fourth pass through the loop, the number in location 24 is decremented to zero. The JMZ instruction detects this condition and causes the program to branch to location 20. We have now escaped from the loop. Next we reload the shifted data from location 25.Finally, we load the second BCD digit. Both digits are now in the accumulator so we can store them in location 26 with the STA (26) instruction. The program is now complete and the HLT instruction terminates it.

These examples show how a computer performs its work. It does it laboriously, one step at a time. The only thing that makes it practical is its high speed operation. With each instruction taking only microseconds, even long complex programs are executed quickly. To an operator, the execution appears almost instantaneous.

Self-Review Questions

21. What is machine language programming? _____

_____

_____

22. List the seven steps of programming.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____

23. A procedure for solving a problem is an _____

_____

24. A graphical description of the problem solution is known as a

_____ _____.

25. What is coding? _____

_____

_____

26. Name the two types of computer instructions and describe each one. _____

_____

_____

_____

_____

_____

_____

27. Name the computer instruction groups according to function.

1. _____
2. _____
3. _____
4. _____
5. _____

28. What is a loop? _____
    _____
    _____

29. The content of the accumulator is 10111010. The CMP instruction is executed. The new accumulator content is _____.

30. Which instruction would you use to down-count the accumulator? _____.
    (Use the Hypothetical Computer Instruction Set.)

31. The content of the accumulator is 45. A JMZ (34) instruction in location 25 is then executed. The next instruction executed is in location _____.

32. Program loops are implemented with the _____ and _____ instructions.

33. The number 0110 0101 is stored in the accumulator. The number in memory location 18 is 1111 0000. The AND (18) instruction is executed. The content of the accumulator becomes _____.

34. Study the program below. At the completion of the program, the content of the accumulator is _____.

```
 0   LDA   (13)
 3   SKO
 4   JMP   (9)
 7   LDA   (14)
 8   HLT
 9   INC
12   JMP   (3)
13   40
14   255
```
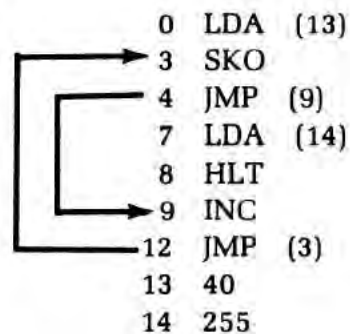
Self-Review Answers

21. Machine language programming is the process of writing programs by using the instruction set of the computer and entering the programs in binary form, one instruction at a time.

22. The seven steps of programming are:

    1. Define the program.
    2. Develop a solution.
    3. Flow chart the problem.
    4. Code the program.
    5. Enter the program into the computer.
    6. Debug the program.
    7. Run the program.

23. A procedure for solving a problem is an **algorithm**.

24. A graphical description of the problem solution is known as a **flow chart**.

25. Coding is the procedure of listing the specific computer instructions sequentially to carry out the algorithm defined by the flow chart.

26. The two types of computer instructions are memory reference and nonmemory reference. A memory reference instruction specifies the location in memory of the data word to be used in the operation indicated by the instruction. A nonmemory reference instruction simply designates an operation to be performed.

27. The computer instruction groups according to function are:

    1. Arithmetic
    2. Logic
    3. Decision making
    4. Data moving
    5. Control

28. A **loop** is a sequence of instructions that is automatically repeated.

29. The complemented accumulator content is 01000101.

30. To down-count the accumulator, you should use DCR (decrement).

31. The next instruction executed is in location 28.

    The JMZ(34) instruction tests for a zero accumulator. The accumulator content is 45; therefore, the program does not branch to location 34. Instead, it executes the next instruction in sequence, which begins in location 28. Remember that the JMZ(34) instruction and its reference address occupies locations 25, 26 and 27.

32. Program loops are implemented with the JMP and JMZ instructions.

33. The content of the accumulator becomes 0110 0000. Consider the two words to be ANDed as inputs to an AND gate as they would appear in a truth table. Then AND each corresponding pair of bits to get the result.

34. The content of the accumulator is 255.

    This program uses the "skip on odd" accumulator instruction (SKO) to test the content of the accumulator. It first loads the content of location 13 into the accumulator. This is the number 40. The SKO then tests for an odd condition by monitoring the LSB. Since 40 is even, the program does not skip. The next instruction in sequence is executed. This is a JMP (9) instruction which causes the program to branch to location 9. Here the INC instruction is executed. The accumulator then becomes 41. The next instruction JMP (3) loops the program back to location 3, where the SKO instruction again tests the accumulator. This time, the content is odd. The instruction in location 4 is now skipped and the next in sequence is executed. This is an LDA (14) which loads 255 into the accumulator. The program then halts.

SOFTWARE

The steps that we have just described make up the procedure typically used in developing application programs for the digital computer. The program may be solving a mathematical equation, sorting and editing a large volume of data, or providing some type of automatic control to an external machine.

These application programs fall into a larger category of computer programs called software. **Software** is a general term used to describe all of the programs used in a digital computer. Besides the specific applications programs, there are many special programs supplied by the computer manufacturer which are used to simplify and speed up the use of the computer. These support programs eliminate much of the drudgery from programming and using a computer. It was determined early in the development of the digital computer that the computer itself with special internal control programs could assume much of the responsibility for the detailed translation of a problem into the binary language of the computer.

Subroutines

Many digital computer manufacturers supply what are called software libraries of subroutines and utility programs. A **subroutine** is a short machine language program that solves a specific problem or carries out some often-used operation. For example, typical subroutines in many minicomputers and microprocessors are the multiply and divide programs. Instead of using the multiply subroutine each time it is required in a problem, the subroutine is stored in the computer memory only once. This saves a substantial amount of memory space. Each time the multiply subroutine is required, a jump instruction in the program causes the program to branch to the multiply subroutine. Once the multiplication has been performed, the computer jumps back to the normal program sequence.

There are many different types of commonly used subroutines. Multiplication and division are two of the most commonly used. Other subroutines include binary to BCD and BCD to binary code conversions. To communicate with external peripheral devices which use the decimal number system and the alphabet, a code such as ASCII is used. Data is entered into the computer in the ASCII code. In order for the computer to process this data, it must first be converted into pure binary numbers. Solutions to computer programs are in the pure binary form. A subroutine is used to convert the binary numbers into the ASCII format and then they are sent to an external peripheral device such as the printer.

Utility Programs

Utility programs refer to the short routines used to run the computer. Input/output programs for specific types of peripheral devices fall into this category. A loader is another utility program. This is a short sequence of instructions that allows data to be loaded into the computer. In order to operate, a computer must be programmed. But to load a program into the computer automatically requires that the loader program exist in the memory to begin with. Such loader programs are often entered manually from the computer front panel. The short loader program in memory then permits longer, more complicated programs to be loaded automatically.

Assembler

The most sophisticated software supplied with most computers are large complex conversion programs called assemblers and compilers. These programs allow the computer to be programmed in a simpler language. Machine language programming is completely impractical for many modern applications. To simplify computer programming and eliminate the need for a knowledge of binary numbers and the computer architecture, computer manufacturers have developed easier methods of programming the computer. These methods involve higher level languages, which are special systems for speeding up the programming process. The higher level language permits someone with no computer expertise whatsoever to use the computer. The higher level language permits the programmer to express his problem as a mathematical equation or in some cases as an English language statement. These equations and statements are then fed to the computer, which then automatically converts them into the binary instructions used to solve the problem.

The simplest form of higher level programming language is called assembly language. This is a method of programming the computer an instruction at a time as you do in machine language programming. However, instead of binary designations for each instruction, short multiletter names called mnemonics are given to each of the computer instructions. These are then written sequentially to form the program. Mnemonics are also given to memory addresses to avoid the use of specific memory locations.

Once the computer program is written in the assembly language, it is then entered into the machine along with an assembler program. The assembler program resides in the computer memory and is used to convert the mnemonics into the binary instruction words that the computer can interpret. As you can see, the assembler is a program that eliminates the necessity of dealing with binary numbers in the digital computer. However, since the machine is still programmed an instruction at a time, it provides wide flexibility in solving a given problem.

Compiler

The compiler, like the assembler, is a complex conversion program that resides in the computer memory. Its purpose is to convert a simplified statement of the program into the binary machine code that the computer can understand. The difference between the compiler and the assembler is that the compiler is capable of recognizing even simpler problem statements.

In one type of compiler programming language know as Fortran, the program can be written as an algebraic equation. The algebraic equation is then entered into the computer through a teletypewriter or via punched cards. The compiler program then analyzes the formula and proceeds to construct a binary program to solve this equation at some location in memory.

Another higher level programming language know as Cobol uses English language statements to describe the problem. These English language statements are punched into cards and then read into the computer memory. The compiler interprets them and converts them into the binary program. Unlike an assembly language program which has a one-to-one correspondence of instruction steps with machine language, a single compiler language program statement often causes many binary instructions to be generated.

There are many different types of higher level programming language used with computers. All of them have the prime function of simplifying the programming procedure. They greatly speed up and expedite the communications with the computer. They allow anyone who is capable of defining his problem to use the computer.

Cross Assemblers and Compilers

There are several special types of higher level programming language that have been developed to aid in programming microprocessors. For simple applications, microprocessors are programmed at the machine language level. However, when longer or more complex programs are required, it is desirable to use an assembler or compiler if it is available. For minicomputers and larger scale computers, compilers and assemblers that reside within the computer memory itself are available to aid in the programming process. However, most microprocessors do not have sufficient memory to accomodate such large complex programs. In addition, the microprocessor is generally to be dedicated to a specific application and, therefore, its memory will only be large enough to hold the application program required. In order to simplify the development of programs for use in a microprocessor, special programs called cross-assemblers and cross-compilers have been developed. These are special programs that reside in the memory of a larger, general purpose digital computer. The application programs are written in these higher level languages and the larger machine then converts the application program into the binary machine language required by the microprocessor. The output of the larger scale computer is generally a paper tape containing the binary program, which is later loaded into the microprocessor memory.

Some of the larger more sophisticated microprocessors have been used as the primary component in a microcomputer that can be used as a software development system for that microprocessor. A large random access memory is added to the microprocessor along with appropriate peripheral devices. Resident assembler programs have been developed for these machines. In this way, the microcomputer based on the microprocessor can be used to develop application programs that will be used later in another system employing the same microprocessor.

Self Review Questions

35. What is software? _____

    _____

    _____

36. What is a subroutine? _____

    _____

    _____

37. Input/output and loader programs are called. _____ programs.

38. The program used to convert an instruction by instruction mnemonic program into binary machine language is called an _____ .

39. What is a compiler? _____

    _____

    _____

40. What are cross assemblers and compilers? _____

    _____

    _____

Self-Review Answers

35. Software is a broad term used to describe all of the programs used in a digital computer.

36. A subroutine is a short, machine language program that solves a specific problem or carries out some often used operation.

37. Input/output and loader programs are called **utility** programs.

38. The program used to convert an instruction-by-instruction mnemonic program into binary machine language is called an **assembler**.

39. A compiler, like an assembler, converts a statement of the program into a binary machine language program. However, a compiler is more complex in that it may convert a single statement into several machine language steps. Whereas, the assembler can only convert each mnemonic instruction into a single machine language step.

40. Cross assemblers and compilers are used with large scale, general purpose computers to develop machine language programs for microprocessors.

MICROPROCESSORS

As indicated earlier, a microprocessor is the simplest and least expensive form of digital computer available. However, this section is more specific. It discusses exactly what a microprocessor is, the types that are available, and how they are used.

Types of Microprocessors

Most microprocessors are the central processing unit (CPU) of a digital computer. That is, microprocessors usually contain the arithmetic-logic and control sections of a small scale digital computer. Most of these microprocessors also contain a limited form of input-output circuitry which permits them to communicate with external equipment. To make the microprocessor a complete computer, external memory and input-output devices must be added. An external read only memory (ROM) is normally used to store the program to be executed. Some read/write, random access memory (RAM) may also be used. The external input-output circuitry generally consists of registers and control gating that buffer the flow of data into and out of the CPU.

Microprocessors come in a wide variety of forms. However, the most popular and widely used microprocessor is a MOS LSI circuit. These circuits are made with both P-channel or N-channel enhancement mode MOS devices. The entire CPU is contained on a single chip of silicon and mounted in either a 16, 24 or 40-pin dual in-line package. Such microprocessors are available with standard word lengths of 4, 8 or 16 bits. Other more sophisticated types of microprocessors are contained within two or more integrated circuit packages. When combined, they form a complete, small scale digital computer.

While most microprocessors are of the single chip MOS variety, there are numerous bipolar microprocessors available. These are inherently faster than the MOS devices but occupy more chip space and consume more power. Where high speed is required, these bipolar devices can be used. A recently developed integrated circuit technology, referred to as integrated injection logic (I^2L), combines both the speed of bipolar devices and the high density characteristics of MOS devices. These new I^2L LSI circuits offer many benefits, and their potential for microprocessor applications is great.

Microprocessors can also be constructed with standard TTL integrated circuits. Standard MSI packages can be combined to construct a small CPU. Figure 6-26 shows a computer of this type. While this kind of microprocessor takes more circuitry and consumes more power, it generally offers several advantages. First, the microprocessor can be constructed to execute a special instruction set designed specifically for the application. With a standard off-the-shelf CPU, the instruction set is fixed. Special instruction sets are often necessary for some applications and they can be readily optimized with a special TTL microprocessor design. Another advantage of an MSI TTL microprocessor is high speed. A standard MOS microprocessor may be too slow for the application. The fastest available MOS microprocessor can execute a single instruction in approximately 2 microseconds. The simpler and less sophisticated MOS microprocessors have instruction execution speeds in the 10 to 50 microsecond region. With a special TTL MSI microprocessor, execution speeds in the nanosecond region are easily obtained.

Figure 6-26
A microprocessor made with TTL MSI
and SSI integrated circuits. This machine
is more powerful than the typical LSI
microprocessor but less powerful than a
full minicomputer. (Photo courtesy Computer Automation Inc.)

In order to use a standard single chip microprocessor, some form of external memory must be used. The program to be executed by the microprocessor is generally stored in a ROM. Data is stored in RAM. Other external components needed to support a microprocessor are an external clock circuit, input-output registers, and peripheral devices.

All single chip microprocessors incorporate a data bus through which all external data transfers take place. This may be a 4 or 8-bit bi-directional bus over which all data transfers between the memory and input-output devices communicate with the CPU. A bus design of this type greatly minimizes the number of interconnections required to connect the microprocessor to the external devices. The limiting factor of such interconnections is the number of pins on the integrated circuit package. The bus organizations keeps the pin count to a minimum, but at the same time requires time sharing of the bus. Since all data transfers between the memory and CPU and between the CPU and the peripheral devices must use the same input-output lines, each operation must take place at a different time.

The input-output devices used with most microprocessors are quite different from those used with larger digital computers. Most larger computers are connected to input-output devices like CRT terminals, teletypewriters, paper tape readers and punches, card readers and line printers. On the other hand, microprocessors are interfaced to devices such as keyboards, 7-segment LED displays, thumbwheel switches, relays, analog-to-digital and digital-to-analog converters, temperature sensors, and other such components.

Applications of Microprocessors

Microprocessors are used primarily for dedicated functions. Rarely are microprocessors used to implement a general purpose digital computer. The program of a microprocessor is usually stored in the read only memory. This means that the program is fixed and dedicated to the specific application.

There are two general applications for modern LSI microprocessors. They can be used as replacements for minicomputers or as replacements for random hard-wired logic. The development of the minicomputer enabled many engineers to design digital computers into special control systems. The minicomputer was dedicated to the control application and its programmable flexibility offered many benefits. But its cost was very high. Some microprocessors have nearly as much computing power and capability as or minicomputer and can replace the minicomputer in many systems. A microprocessor has the advantage of smaller size, lower cost, and lower power consumption.

Another common use for the microprocessor is as an alternative to standard hard-wired digital logic circuits. Equipment customarily constructed with logic gates, flip-flops, counters, and other SSI and MSI circuits can often be implemented with a single microprocessor. All of the standard logic functions such as logic operations, counting and shifting can be readily carried out by the microprocessor through programming. The microprocessor will execute instructions and sort subroutines that perform the same logic functions.

Many benefits result from using the microprocessor in replacing hard-wired random logic systems. Some of these advantages are: (1) reduced development time and cost; (2) reduced manufacturing time and cost; (3) enhanced product capability; (4) improved reliability.

Development time and cost can be significantly reduced when a microprocessor is used. The design procedures used with standard logic circuits are completely eliminated. Much of the breadboarding, cut-and-try and prototype construction is completely eliminated. Design changes can be readily incorporated and new functions implemented by simply changing the program. With a microprocessor, the logic and control functions are implemented with programs. The program can be written and entered into memory and then tested. System changes are easy to make by simply rewriting the program. Unique functions can be readily added by increasing the size of the program. In many cases, the system can be made self-checking by programming special diagnostic routines.

Development time is further reduced because a single integrated circuit microprocessor usually replaces many other integrated circuits. This reduces wiring and interconnections and simplifies printed circuit board layout. Often the printed circuit board will be significantly smaller with a microprocessor system. Power consumption and cooling are also usually simplified. The benefit of reduced development time and cost, of course, is that the product can come to market or be applied sooner.

Manufacturing costs are also reduced as a result of replacing random logic with a microprocessor. Fewer integrated circuits and smaller printed circuit boards are required to construct the system. Therefore, less time and materials are required to assemble the equipment. The programmed nature of the microprocessor system also makes it easier to test and debug than an equivalent hard-wired system.

Enhanced product capability is another benefit of using the microprocessor to replace hard-wired logic. The power of a digital system implemented with a microprocessor is limited strictly by the imagination of the designer. Many unique features and capabilities can be incorporated into the design by simply adding to the program. The incremental cost for adding such features to a microprocessor system is small compared to that of a hard-wired logic system. The ROM used to store the program usually contains extra room for program additions. Therefore, it is very easy to add special features. Many of these special or unique features would be difficult to incorporate in a random hard-wired logic design because of the extra design time, the complexity, and the additional cost. When a microprocessor is used, no additional parts or significant amount of design time are required to add them. The more unique and special features that a product can incorporate the better it performs and the more competitive it will be in the marketplace.

Another benefit of using the microprocessor to implement digital systems is increased reliability. Whenever the number of integrated circuits and wiring interconnections are reduced in a system, reliability increases significantly. Most system failures result from the failure of an integrated circuit or from an interconnection. The number of integrated circuits and interconnections are greatly reduced in going from a standard hard-wired logic system to a microprocessor system. Increased reliability means fewer failures and leads to a corresponding reduction in both warranty and service costs.

The benefits of using a microprocessor are so significant that they will soon replace most random hard-wired logic designs. But the biggest present disadvantage of using a microprocessor is the designer's lack of programming knowledge. Very little circuit or logic design is required to implement a system with a microprocessor. Instead, the primary skill required is digital computer programming. Most engineers and digital

designers were not trained in this subject, and therefore, initial design attempts with microprocessors may be slow and frustrating. However, as microprocessors are more widely used and their benefits recognized, engineers and designers will learn programming and begin to implement their systems with these devices.

Where are Microprocessors Used?

There are so many applications for microprocessors that it is difficult to classify and list them. However, to give you a glimpse at the many diverse uses for these devices, consider some of the applications where they are now being used.

1. Electronic Cash Registers
2. Electronic Scales
3. Electrical Appliance Controls
4. Automotive Controls
5. Traffic Signal Controllers
6. Machine Tool Controls
7. Programmable Calculators
8. Automatic Test Equipment
9. Data Communications Terminals
10. Process Controllers
11. Electronic Games
12. Data Collection

These are only a few of the many applications presently implemented with microprocessors. Just keep in mind that the microprocessor can be used in any other application where hard-wired standard logic systems are now used. In addition, microprocessors can also be used as the CPU in a small general purpose microcomputer or minicomputer.

A typical application for a microprocessor is illustrated in Figure 6-27. Here, the microprocesser is used in an electronic scale for a grocery market. The item to be weighed is placed on the scale. A transducer and analog-to-digital converter convert the weight into a binary word that is read into the CPU under program control. A clerk enters the price per pound via the keyboard. This too is read into the CPU. Then the CPU computes the price by multiplying the weight by the price per unit of weight. Then the total price is displayed on a 7-segment LED readout and printed on a ticket. All of this takes place under the control of the dedicated program stored in the ROM. Note the single 8-bit bi-directional data bus over which all data transfers take place.
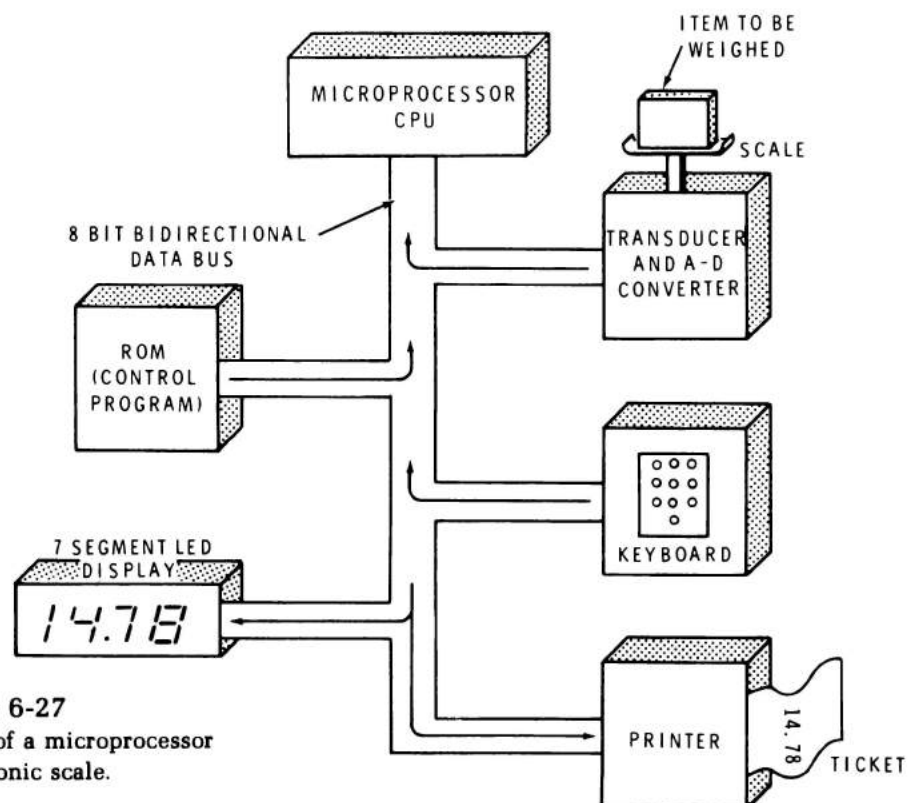
Figure 6-27
Typical application of a microprocessor
in an electronic scale.

Designing with Microprocessors

As indicated earlier, microprocessors can be used in two general ways. First, they can be used to replace minicomputers for dedicated control functions. Second, microprocessors can be used to replace standard hard-wired random logic systems. This section provides you with some guidelines to help you decide when and where a microprocessor should be used.

Microprocessors are generally much slower and less sophisticated than the typical minicomputer. But despite these limitations, microprocessors can often be used to replace minicomputers in some systems. The reason for this is that most minicomputers used in control systems are not used to their full capability. In a sense, they are a case of overkill. Many control systems used the minicomputer simply because of the ease with which the control can be changed by modifying the program. The significantly higher cost has been traded off for the convenience of system modification. In these applications, the microprocessor can usually handle the control functions as well as the minicomputer. A careful study must be made in such designs to see when a microprocessor can replace a minicomputer. There are many trade-offs to consider (speed, cost, etc.). Keep in mind that microprocessor development is in its infancy. Many technological improvements will be made over the years, causing the microprocessor to further approach the capabilities of today's minicomputer.

The microprocessor is a design alternative which should be considered in the early design stages of any digital system. The benefits of a microprocessor over standard hard-wired designs is significant in the larger, more sophisticated digital systems. As a general guideline, a microprocessor can be used beneficially if it will replace from thirty to fifty standard MSI and SSI TTL integrated circuits. If a preliminary design indicates that this many TTL integrated circuits must be used, a microprocessor should be considered. Unless the speed limitaton of the microprocessor is a factor, all of the benefits mentioned earlier will result by using the microprocessor.

Another way to equate a microprocessor design with the more conventional hard-wired logic design, is to compare the number of gates in a hard-wired design with the number of bits of memory required by a microprocessor system. It takes approximately 8 to 16 bits of memory in a microprocessor system to replace a single gate. Since most read only memories used to store the program for a microprocessor can contain as many as 16,384 bits, such a memory can replace from 1000 to 2000 gates. Depending upon the number of gates per SSI or MSI package, this can represent a replacement of hundreds of integrated circuit packages. A 16,384 (16K) bit ROM in a single 40-pin IC package, for example, can replace one hundred to four hundred 14, 16, or 24-pin SSI and MSI packages. This is a significant saving.

At this point, you may still have some doubts about the ability of a microprocessor to replace standard hard-wired logic functions. It may be difficult for you to imagine how a microprocessor can perform the functions you are so used to implementing with SSI and MSI packages. Therefore, we will consider all of the standard logic functions and illustrate how a microprocessor can perform them.

The microprocessor can readily perform all of the standard logical functions such as AND, OR, and Exclusive OR. It usually does this by executing the instructions designed for this purpose. Logical operations are generally performed on data stored in memory and in the accumulator register, with the result appearing in the accumulator. Suppose that you wanted to perform the NAND function on two 8-bit words. Using the instruction set in Table I, we could write the following program. Assume that the two words to be NANDed are stored in locations 8 and 9.

$$
\begin{array}{lll}
0 & \text{LDA} & (8) \\
3 & \text{AND} & (9) \\
6 & \text{CMP} & \\
7 & \text{HLT} & \\
\end{array}
$$

The first instruction loads the first word into the accumulator. The second instruction performs the AND function with the word in the accumulator and the word in location 9. The result appears in the accumulator. Finally, this result is complemented to form the NAND function. This simple example illustrates the procedure you use to implement any Boolean function.

Arithmetic operations are also readily performed by a microprocessor. Special adders, subtractors and other arithmetic circuits are not required because all microprocessors can perform arithmetic operations through programming. Multiplication and division operations are carried out by subroutines. Even the higher math functions such as square root, trigonometric functions, and logarithms can be computed with subroutines. Many special algorithms have been developed for solving these higher mathematical functions with digital computers. To handle very large or very small numbers or to improve the accuracy of computation, multiple precision arithmetic subroutines are also available. Number size is limited by the number of bits in the basic computer data word. However, several computer words can be used to represent a quantity as large or as small as needed. Special programs can then be written to manipulate this data just as if it were represented by a single smaller word.

An example of a programmed arithmetic operation is shown in Figure 6-28. This flow chart illustrates the procedure for multiplying two positive numbers, A and B, by repeated addition. A is added B times to produce the product. The program to implement this algorithm is given below. The numbers to be multiplied are stored in locations 31 and 32. The product or answer is stored in locations 33.

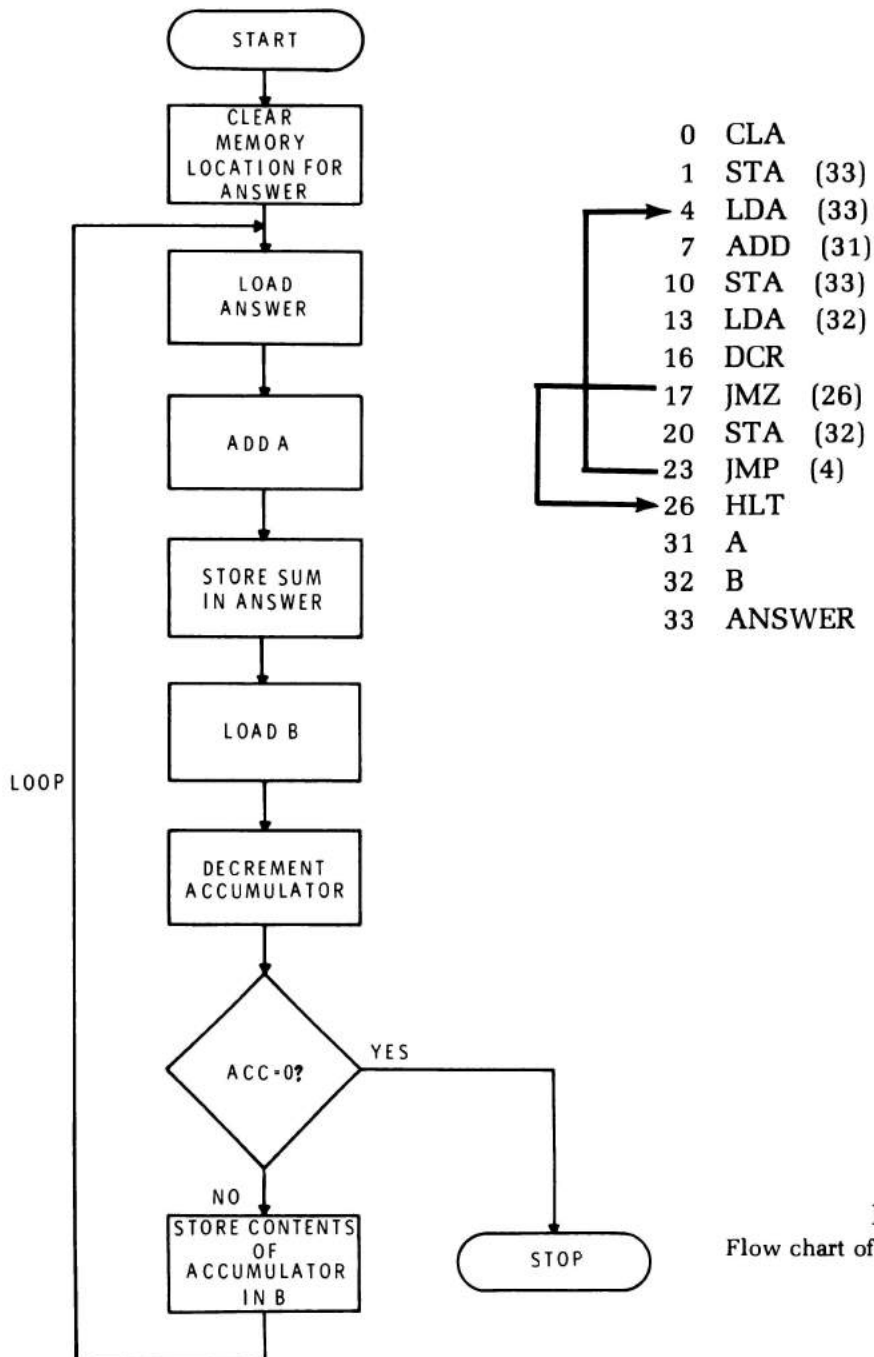|		
0	CLA	
1	STA	(33)
4	LDA	(33)
7	ADD	(31)
10	STA	(33)
13	LDA	(32)
16	DCR	
17	JMZ	(26)
20	STA	(32)
23	JMP	(4)
26	HLT	
31	A	
32	B	
33	ANSWER	

Figure 6-28
Flow chart of multiplication subroutine.

The first two instructions are used to clear the memory location where the ANSWER is to be stored. CLA resets the accumulator to zero and the STA instruction writes zeros in memory location 33. Next, the LDA instruction loads the contents of 33 (zero) into the accumulator. Then we add A to it with the ADD instruction. We then restore the partial product in location 33. We then load the content of location 32 (B) into the accumulator and subtract one from it with the decrement instruction DCR. We use a jump on zero (JMZ) instruction to see if the accumulator is zero. If it is not, we restore the accumulator content in location 32. The jump instruction creates a loop that returns us to the LDA (33) instruction. The entire sequence is then repeated. This continues until A has been added B times. Each time we add A to the answer, we subtract one from B. This permits us to keep track of how many times A has been added. When A has been added B times, the content of location 32 is again reduced by one, producing a zero result. The JMZ instruction tests for zero. The correct product is contained in location 33 at this time. The JMZ causes the program to branch to location 26, where the HALT instruction is executed to stop the program.
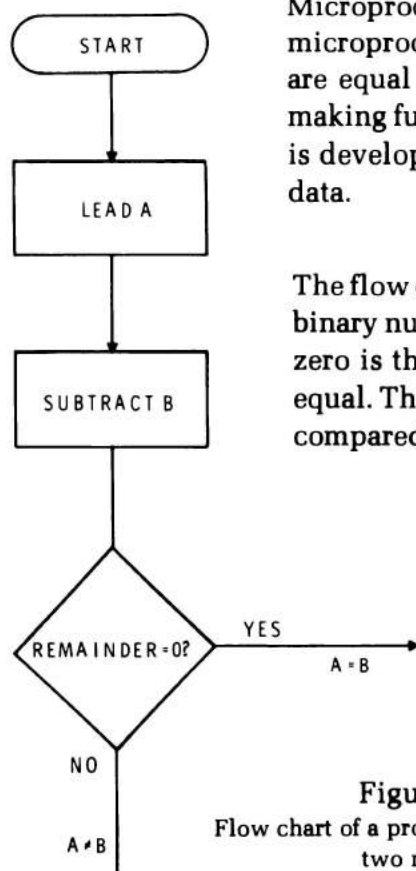
Microprocessors can also be used to make decisions. For example, the microprocessor can compare two binary numbers and determine if they are equal or if one is greater than or less than another. This decision-making function permits the microprocessor to evaluate information as it is developed and to modify its operation according to the values of the data.

The flow chart in Figure 6-29 illustrates one algorithm for comparing two binary numbers. Here, one number is subtracted from the other. A test for zero is then made. If the remainder is zero, of course, the numbers are equal. The program below implements this algorithm. The numbers to be compared are stored in locations 15 and 16.

```
 0   LDA  (15)
 3   SUB  (16)
 6   JMZ  (23)
 9   next instruction (A ≠ B)
23   next instruction (A = B)
```


Figure 6-29
Flow chart of a procedure for comparing
two numbers.

If the numbers are equal, the program branches to location 23. If the numbers are not equal, the program continues in its normal, sequential manner.

Another common logic function that is readily implemented with a microprocessor is counting. The microprocessor can count external events or a frequency standard. External events are counted by applying them to the interrupt line on the microprocessor. As each event occurs, an interrupt is generated with the microprocessor. This causes the microprocessor to jump to a subroutine that will increment the accumulator register or add one to some memory location. Up or down counters are readily implemented with the increment and decrement accumulator instructions. Decision-making techniques can be used to detect when a specific count is reached or to count quantities larger than the computer word size permits. For example, with an 8-bit data word in a microprocessor, the maximum count that the accumulator can handle is 1111 1111 or 255. To count to higher values, a program can be written to indicate each time the counter overflows.

The program below illustrates a method of detecting a count of 153. The flow chart in Figure 6-30 shows the approach.

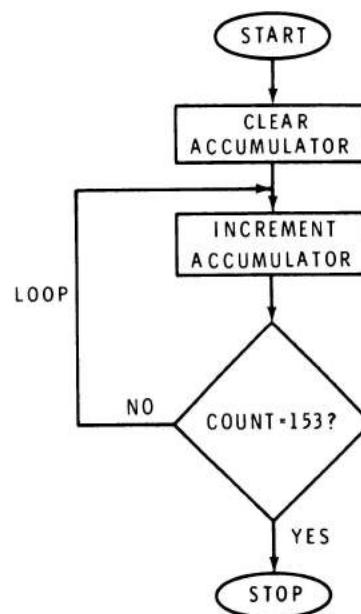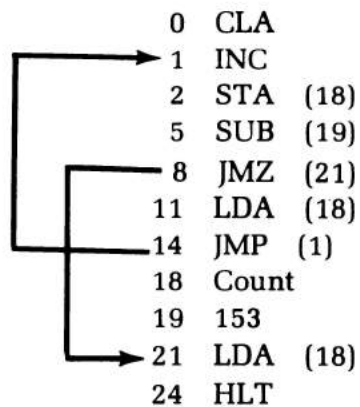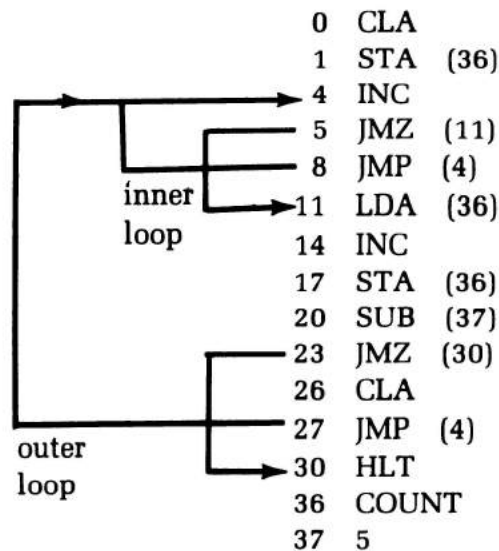|		
0	CLA	
1	INC	
2	STA	(18)
5	SUB	(19)
8	JMZ	(21)
11	LDA	(18)
14	JMP	(1)
18	Count	
19	153	
21	LDA	(18)
24	HLT	

Figure 6-30
Flow chart showing a method of detecting a count of 153.

The first instruction clears the accumulator. The accumulator is then incremented by the INC instruction, and the count is stored in location 18. The count is then compared by subtracting 153 and testing for zero. If a non-zero result occurs, the count is retrieved with the LDA instruction and the program loops back to the increment instruction. This loop continues until a count of 153 is reached. When the JMZ instruction detects the zero condition, the program branches to location 21, where the count is loaded and the program halts.

To count to numbers higher than 255, the program below can be used. See the flow chart in Figure 6-31 for an explanation of the procedure. This program counts in multiples of 256. Note that the program has two loops. The inner loop determines when a count of 256 occurs, while the outer loop determines the number of times that the inner loop occurs. The total count then is the product of the number of times the inner loop occurs and the count in location 37, in this case 5. The program halts on a count of 5 × 256 or 1280.

```
           0   CLA
           1   STA   (36)
           4   INC
           5   JMZ   (11)
           8   JMP   (4)
 inner    11   LDA   (36)
 loop     14   INC
          17   STA   (36)
          20   SUB   (37)
          23   JMZ   (30)
          26   CLA
          27   JMP   (4)
 outer    30   HLT
 loop     36   COUNT
          37   5
```
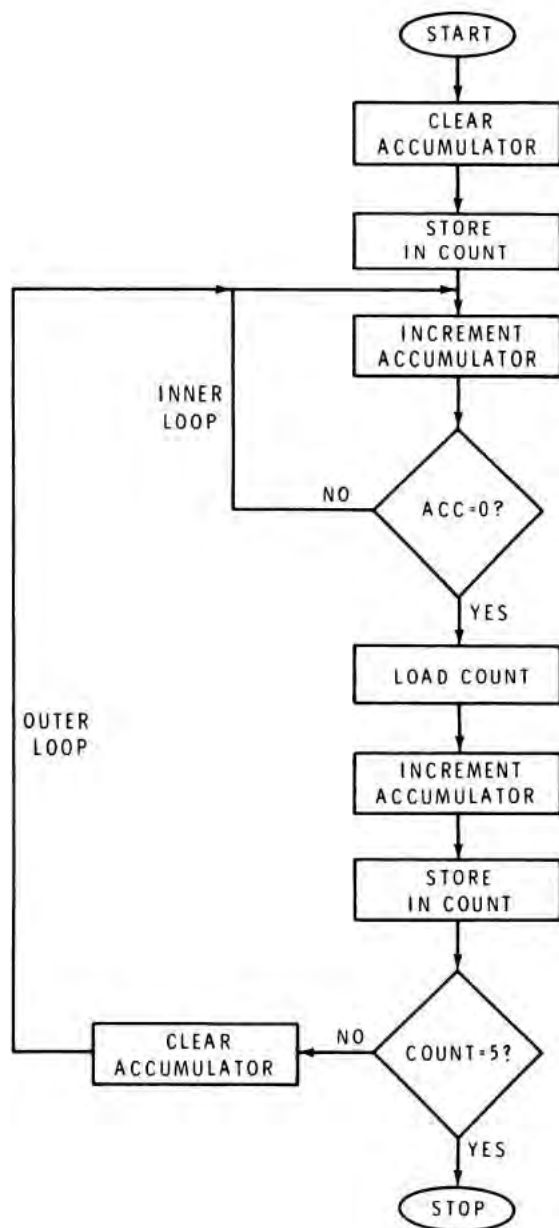
Figure 6-31
Flow chart illustrating a program to count
to 1280 (5 × 256).

The first instruction clears the accumulator. The next instruction writes zero into memory location 36 which we call COUNT. The content of location 36 tells us how many times the inner loop is repeated. These first two instructions initialize the circuitry prior to starting the count. The program then begins the count by executing the increment instruction in location 4. The JMZ (11) then tests for zero. If the accumulator is not zero, the JMP instruction is executed, creating a loop that returns the program to the INC instruction. The loop is repeated 255 times, at which time the accumulator is 1111 1111. The INC instruction is executed a 256th time and the accumulator recycles to 0000 0000. Then the JMZ (11) instruction again tests for zero. This time the program branches to location 11 where the LDA (36) instruction is performed. This loads COUNT (which is initially zero). COUNT is then incremented to indicate that a count of 256 has occurred. COUNT is then restored by the STA (36) in location 17. Next, the program tests to see if COUNT is 5. It subtracts 5 from COUNT. If the remainder is not zero, the accumulator is cleared and the program loops back to the beginning, where the inner loop is again repeated. When COUNT becomes 5, the inner loop has been repeated 5 times, indicating a count of 1280. The program then branches via the JMZ (30) instruction to a HLT.
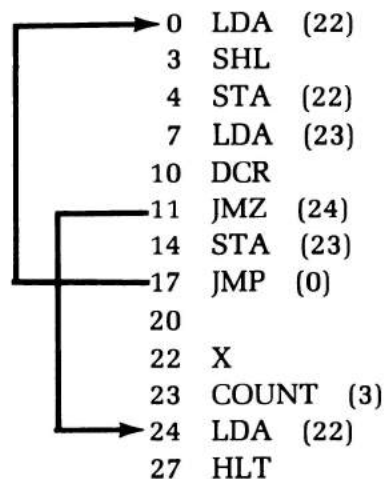
The microprocessor can also generate timed output pulses, in several ways. The simplest method is to store a series of binary numbers in sequential memory locations with the proper bit designations. These can then be read out of memory, one at a time, and sent to the output data bus. As the binary words change, the output bits change and generate any desired sequence of timing pulses. The rate of occurence of these pulses depends upon the speed of the microprocessor. Longer timed output pulses can also be generated by producing internal timing delays. This can be done by programming counting loops like the ones just illustrated. The time delays are a product of the instruction execution speed and the desired count. With a count of 153 and an instruction execution speed of 12.5 microseconds, the total delay would be $153 \times 12.5 = 1912.5$ microseconds. The binary words in memory could be outputted every 1.9125 milliseconds.

Self-Review Questions

41. The most popular and widely used microprocessor is a _____ _____ circuit.

42. Which of the following best describes a microprocessor?

    A. A general purpose digital computer.
    B. A special purpose digital computer.

43. What are the two general applications for microprocessors?

    1. _____
    2. _____

44. List the four benefits of using microprocessors to replace hard-wired logic.

    1. _____
    2. _____
    3. _____
    4. _____

45. Write a program showing how you would multiply a number (X) in location 22 by 8. Use the instruction set in Table I. Start your program in location 0. (Hint: A shift-left operation multiplies by 2.)

Self-Review Answers

41. The most popular and widely used microprocessor is a MOS LSI circuit.

42. B — A microprocessor is best described as a special purpose digital computer.

43. The two general applications for microprocessors are:

    1. To replace minicomputers.
    2. To replace hard-wired logic.

44. The four benefits of using microprocessors to replace hard-wired logic are:

    1. Reduced development time and cost.
    2. Reduced manufacturing time and cost.
    3. Enhanced product capability.
    4. Improved reliability.

45. The flow chart in Figure 6-32 shows one method of multiplying the content of location 22 (X) by 8. The program is given below.

    ```
     0   LDA    (22)
     3   SHL
     4   STA    (22)
     7   LDA    (23)
    10   DCR
    11   JMZ    (24)
    14   STA    (23)
    17   JMP    (0)
    20
    22   X
    23   COUNT  (3)
    24   LDA    (22)
    27   HLT
    ```

    Each time the number X is shifted left, it is effectively multiplied by 2. Three shifts produce multiplication by 8. A counter and decision-making loop determine when three shifts occur.
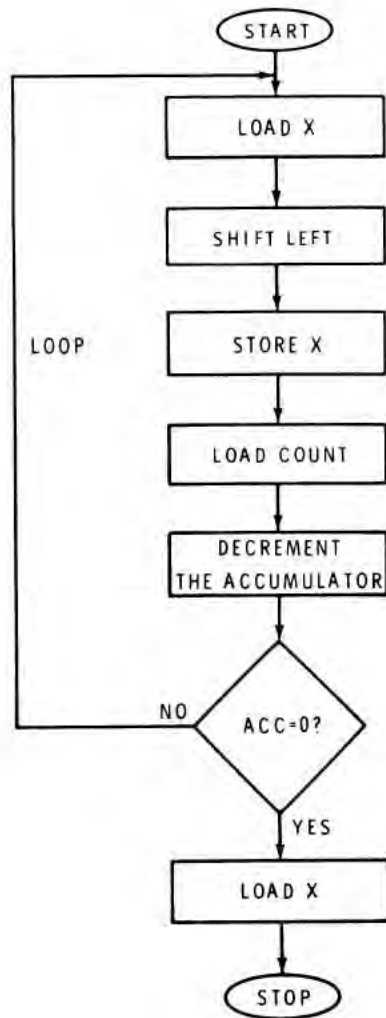
Figure 6-32
Flow Chart illustrating the procedure for
multiplying a number by 8 by shifting.