Analog Signals and Components

8

Glossary

Active Region—The region in the characteristic curve of an analog device in which the signal is amplified linearly.

Amplification—The process of increasing the size of a signal. Also called gain.

Analog signal—A signal, usually electrical, that can have any amplitude (voltage or current) value and exists at any point in time.

Anode—The element of an analog device that accepts electrons.

Base—The middle layer of a bipolar transistor, often the input.

Biasing—The addition of a dc voltage or current to a signal at the input of an analog device, which changes the signal's position on the characteristic curve.

Bipolar Transistor—An analog device made by sandwiching a layer of doped semiconductor between two layers of the opposite type: PNP or NPN.

Buffer—An analog stage that prevents loading of one analog stage by another.

Cascade—Placing one analog stage after another to combine their effects on the signal.

Cathode—The element of an analog device that emits electrons.

Characteristic Curve—A plot of the relative responses of two or three analog-device parameters, usually output with respect to input.

Clamping—A nonlinearity in amplification where the signal can be made no larger.

Collector—One of the outer layers of a bipolar transistor, often the output.

Compensation—The process of counteracting the effects of signals that are inadvertently fed back from the output to the input of an analog system. The process increases stability and prevents oscillation.

Cutoff Region—The region in the characteristic curve of an analog device in which there is no current through the device. Also called the OFF region.

Diode—A two-element vacuum tube or semiconductor with only a cathode and an anode (or plate).

Drain—The connection at one end of a field-effect-transistor channel, often the output.

Electron—A subatomic particle that has a negative charge and is the basis of electrical current.

Emitter—One of the outer layers of a bipolar transistor, often the reference.

Field-Effect Transistor (FET)—An analog device with a semiconductor channel whose width can be modified by an electric field. Also called a unipolar transistor.

Gain—see ***Amplification***.

Gain-Bandwidth Product—The interrelationship between amplification and frequency that defines the limits of the ability of a device to act as a linear amplifier. In many amplifiers, gain times bandwidth is approximately constant.

Gate—The connection at the control point of a field-effect transistor, often the input.

Grid—The vacuum-tube element that controls the electron flow from cathode to plate. Additional grids in some tubes perform other control functions to improve performance.

Hole—A positively charged "particle" that results when an electron is removed from an atom in a semiconductor crystal structure.

Integrated Circuit (IC)—A semiconductor device in which many components, such as diodes, bipolar transistors, field-effect transistors, resistors and capacitors are fabricated to make an entire circuit.

Junction FET (JFET)—A field-effect transistor that forms its electric field across a PN junction.

Linearity—The property found in nature and most analog electrical circuits that governs the processing and combination of signals by treating all signal levels the same way.

Load Line—A line drawn through a family of characteristic curves that shows the operating points of an analog device for a given output load impedance.

Loading—The condition that occurs when a cascaded analog stage modifies the operation of the previous stage.

Metal-Oxide Semiconductor (MOSFET)—A field-effect transistor that forms its electric field through an insulating oxide layer.

N-Type Impurity—A doping atom with an excess of electrons that is added to semiconductor material to give it a net negative charge.

Noise—Any unwanted signal.

Noise Figure (NF)—A measure of the noise added to a signal by an analog processing stage.

Operational Amplifier (op amp)—An integrated circuit that contains a symmetrical circuit of transistors and resistors with highly improved characteristics over other forms of analog amplifiers.

Oscillator—An unstable analog system, which causes the output signal to vary spontaneously.

P-Type Impurity—A doping atom with an excess of holes that is added to semiconductor material to give it a net positive charge.

Peak Inverse Voltage (PIV)—The highest voltage that can be tolerated by a reverse biased PN junction before current is conducted.

Pentode—A five-element vacuum tube with a cathode, a control grid, a screen grid, a suppressor grid, and a plate.

Plate—See anode, usually used with vacuum tubes.

PN Junction—The region that occurs when P-type semiconductor material is placed in contact with N-type semiconductor material.

Saturation Region—The region in the characteristic curve of an analog device in which the output signal can be made no larger. See ***Clamping***.

Semiconductor—An elemental material whose current conductance can be controlled.

Signal-To-Noise Ratio (SNR)—The ratio of the strength of the desired signal to that of the unwanted signal (noise).

Slew Rate—The maximum rate at which a signal may change levels and still be accurately amplified in a particular device.

Source—The connection at one end of the channel of a field-effect transistor, often the reference.

Superposition—The natural process of adding two or more signals together and having each signal retain its unique identity.

Tetrode—A four-element vacuum tube with a cathode, a control grid, a screen grid, and a plate.
Triode—A three-element vacuum tube with a cathode, a grid, and a plate.
Unipolar Transistor—see ***Field-Effect Transistor (FET)***.
Zener Diode—A PN-junction diode with a controlled peak inverse voltage so that it will start conducting current at a preset reverse voltage.

Introduction

This chapter, written by Greg Lapin, N9GL, treats analog signal processing in two major parts. Analog signals behave in certain well defined ways regardless of the specific hardware used to implement the processing. Signal processing involves various electronic stages to perform functions such as amplifying, filtering, modulation and demodulation. A piece of electronic equipment, such as a radio, cascades a number of these circuits. How these stages interact with each other and how they affect the signal individually and in tandem is the subject of the first part of this chapter.

Implementing analog signal processing functions involves several types of active components. An active electronic component is one that requires a power source to function, and is distinguished in this way from passive components (such as resistors, capacitors and inductors) that are described in the **DC Theory and Resistive Components** chapter and the **AC Theory and Reactive Components** chapter. The second part of this chapter describes the various technologies that implement active devices. Vacuum tubes, bipolar semiconductors, field-effect semiconductors and integrated semiconductor circuitry comprise a wide spectrum of active devices used in analog signal processing. Several different devices can perform the same function. The second part of the chapter describes the physical basis of each device. Understanding the specific characteristics of each device allows you to make educated decisions about which device would be best for a particular purpose when designing analog circuitry, or understanding why an existing circuit was designed in a particular way.

Analog Signal Processing

LINEARITY

The term, *analog signal*, refers to the continuously variable voltage of which all radio and audio signals are made. Some signals are man-made and others occur naturally. In nature, analog signals behave according to laws that make radio communication possible. These same laws can be put to use in electronic instruments to allow us to manipulate signals in a variety of ways.

The premier properties of signals in nature are *superposition* and *scaling*. Superposition is the property by which signals combine. If two signals are placed together, whether in a circuit, in a piece of wire, or even in air, they become one combined signal that is the sum of the individual signals. This is to say that at any one point in time, the voltage of the combined signal is the sum of the voltages of the two original signals at the same time. In a linear system any number of signals will add in this way to give a single combined signal.

One of the more important features of superposition, for the purposes of signal processing, is that signals that have been combined can be separated into their original components. This is what allows signals that have been contaminated with noise to be separated from the noise, for example.

Amplification and attenuation scale signals to be larger and smaller, respectively. The operation of scaling is the same as multiplying the signal at each point in time by a constant value; if the constant is greater than one then the signal is amplified, if less than one then the signal is attenuated.

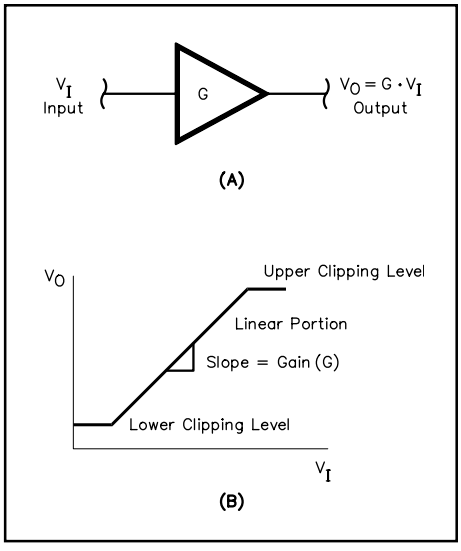

Fig 8.1 — Generic amplifier. (A) Symbol. For the linear amplifier, gain is the constant value, G, and the output voltage is equal to the input voltage times G; (B) Transfer function, input voltage along the x-axis is converted to the output voltage along the y-axis. The linear portion of the response is where the plot is diagonal; its slope is equal to the gain, G. Above and below this range are the clipping limits, where the response is not linear and the output signal is clipped.

Linear Operations

Any operation that modifies a signal and obeys the rules of superposition and scaling is a *linear operation*. The most basic linear operation occurs in an amplifier, a circuit that increases the amplitude of a signal. Schematically, a generic amplifier is signified by a triangular symbol, its input along the left face and its output at the point on the right (see **Fig 8.1**). The linear amplifier multiplies every value of a signal by a constant value. Amplifier gain is often expressed as a multiplication factor (x 5, for example).

$$\text{Gain} = \frac{V_o}{V_i} \qquad (1)$$

where V_o is the output voltage from an amplifier when an input voltage, V_i, is applied.

Ideal linear amplifiers have the same gain for all parts of a signal. Thus, a gain of 10 changes 10 V to 100 V, 1 V to 10 V and –1 V to –10 V. Amplifiers are limited by their dynamic range and frequency response, however. An amplifier can only produce output levels that are within the range of its power supply. The power-supply voltages are also called the *rails* of an amplifier. As the amplified output approaches one of the rails, the output will not go beyond a given voltage that is near the rail. The output is limited at the *clipping level* of an amplifier. When an amplifier tries to amplify a signal to be larger than this value, the output remains at this level; this is called output *clipping*. Clipping is a nonlinear effect; an amplifier is considered linear only between its clipping levels. See Fig 8.1.

Another limitation of an amplifier is its frequency response. Signals within a range of frequencies are amplified consistently but outside that range the amplification changes. At higher frequencies an amplifier acts as a low-pass filter, decreasing amplification with increasing frequency. For lower frequencies, amplifiers are of two kinds: dc and ac coupled. A dc coupled amplifier equally amplifies signals with frequencies down to dc. An ac coupled amplifier acts as a high-pass filter, decreasing amplification as the frequency decreases toward dc.

The combination of gain and frequency limitations is often expressed as a *gain-bandwidth product*. At high gains many amplifiers work properly only over a small range of frequencies. In many amplifiers, gain times bandwidth is approximately constant. As gain increases, bandwidth decreases, and vice versa. Another similar descriptor is called *slew rate*. This term describes the maximum rate at which a signal can change levels and still be accurately amplified in a particular device. There is a direct correlation between the signal-level rate of change and the frequency content of that signal.

Feedback and Oscillation

The stability of an amplifier refers to its ability to provide gain to a signal without tending to oscillate. For example, an amplifier just on the verge of oscillating is not generally considered to be "stable." If the output of an amplifier is fed back to the input, the feedback can affect the amplifier stability. If the amplified output is added to the input, the output of the sum will be larger. This larger output, in turn, is also fed back. As this process continues, the amplifier output will continue to rise until the amplifier cannot go any higher (clamps). Such *positive feedback* increases the amplifier gain, and is called *regeneration*.

Most practical amplifiers have intrinsic feedback that is unavoidable. To improve the stability of an amplifier, *negative feedback* can be added to counteract any unwanted positive feedback. Negative feedback is often combined with a phase-shift *compensation* network to improve the amplifier stability.

The design of feedback networks depends on the desired result. For amplifiers, which should not oscillate, the feedback network is customized to give the desired frequency response without loss of stability. For oscillators, the feedback network is designed to create a steady oscillation at the desired frequency.

Filtering

A filter is a common linear stage in radio equipment. Filters are characterized by their ability to selectively attenuate certain frequencies (stop band) while passing or amplifying others (pass band). Passive filters are described in the **Filters and Projects** chapter. Filters can also be designed using active devices. All practical amplifiers are low-pass filters or band-pass filters, because the gain decreases as the frequency increases beyond their gain-bandwidth products.

Summing Amplifiers

In a linear system, nature does most of the work for us when it comes to adding signals; placing two signals together naturally causes them to add. When processing signals, we would like to control the summing operation so the signals do not distort. If two signals come from separate stages and they are connected, the stages may interact, causing both stages to distort their signals. Summing amplifiers generally use a resistor in series with each stage, so the resistors connect to the common input of the following stage. **Fig 8.2** illustrates the resistors connecting to a summing amplifier. Ideally, any time we wanted to combine signals (for example, combining an audio signal with a PL tone in a 2 m FM transmitter prior to modulating the RF signal) we could use a summing amplifier.

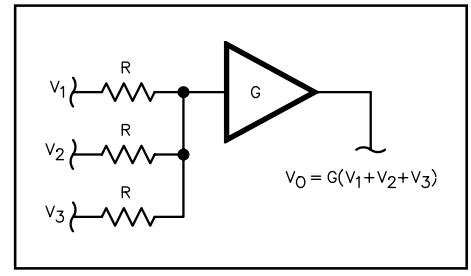

Fig 8.2 — Summing amplifier. The output voltage is equal to the sum of the input voltages times the amplifier gain, G. As long as the resistance values, R, are equal and the amplifier input impedance is much higher, the actual value of R does not affect the output signal.

Buffering

It is often necessary to isolate the stages of an analog circuit. This isolation reduces the loading, coupling and feedback between stages. An intervening stage, called a *buffer*, is often used for this purpose. A buffer is a linear circuit that is a type of amplifier. It is often necessary to change the characteristic impedance of a circuit between stages. Buffers can have high values of amplification but this is unusual. A buffer performs impedance transformations most efficiently when it has a low or unity gain. **Fig 8.3** shows common forms of buffers with low-impedance outputs: the cathode follower using a triode tube, the emitter follower using a bipolar transistor, the source follower using a field-effect transistor and the voltage follower, using an operational amplifier.

In some circuits, notably power amplifiers, the desired goal is to deliver a maximum amount of power to the output device (such as a speaker or an antenna). Matching the amplifier output impedance to the output-device impedance provides maximum power transfer. A buffer amplifier may be just the circuit for this type of application. Such amplifier circuits must be carefully designed to avoid distortion.

Amplitude Modulation/Demodulation

Voice signals are transmitted over the air by amplitude modulating them on higher frequency carrier signals (see the **Mixers** chapter). The process of amplitude modulation can be mathematically described as the multiplication (product) of the voice signal and the carrier signal. Multiplication is a linear process since amplitude modulating the sum of two audio signals produces a signal that is identical to the sum of amplitude modulating each audio signal individually. When two equal-strength SSB signals are transmitted on the same frequency, the observer hears both of the voices simultaneously. Another aspect of the linear behavior of amplitude modulation is that amplitude-modulated signals can be demodulated to be exactly in their original form. Amplitude demodulation is the converse of amplitude modulation, and is represented as a division operation.

In the linear model of amplitude modulation, the signal to be modulated (such as the audio signal in

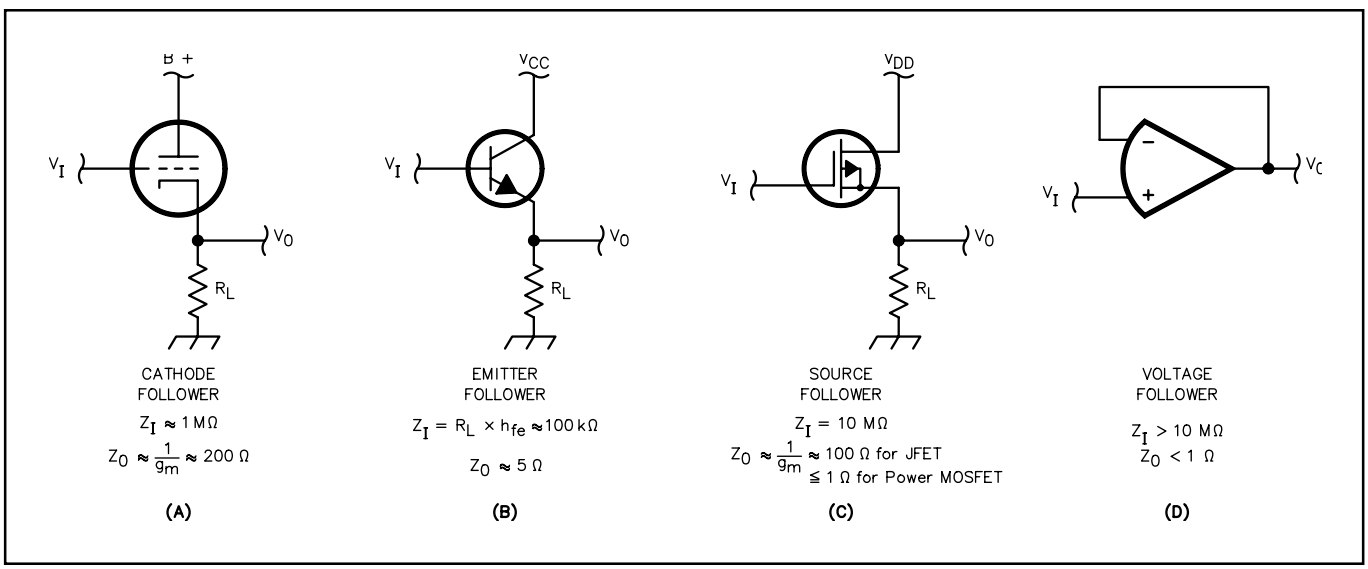

Fig 8.3 — Common buffer stages and some typical input (Z_I) and output (Z_O) impedances. (A) Cathode follower, made with triode tube; (B) Emitter follower, made with NPN bipolar transistor; (C) Source follower, made with FET; and (D) Voltage follower, made with operational amplifier. All of these buffers are terminated with a load resistance, R_L, and have an output voltage that is approximately equal to the input voltage (gain ≈ 1).

an AM transmitter) is shifted in frequency by multiplying it with the carrier. The modulated waveform is considered to be a linear function of the signal. The carrier is considered to be part of a time-varying linear system and not a second signal.

A curious trait of amplitude modulation is that it can be performed nonlinearly. Each nonlinear form of amplitude modulation generates the desired linear product term in addition to other unwanted terms that must be removed. Accurate analog multipliers and dividers are difficult and expensive to fabricate. Two common nonlinear amplitude modulating schemes are much simpler to implement but have disadvantages as well.

Power-law modulators generate many frequencies in addition to the desired ones. These unwanted frequencies, often called *intermodulation products*, steal energy from the desired *first order product*. The unwanted signals must be filtered out. The inefficiency of this process makes this type of modulator good only for low-level modulation, with additional amplification required for the modulated signal. A *square-law modulator* can be implemented with a single FET, biased in its saturation region, as the only active component.

Switching modulators are more efficient and provide high-level modulation. A single active device acts as a switch to turn the signal on and off at the carrier frequency. Both the signal and the carrier must be amplified to relatively high levels prior to this form of modulation. The modulated carrier must be filtered by a tank circuit to remove unwanted frequency components generated by the switching artifacts.

Nonlinear demodulation of an amplitude-modulated signal can be realized with a single diode. The diode rectifies the signal (a nonlinear process) and then the nonlinear products are filtered out before the desired signal is recovered.

NONLINEAR OPERATORS

All signal processing doesn't have to be linear. Any time that we treat various signal levels differently, the operation is called *nonlinear*. This is not to say that all signals must be treated the same for a circuit to be linear. High frequency signals are attenuated in a low-pass filter while low frequency signals are not, yet the filter can be linear. The distinction is that all voltages of the high-frequency signal are attenuated by the same amount, thus satisfying one of the linearity conditions. What if we do not want to treat all voltage levels the same way? This is commonly desired in analog signal processing for clipping, rectification, compression, modulation and switching.

Clipping and Rectification

Clipping is the process of limiting the range of signal voltages passing through a circuit (in other words, *clipping* those voltages outside the desired range off of the signals). There are a number of reasons why we would like to do this. Clipping generally refers to the process of limiting the positive and negative peaks of a signal. We might use this technique to avoid overdriving an amplifier, for example. Another type of clipping results in rectification. The rectifier clips off all voltages of one polarity (positive or negative) and allows only the other polarity through, thus changing ac to pulsating dc (see the **Power Supplies and Projects** chapter). Another use of clipping is when only one signal polarity is allowed to drive an amplifier input; a clipping stage precedes the amplifier to ensure this.

Logarithmic Amplification

It is sometimes desirable to amplify a signal logarithmically, which means amplifying low levels more than high levels. This type of amplification is often called *signal compression*. Speech compression is sometimes used in audio amplifiers that feed modulators. The voice signal is compressed into a small range of amplitudes, allowing more voice energy to be transmitted without overmodulation (see the **Modulation Sources** chapter).

ANALOG BUILDING BLOCKS

Many types of electronic equipment are developed by combining basic analog signal processing circuits or "building blocks." This section describes several of these building blocks and how they are combined to perform complex functions. Although not all basic electronic functions are discussed here, the characteristics of combining them can be applied generally.

An analog building block can contain any number of discrete components. Since our main concern is the effect that circuitry has on a signal, we often describe the building block by its actions rather than its specific components. For this reason, an analog building block is often referred to as a *two-port network* or a *black box*. Two basic properties of analog networks are of principal concern: the effect that the network has on an analog signal and the interaction that the network has with the circuitry surrounding it. The two network ports are the input and output connections. The signal is fed into the input port, is modified inside the network and then exits from the output port.

An analog network modifies a signal in a specific way that can be described mathematically. The output is related to the input by a *transfer function*. The mathematical operation that combines a signal with a transfer function is pictured symbolically in **Fig 8.4**. The output signal, w(t), has a value that changes with time. The output signal is created by the action of an analog transfer function, h(t), on the input signal, g(t).

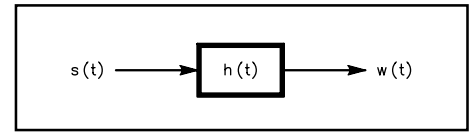

Fig 8.4 — Linear function block. The output signal, w(t) is produced by the action of the transfer function, h(t) on the input signal s(t).

While it is not necessary to understand transfer functions mathematically to work with analog circuits, it is useful to realize that they describe how a signal interacts with other signals in an electronic system. In general, the output signal of an analog system depends not only on the input signal at the same time, but also on past values of the input signal. This is a very important concept and is the basis of such essential functions as analog filtering.

Cascading Stages

If an analog circuit can be described with a transfer function, a combination of analog circuits can also be described similarly. This description of the combined circuits depends upon the relationship between the transfer functions of the parts and that of the combined circuits. In many cases this relationship allows us to predict the behavior of large and complex circuits from what we know about the parts that make them up. This aids in the design and analysis of analog circuits.

When two analog circuits are cascaded (the output signal of one stage becomes the input signal to the next stage) their transfer functions are combined. The mechanism of the combination depends on the interaction between the stages. The ideal case is when there is no interaction between stages. In other words, the action of the first stage is unchanged, regardless of whether or not the second stage follows it. Just as the signal entering the first stage is modified by the action of the first transfer function, the ideal cascading of analog circuits results in changes produced only by the individual transfer functions. For any number of stages that are cascaded, the combination of their transfer functions results in a new transfer function. The signal that enters the circuit is changed by the composite transfer function, to produce the signal that exits the cascaded circuits.

Cascaded Buffers

Buffer stages that are made with single active devices can be more effective if cascaded. Two types of such buffers are in common use. The *Darlington pair* is a cascade of two common-collector transistors as shown in **Fig 8.5.** (The various amplifier configurations will be described

later in this chapter.) The input impedance of the Darlington pair is equal to the load impedance times the current gain, h_{FE}. The current gain of the Darlington pair is the product of the current gains for the two transistors.

$$Z_I = Z_{LOAD} \times h_{FE1} \times h_{FE2} \quad (2)$$

For example, if a typical bipolar transistor has $h_{FE} = 100$ and a circuit has a $Z_{LOAD} = 15\ k\Omega$, a pair of these transistors in the Darlington-pair configuration would have:

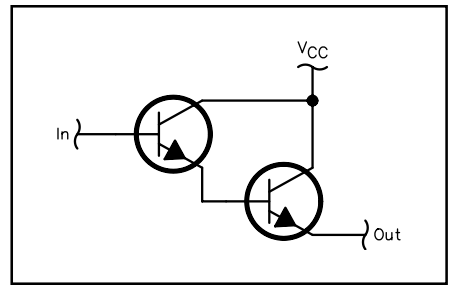

Fig 8.5 — Darlington pair made with two emitter followers. Input impedance, Z_I, is far higher than for a single transistor and output impedance, Z_O, is nearly the same as for a single transistor. DC biasing has been omitted for simplicity.

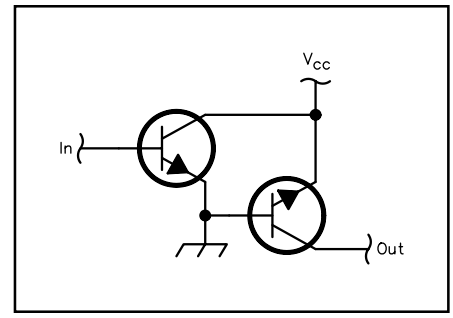

Fig 8.6 — Cascode pair made with two NPN bipolar transistors has a medium input impedance and high output impedance. DC biasing has been omitted for simplicity.

$$Z_I = 15\ k\Omega \times 100 \times 100 = 150\ M\Omega.$$

The shunt capacitance at the input of real transistors can lower the actual impedance as the frequency increases.

A common-emitter amplifier followed by a common-base amplifier is called a *cascode buffer* (see **Fig 8.6**). Cascodes are also made with FETs by following a common-source amplifier by a common-gate configuration. The input impedance and current gain of the cascode are approximately the same as those of the first stage. The output impedance is much higher than that of a single stage. Cascode amplifiers have excellent input/output isolation (very low unwanted feedback) and this can provide high gain with good stability. An example of a cascode buffer made with bipolar transistors has moderate input impedance, $Z_I = 1\ k\Omega$, high current gain, $h_{FE} = 50$ and high output impedance, $Z_O = 1\ M\Omega$. There is very little reverse internal feedback in the cascode design, making it very stable, and the amplifier design has little effect on external tuning components. Cascode circuits are often used in tuned amplifier designs for these reasons.

Interstage Loading and Impedance Matching

If the transfer function of a stage changes when it is cascaded with another stage, we say that the second stage has *loaded* the first stage. This often occurs when an appreciable amount of current passes from one stage to the next.

Every two-port network can be further defined by its input and output impedance. The input impedance is the opposition to current, as a function of frequency, that is seen when looking into the input port of the network. Likewise, the output impedance is similarly defined when looking back into a network through its output port. Interstage loading is related to the relative output impedance of a stage and the input impedance of the stage that is cascaded after it.

In some applications the goal is to transfer a maximum amount of power. In an RF amplifier, the impedance at the input of the transmission line feeding an antenna is transformed by means of a matching network to produce the resistance the amplifier needs in order to efficiently produce RF power.

In contrast, it is the goal of most analog signal processing circuitry to modify a signal rather than to deliver large amounts of energy. Thus, an impedance-matched condition may not be what is desired. Instead, current between stages can be minimized by having mismatched impedances. Ideally, if the output impedance of a network approaches zero ohms and the input impedance of the following stage is very high, very little current will pass between the stages, and interstage loading will be negligible.

Noise

Generally we are only interested in specific man-made signals. Nature allows many signals to combine, however, so the desired signal becomes combined with many other unwanted signals, both man-made and naturally occurring. The broadest definition of noise is any signal that is not the one in which we are interested. One of the goals of signal processing is to separate desired signals from noise.

One form of noise that occurs naturally and must be dealt with in low-level processing circuits is called *thermal noise*, or *Johnson noise*. Thermal noise is produced by random motion of free electrons in conductors and semiconductors. This motion increases as temperature increases, hence the name. This kind of noise is present at all frequencies and is proportional to temperature. Naturally occurring noise can be reduced either by decreasing the bandwidth or by reducing the temperature in the system. Thermal noise voltage and current vary with the circuit impedance, according to Ohm's Law. Low-noise-amplifier-design techniques are based on these relationships (see the **Amplifiers** chapter).

Analog signal processing stages are characterized in part by the noise they add to a signal. A distinction is made between enhancing existing noise (such as amplifying it) and adding new noise. The noise added by analog signal processing is commonly quantified by the *noise factor, f*. Noise factor is the ratio of the total output noise power (thermal noise plus noise added by the stage) to the input noise power when the termination is at the standard temperature of 290 K (17°C). When the noise factor is expressed in dB, we often call it *noise figure, NF*. NF is calculated as:

$$NF = 10 \log \frac{P_{NO}}{A\, P_{NTH}} \qquad (3)$$

where:

P_{NO} = total noise output power,

A = amplification gain, and

P_{NTH} = input thermal noise power.

The noise factor can also be calculated as the difference between the input and output signal-to-noise ratios (SNR), with SNR expressed in dB.

In a system of many cascaded signal processing stages, each stage affects the noise of the system. The noise factor of the first stage dominates the noise factor of the entire system. Designers try to optimize system noise factor by using a first stage with a minimum possible noise factor and maximum possible gain. A circuit that overloads is often as useless as one that generates too much noise. See the **Transceivers** chapter for more information about circuit noise.

Analog Devices

There are several different kinds of components that can be used to build circuits for analog signal processing. The same processing can be performed with vacuum tubes, bipolar semiconductors, field-effect semiconductors or integrated circuitry, each with its own advantages and disadvantages.

TERMINOLOGY

A similar terminology is used for most active electronic devices. The letter V stands for voltages and I for currents. Voltages generally have two subscripts indicating the terminals the voltage is measured between (V_{BE} is the voltage between the base and the emitter of a bipolar transistor). Currents have a single subscript indicating the terminal that the current flows into (I_P is the current into the plate of a vacuum tube). If the current flows out of the device, it is generally indicated with a negative sign. Power supply voltages have two subscripts that are the same, indicating the terminal to which the voltage is applied (V_{DD} is the power supply voltage applied to the drain of a field-effect transistor). A transfer characteristic is a ratio of an output parameter to an input parameter, such as output current divided by input current. Transfer characteristics are represented with letters, such as h, s, y or z. Resistance is designated with the letter r, and impedance with the letter Z. For example, r_{DS} is resistance between drain and source of an FET and Z_i is input impedance. In some designators, values differ for dc and ac signals. This is indicated by using capital letters in the subscripts for dc and lower-case subscripts for ac. For example, the common-emitter dc current gain for a bipolar transistor is designated as h_{FE}, and h_{fe} is the ac current gain. Qualifiers are sometimes added to the subscripts to indicate certain operating modes of the device. SS for saturation, BR for breakdown, ON and OFF are all commonly used.

The abbreviations for tubes existed before these standards were adopted so some tube-performance descriptors are different. For example, B+ is usually used for the plate bias voltage. Since integrated circuits are collections of semiconductor components, the abbreviations for the type of semiconductor used also apply to the integrated circuit. V_{CC} is a power supply voltage for an integrated circuit made with bipolar transistor technology.

Amplifier Types

Amplifier configurations are described by the *common* part of the device. The word "common" is used to describe the connection of a lead directly to a reference. The most common reference is ground, but positive and negative power sources are also valid references. The type of reference used depends on the type of device (vacuum tube, transistor [NPN or PNP], FET [P-channel or N-channel]), which lead is common and the range of signal levels. Once a common lead is chosen, the other two leads are used for signal input and output. Based on the biasing conditions, there is only one way to select these leads. Thus, there are three possible amplifier configurations for each type of three-lead device.

The operation of an amplifier is specified by its gain. A gain in this sense is defined as the change (Δ) in the output parameter divided by the corresponding change in the input parameter. If a particular device measures its input and output as currents, the gain is called a current gain. If the input and output are voltages, the amplifier is defined by its voltage gain. If the input is a voltage and the output is a current, the ratio is called the *transconductance*.

Characteristic Curves

Analog devices are described most completely with their *characteristic curves*. Almost all devices that we deal with are nonlinear over a wide range of operating parameters. We are often interested in using a device only in the region that approximates a linear response. The characteristic curve is a plot of the interrelationships between two or three variables. The vertical (y) axis parameter is the output, or result of the device being operated with an input parameter on the horizontal (x) axis. Often the output is the

result of two input values. The first input parameter is represented along the x axis and the second input parameter by several curves, each for a different value. For example, a vacuum tube characteristic curve may have the plate current along the y axis, the grid voltage along the x axis and several curves, each representing a different value of the plate bias voltage (see **Fig 8.7**).

The parameters plotted in the characteristic curve depend on how the device will be used. The common amplifier configuration defines the input and output leads, and their relationship is diagrammed by the curves. Device parameters are usually derived from the characteristic curve. To calculate a gain, the operating region of the curve is specified, usually a straight portion of the curve if linear operation is desired. Two points along that portion of the curve are selected, each defined by its location along the x and y axes. If the two points are defined by (x_1,y_1) and (x_2,y_2), the slope, m, of the curve, which can be a gain, a resistance or a conductance, is calculated as:

$$m = \frac{\Delta y}{\Delta x} = \frac{y_1 - y_2}{x_1 - x_2} \qquad (4)$$

A characteristic curve that plots device output voltage and current along the x and y axes permits the inclusion of an additional curve. The *load line* is a straight line with a slope that is equal to the load impedance. The intersections between the load line and the characteristic curves indicate the operating points for that circuit. Load lines are only applicable to output characteristic plots; they cannot be used with input or transfer (input versus output) characteristic curves.

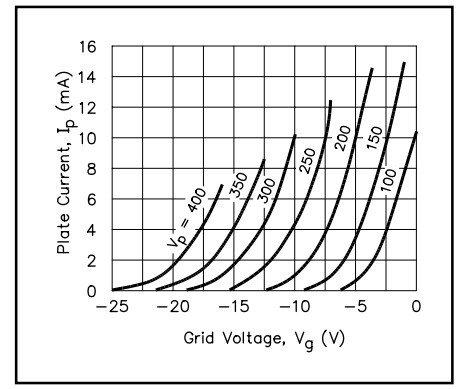

Fig 8.7 — Tube characteristic curve. Input signal is the grid voltage, V_g, along the x-axis and the output signal is the plate current, I_p, along the y-axis. Different curves are plotted for various values of plate bias voltage, V_p (also called B+).

BIASING

The operation of an analog signal processing device is greatly affected by which portion of the characteristic curve is used to do the processing. As an example, consider the vacuum tube characteristic curves in **Fig 8.8** and **Fig 8.9**.

The relationship between the input and the output of a tube amplifier is illustrated in Fig 8.8. The input signal (a sine wave in this example) is plotted in the vertical direction and below the graph. For a grid bias level of –5 V, the sine wave

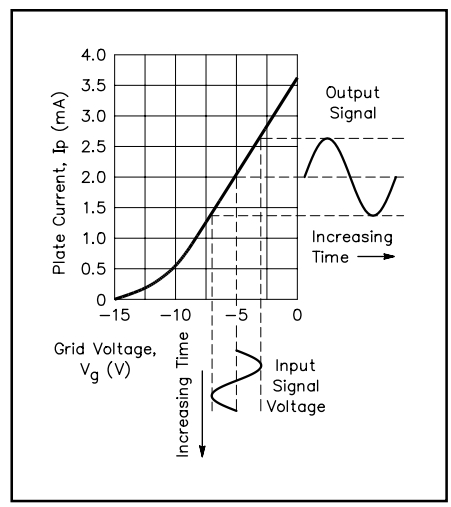

Fig 8.8 — Determination of output signal (to the right of the plot) for a given input signal (below the plot, turned on its side) with a tube characteristic curve plotted for a given plate bias. Note that the grid bias voltage, –5 V, causes the entire range of the input signal to be mapped onto the linear (diagonal straight line) portion of the characteristic curve. The output signal has the same shape as the input signal except that it is larger in amplitude.

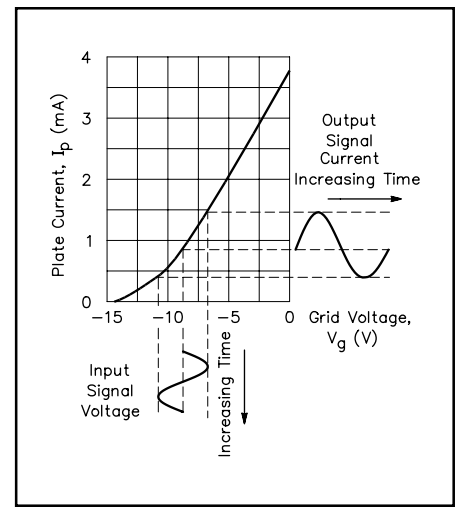

Fig 8.9 — Same characteristic curve and input signal as in Fig 8.8 except the grid bias voltage is now about –8.75 V. The input signal falls on the curved (nonlinear) portion of the plot and causes distortion in the output signal. Note how the upper portion of the output sine wave was amplified more than the lower portion.

causes the grid voltage, V_g, to deviate between –3 and –7 V. These values correspond to a range of plate currents, I_p, between 1.4 and 2.6 mA. With a plate bias of 200 V and a load resistance, R_p, of 50 kΩ, the corresponding change in plate voltage, V_p, is between 70 and 130 V. Thus, this triode amplifier configuration changes a range of 4 V at the input to 60 V at the output. Also there is a change of output-signal voltage polarity; this amplifier both amplifies the signal magnitude 15 times and shifts the phase of the signal by 180°.

In the previous example the signal was biased so that it fell on a linear (straight) portion of the characteristic curve. If a different bias voltage is selected so that the signal does not fall on a linear portion of the curve, the output signal will be a distorted version of the input signal. This is illustrated in Fig 8.9. The input signal is amplified within a curved region of the characteristic curve. The positive part of the signal is amplified more than the negative part of the signal. Proper biasing is crucial to ensure amplifier linearity.

Input biasing serves to modify the relative level (dc offset) of the input signal so that it falls on the desired portion of the characteristic curve. Devices that perform signal processing (vacuum tubes, diodes, bipolar transistors, field-effect transistors and operational amplifiers) usually require appropriate input signal biasing.

Manufacturers' Data Sheets

Manufacturer's data sheets list device characteristics, along with the specifics of the part type (polarity, semiconductor type), identification of the pins, and the typical use (such as small signal, RF, switching or power amplifier). The pin identification is important because, although common package pinouts are normally used, there are exceptions. Manufacturers may differ slightly in the values reported, but certain basic parameters are listed. Different batches of the same devices are rarely identical, so manufacturers specify the guaranteed limits for the parameters of their device. There are usually three columns of values listed in the data sheet. For each parameter, the columns may list the guaranteed minimum value, the guaranteed maximum value and/or the typical value.

Another section of the data sheet lists ABSOLUTE MAXIMUM RATINGS, beyond which device damage may result. For example, the parameters listed in the ABSOLUTE MAXIMUM RATINGS section for a solid-state device are typically voltages, continuous currents, total device power dissipation (P_D) and operating- and storage-temperature ranges.

Rather than plotting the characteristic curves for each device, the manufacturer often selects key operating parameters that describe the device operation for the configurations and parameter ranges that are most commonly used. For example, a bipolar transistor data sheet might include an OPERATING PARAMETERS section. Parameters are listed in an OFF CHARACTERISTICS subsection and an ON CHARACTERISTICS subsection that describe the conduction properties of the device for dc voltages. The SMALL-SIGNAL CHARACTERISTICS section often contains the guaranteed minimum Gain-Bandwidth Product (f_T), the guaranteed maximum output capacitance, the guaranteed maximum input capacitance and the guaranteed range of the transfer parameters applicable to a given device. Finally, the SWITCHING CHARACTERISTICS section lists absolute maximum ratings for Delay Time (t_d), Rise Time (t_r), Storage Time (t_s) and Fall Time (t_f). Other types of devices list characteristics important to operation of that specific device.

When selecting equivalent parts for replacement of specified devices, the data sheet provides the necessary information to tell if a given part will perform the functions of another. Lists of equivalencies generally only specify devices that have nearly identical parameters. There are usually a large number of additional devices that can be chosen as replacements. Knowledge of the circuit requirements adds even more to the list of possible replacements. The device parameters should be compared individually to make sure that the replacement part meets or exceeds the parameter values of the original part required by the circuit. Be aware that in some applications a far superior part may fail as a replacement, however. A transistor with too much gain could easily oscillate if there were insufficient negative feedback to ensure stability.

VACUUM TUBES

Current is generally described as the flow of electrons through a conductor, such as metal. The vacuum tube controls the flow of electrons in a vacuum, which is analogous to a faucet that adjusts the flow of a fluid. The British commonly refer to vacuum tubes as *valves*. Although the physics of the operation of vacuum tubes varies greatly from that of semiconductors, there are many similarities in the way that they behave in analog circuits.

Thermionic Theory

Metals are elements that are characterized by their large number of free electrons. Individual atoms do not hold onto all of their electrons very tightly, and it is relatively easy to dislodge them. This property makes metals good conductors of electricity. Under electrical pressure (voltage), electrons collide with metal atoms, dislodging an equal number of free electrons from the metal. These collide with adjoining metal atoms to continue the process, resulting in a flow of electrons.

It is also possible to cause the free electrons to be emitted into space if enough energy is added to them. Heat is one way of adding energy to metal atoms, and the resulting flow of electrons into space is called *thermionic emission*. It is important to remember that the metal atoms don't permanently lose electrons; the emitted electrons are replaced by others that come from an electrical connection to the heated metal. Thus, an electron that flows into the heated metal collides with and is captured by a metal atom, knocking loose a highly energized electron that is emitted into space.

In a vacuum, there are no other atoms with which the emitted electron can collide, so it follows a straight path until it collides with another atom. A *vacuum tube* has nearly all of the air evacuated from it, so the emitted electrons proceed unhindered to another piece of metal, where they continue to move as part of the electrical current.

Components of a Vacuum Tube

A basic vacuum tube contains at least two parts: a *cathode* and a *plate*. The electrons are emitted from the *cathode*. The cathode can either be heated directly by passing a large dc current through it, or it can be located adjacent to a heating element. Although ac currents can also be used to directly heat cathodes, if any of the ac voltage mixes with the signal, ac hum will be introduced into the output. If the ac heater supply voltage can be obtained from a center tapped transformer, and the center tap is connected to the signal ground, hum can be minimized. Cathodes are made of substances that have the highest emission of electrons for the lowest temperatures and voltages. Tungsten, thoriated tungsten and oxide-coated metals are commonly used.

Every vacuum tube needs a receptor for the emitted electrons. After moving though the vacuum, the electrons are absorbed by the *plate*. Since the plate receives electrons, it is also called the *anode*. Each electron has a negative charge, so a positively biased plate will attract the emitted electrons to it, and a current will result. For every electron that is accepted by the plate, another electron flows into the cathode; the plate and cathode currents must be the same. As the plate voltage is increased, there is a larger electrical field attracting electrons, causing more of them to be emitted from the cathode. This increases the current through the tube. This relationship continues until a limit is reached where further increases to the electrical field do not cause any more electrons to be emitted. This is the *saturation point* of the vacuum tube.

A vacuum tube that contains only a cathode and a plate is called a *diode tube* (di- for two components). See **Fig 8.10**. The diode tube is similar to a semiconductor diode since it allows current to pass in only one direction; it is used as a rectifier. When the plate voltage becomes negative, the electrical field that is set up repels electrons, preventing them from being emitted from the cathode.

To amplify signals, a vacuum tube must also contain a control *grid*. This name comes from its

physical construction. The grid is a mesh of wires located between the cathode and the plate. Electrons from the cathode pass between the grid wires on their way to the plate. The electrical field that is set up by the voltage on these wires affects the electron flow from cathode to plate. A negative grid voltage sets up an electrical field that repels electrons, decreasing emission from the cathode because of the higher energy needed for the electrons to escape from their atoms into the vacuum. A positive grid voltage will have the opposite effect. Since the plate voltage is always positive, however, grid voltages are usually negative. The more negative the grid, the less effective the electrical field from the plate will be at attracting electrons from the cathode.

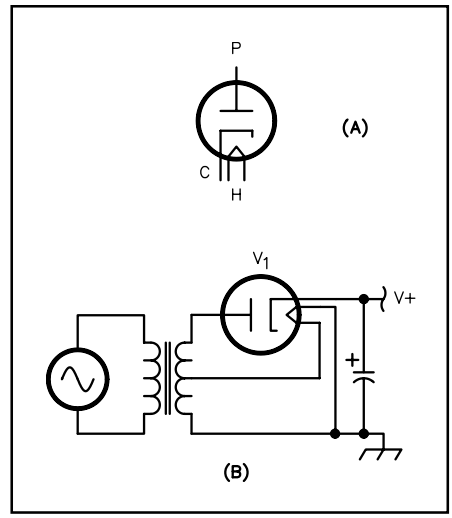

Fig 8.10 — Vacuum tube diode. (A) Schematic symbol detailing heater (H), cathode (C) and plate (P). (B) Power supply circuit using diode as a half wave rectifier.

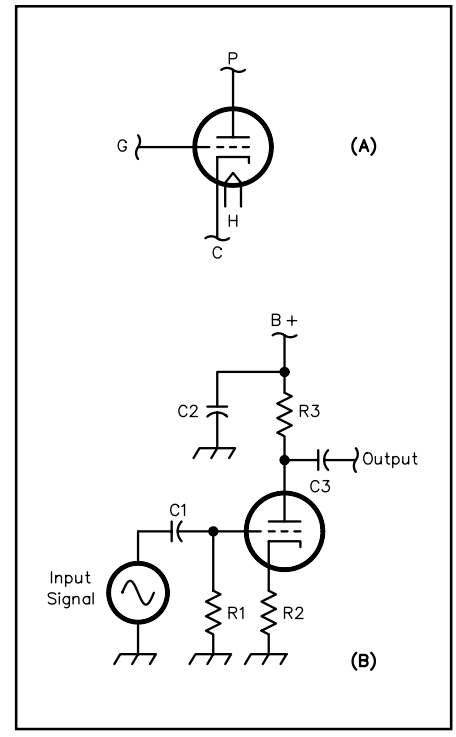

Fig 8.11 — Vacuum tube triode. (A) Schematic symbol detailing heater (H), cathode (C), grid (G) and plate (P). (B) Audio amplifier circuit using a triode. C1 and C3 are dc blocking capacitors for the input and output signals to isolate the grid and plate bias voltages. C2 is a bypass filter capacitor to decrease noise in the plate bias voltage, B+. R1 is the grid bias resistor, R2 is the cathode bias resistor and R3 is the plate bias resistor. Note that although the cathode and grid bias voltages are positive with respect to ground, they are still negative with respect to the plate.

Vacuum tubes containing a cathode, a grid and a plate are called *triode* tubes (tri- for three components). See **Fig 8.11**. They are generally used as amplifiers, particularly at frequencies in the HF range and below. Characteristic curves for triodes normally relate grid bias voltage and plate bias voltage to plate current for the triode (Fig 8.7). There are three descriptors of a tube's performance that can be derived from the characteristic curves. The *plate resistance*, r_p, describes the resistance to the flow of electrons from cathode to plate. The r_p is calculated by selecting a vertical line in the characteristic curve and dividing the change in plate-to-cathode voltage (ΔV_p) of two of the lines by the corresponding change in plate current (ΔI_p).

$$r_p = \frac{\Delta V_p}{\Delta I_p} \qquad (5)$$

The ratio of change in plate voltage (ΔV_p) to the change in grid-to-cathode voltage (ΔV_g) for a given plate current is the *amplification factor* (μ). Amplification factor is calculated by selecting a horizontal line in the characteristic curve and dividing the difference in plate voltage of two of the lines by the difference in grid voltages that corresponds to the same points.

$$\mu = \frac{\Delta V_p}{\Delta V_g} \qquad (6)$$

Triode amplification factors range from 10 to about 100.

The plate current flows to the plate bias supply, so the output from a triode amplifier is often expressed

as the voltage that is developed as this current passes through a load resistor. The value of the load resistance affects the tube amplification, as illustrated by the dynamic characteristic curves in **Fig 8.12**, so the tube μ does not fully describe its action as an amplifier. *Grid-plate transconductance* (g_m) takes into account the change of amplification due to load resistance. The slope of the lines in the characteristic curve represents g_m. (Since the various lines are nearly parallel in the linear operating region, they have about the same slope.)

$$g_m = \frac{\Delta I_p}{\Delta V_g} \qquad (7)$$

This ratio represents a conductance, which is measured in siemens. Triodes have g_m values that range from about 1000 to several thousand microsiemens, the higher values indicating greater possible amplification.

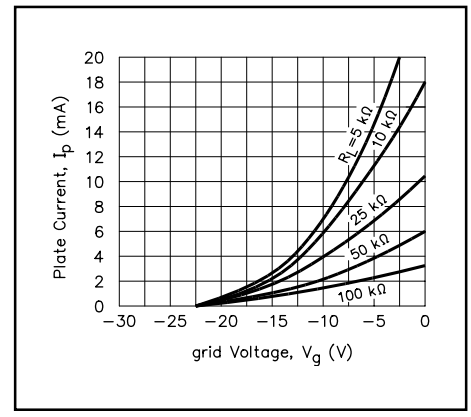

Fig 8.12 — Vacuum tube dynamic characteristic curve. This corresponds to the V_p = 300 line in Fig 8.7 with different values of load resistance. This shows how the tube will behave when cascaded to circuits with different input impedances.

The input impedance of a vacuum tube amplifier is directly related to the grid current. Grid current varies with grid voltage, increasing as the voltage becomes more positive. The normal operation uses a negative grid-bias voltage, and the input impedance can be in the megohm range for very negative grid bias values. This is limited by the desired operating point on the characteristic curve, however, as illustrated in Figs 8.8 and 8.9. The output impedance of the amplifier is a function of the plate resistance, r_p, in parallel with the output capacitance. Typical output impedance is on the order of hundreds of ohms.

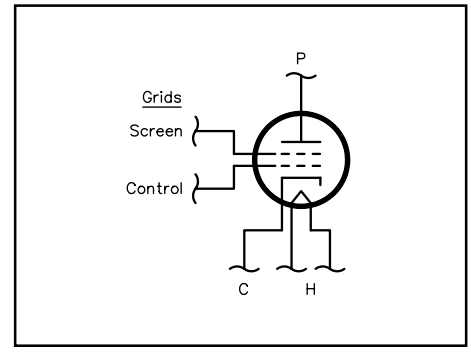

Fig 8.13 — Vacuum tube tetrode. Schematic symbol detailing heater (H), cathode (C), the two grids: control and screen and plate (P).

The physical configuration of the components within the vacuum tube appear as conductors that are separated by an insulator (in this case, the vacuum). This description is very similar to that of a capacitor. The capacitance between the cathode and grid, between the grid and plate, and between the cathode and plate can be large enough to affect the operation of the amplifier at high frequencies. These capacitances, which are usually on the order of a few picofarads, can limit the frequency response of a vacuum tube amplifier and can also provide signal feedback paths that may lead to unwanted oscillation. Neutralizing circuits are sometimes used to counteract the effects of internal capacitances and to prevent oscillations.

The grid-to-plate capacitance is the chief source of unwanted signal feedback. A special form of vacuum tube has been developed to deal with the grid-to-plate capacitance. A second grid, called a *screen grid*, is inserted between the original grid (now called a *control grid*) and the plate. The additional tube component leads to the name for this new tube—*tetrode* (tetra- for four components). See **Fig 8.13**. The screen grid reduces the capacitance between the control grid and the plate, but it also reduces the electrical field from the plate that attracts electrons from the cathode. Like the control grid, the screen grid is made of a wire mesh and electrons pass through the spaces between the wires to get to the plate. The bias of the screen grid is positive with respect to the cathode, in order to enhance the attraction of electrons from the cathode. The electrons accelerate toward the screen grid and most of them pass through the spaces and continue to accelerate until they reach the plate. The presence of the screen grid adversely affects the overall efficiency of the tube, since some of the electrons strike the grid wires. A

bypass capacitor with a low reactance at the frequency being amplified by the vacuum tube is generally connected between the screen grid and the cathode.

A special form of tetrode concentrates the electrons flowing between the cathode and the plate into a tight beam. The decreased electron-beam area increases the efficiency of the tube. *Beam tetrodes* permit higher plate currents with lower plate voltages and large power outputs with smaller grid driving power. RF power amplifiers are usually made with this type of vacuum tube.

Another unwanted effect in vacuum tubes is the emission of electrons from the plate. The electrons flowing within the tube have so much energy that they are capable of dislodging electrons from the metal atoms in the plate. These *secondary emission* electrons are repelled back to the plate by the negative bias of the grid in a triode and are of no concern. In the tetrode, the screen grid is positively biased and attracts the secondary emission electrons, causing a reverse current from the plate to the screen grid.

A third grid, called the *suppressor grid*, can be added between the screen grid and the plate. This overcomes the effects of secondary emission in tetrodes. A vacuum tube with three grids is called a *pentode* (penta- for five components). See **Fig 8.14**. The suppressor grid is negatively biased with respect to the screen grid and the plate. In some tube designs it is internally connected to the cathode. The suppressor grid repels the secondary emission electrons back to the plate.

As the number of grids is increased between the cathode and the plate, the effect of the electrical field from the positive plate voltage at the cathode is decreased. This limits the number of electrons that can be emitted from the cathode and the characteristic curves tend to flatten out as the grid bias becomes less negative. This flattening is another nonlinearity of the tube as an amplifier, since the response saturates at a given plate current and will go no higher. Tube saturation can be used advantageously in some circuits if a constant current source is desired, since the current does not change within the saturation region regardless of changes in plate voltage.

Types of Vacuum Tube Amplifiers

The descriptions of vacuum tube amplifiers up to this point have been for only one configuration, the common cathode, where the cathode is connected to the signal reference point, the grid is the input and the plate is the output. Although this is the most common configuration of the vacuum tube as an amplifier, other configurations exist. If the signal is introduced into the cathode and the grid is at a reference level (still negatively biased but with no ac component), with the output at the plate, the amplifier is called a *grounded-grid* (**Fig 8.15**). This amplifier is characterized by a very low input impedance, on the order of a few hundred ohms, and a low output impedance, that is

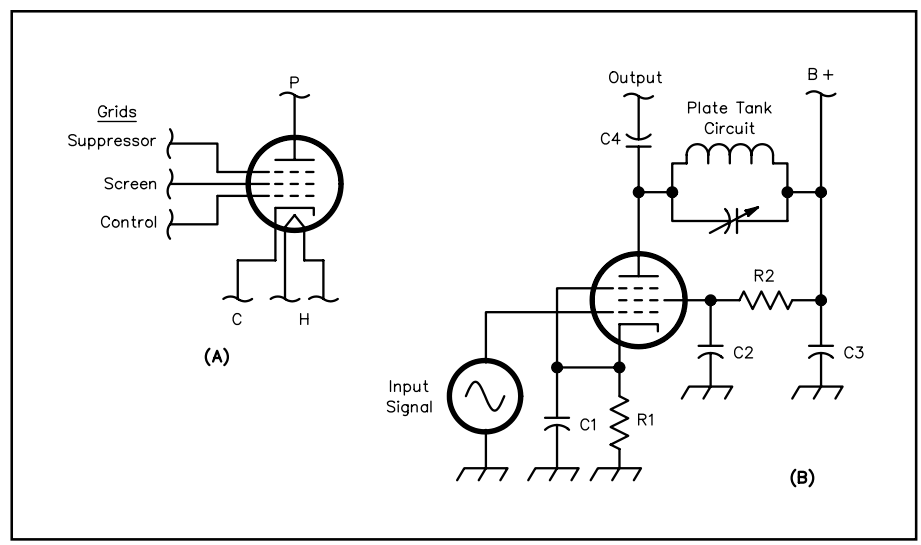

Fig 8.14 — Vacuum tube pentode. (A) Schematic symbol detailing heater (H), cathode (C), the three grids: control, screen and suppressor, and plate (P). (B) RF amplifier circuit using a pentode. C1, C2 and C3 are bypass (filter) capacitors and C4 is a dc blocking capacitor to isolate the plate bias voltage from the output signal. R1 is the cathode bias resistor and R2 is the screen voltage dropping resistor. The plate tank circuit is tuned to the desired frequency bandpass. As is common, the heater circuit is not shown.

mainly determined by the plate resistance of the tube.

The third configuration is called the *cathode follower* (**Fig 8.16**). The plate is the common element, the grid is the input and the cathode is the output. This type of amplifier is often used as a buffer stage due to its high input impedance, similar to that of the common cathode amplifier, and its very low output impedance. The output impedance (Z_o) can be calculated from the tube characteristics as:

$$Z_o = \frac{r_p}{1+\mu} \qquad (8)$$

where:

r_p = tube plate resistance

μ = tube amplification factor.

For a close approximation, we can simplify this equation as:

$$Z_o \approx \frac{r_p}{\mu} = \frac{1}{g_m}$$

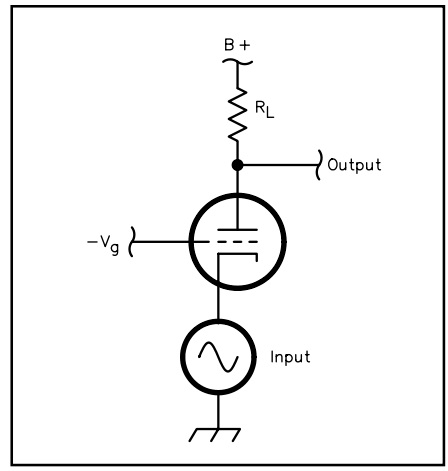

Fig 8.15 — Grounded grid amplifier schematic. The input signal is connected to the cathode, the grid is biased to the appropriate operating point by a dc bias voltage, – V_G, and the output voltage is obtained by the voltage drop through R_L that is developed by the plate current, I_P.

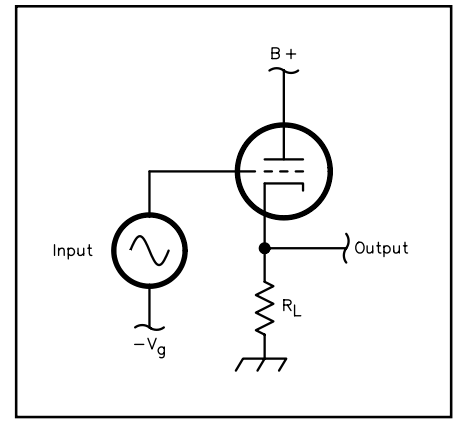

Fig 8.16 — Cathode follower schematic. The input signal is biased by –V_G and fed into the grid. The plate bias, B+ is fed directly into the plate terminal. The output is derived by the cathode current (which is equal to the plate current, I_P) dropping the voltage through the load resistor, R_L.

Other Types of Tubes

Vacuum tube identifiers do not generally indicate what type of tube the device is. The format is typically a number, one or two letters and a number (such as 6AU6 or 12AT7). The first number in the identifier indicates the heater voltage (usually either 6 or 12 V). The last number often indicates the number of elements, including the heater. Some tubes also have an additional letter following the identifier (usually A or B) that indicates a revision of the tube design that represents an improvement in its operating parameters. There are also tubes that do not follow this naming convention, many of which are power amplifiers or military-type tubes (such as 6146 and 811).

To reduce stray reactances, some tubes do not have the plate connection in the tube base, where all the other connections are located. Rather, a connection is made at the top of the tube through a metallic cap. This requires an additional connector for the plate circuitry.

Tubes may share components in a single envelope to reduce size and incidental power requirements. A very common example of this is the dual triode tube (such as 12AT7 or 12AU7) that contains a single heater circuit and two complete triode tubes in the same device. Other configurations of multiple devices contained in a single vacuum tube also exist. The 6GW8 and 6EA8 tubes each contain both a triode and a pentode. The 6BN8 contains three distinct devices, one triode and two diodes.

Most common vacuum tubes are encased in glass. It is also possible to encase them in metal or ceramic materials to attain higher tube power and smaller size. Since heat dissipation from the plate is one of the major limiting factors for vacuum tube power amplifiers, the alternate materials remove heat more efficiently. These tubes can be cooled by convection, with the casing connected to a large heat sink, or with water flowing past the tube for hydraulic cooling.

A variation of the vacuum tube that is widely used in oscilloscopes and television monitors is the *cathode ray tube (CRT)*, diagrammed in **Fig 8.17**. The CRT has a cathode and grid much like a triode tube. The plate, usually referred to as the *anode* in this device, is designed to accelerate the electrons to very high velocities, with anode voltages that can be as high as tens of thousands of volts. The anode of the CRT differs from the plates of other vacuum tubes, since it is designed as a set of plates that are parallel to the electron beam. The anode voltage accelerates the electrons but does not absorb them. The electron beam passes by the anode and continues to the face of the tube. The cathode, grid and anode are all located in the neck of the CRT and are collectively referred to as the *electron gun*.

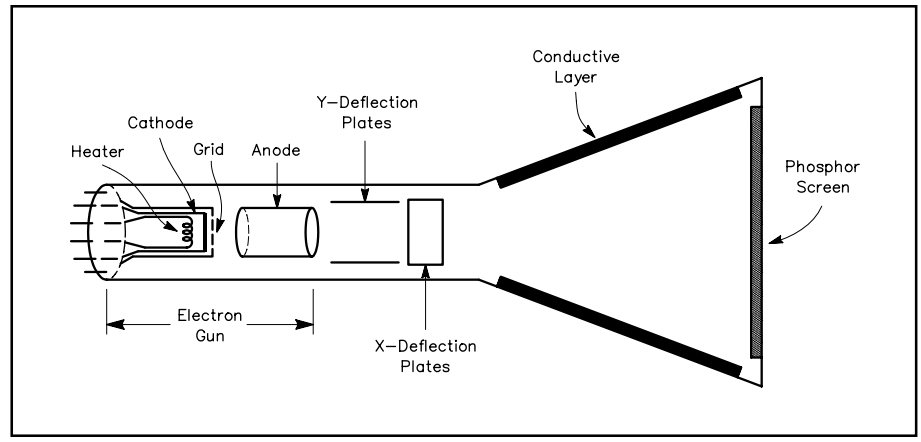

Fig 8.17 — Cross section of CRT. The electron gun generates a stream of electrons and is made up of a heater, cathode, grid and anode (plate). The electron beam passed by two pairs of deflection plates that deviate the path of the beam in the vertical (y) direction and then the horizontal (x) direction. The deflected electron beam strikes a phosphor screen and causes it to glow at that spot. Any electrons that bounce off the screen are absorbed by the conductive layer along the sides of the tube, preventing spurious luminescence.

The electron beam is deflected from its path by either magnetic deflectors that surround the yoke of the tube or by electrostatic deflection plates that are built into the tube neck just beyond the electron gun. A CRT typically has two sets of deflectors: vertical and horizontal. When a potential is applied to a set of deflectors, the passing electron beam is bent, altering its path. In an oscilloscope, the time base typically drives the horizontal deflectors and the input signal drives the vertical deflectors, although in many oscilloscopes it is possible to connect another input signal to the horizontal deflectors to obtain an X-Y, or vector, display. In televisions and some computer monitors, the deflectors are typically driven by a raster generator. The horizontal deflectors are driven by a sawtooth pattern that causes the beam to move repeatedly from left to right and then retrace quickly to the left. The vertical deflectors are driven by a slower sawtooth pattern that causes the beam to move repeatedly from top to bottom and then retrace quickly to the top. The relative timing of the two sawtooth patterns is such that the beam scans from left to right, retraces to the left and then begins the next horizontal trace just below the previous one.

Beyond the deflectors, the CRT flares out. The front face is coated with a phosphorescent material that glows when struck by the electron beam. To prevent spurious phosphorescence, a conductive layer along the sides of the tube absorbs any electrons that reflect off the glass.

Vector displays have better resolution than raster scanning. The trace lines are clearer, which is the reason oscilloscope displays use this technique. It is faster to fill the screen using raster scanning, however. This is why TVs use raster scanning.

Some CRT tubes are designed with multiple electron beams. The beams are sometimes generated by different electron guns that are placed next to each other in the neck of the tube. They can also be generated by splitting the output of a single electron gun into two or more beams. Very high quality oscilloscopes use two electron beams to trace two input channels rather than the more common method of alternating a single beam between the two inputs. Color television tubes use three electron beams for the three primary colors (red, green and blue). Each beam is focused on only one of these colored phosphors, which are interleaved on the face of the tube. A metal shadow mask keeps the colors separate as the beams scan across the tube.

A variation of the CRT is the *vidicon tube*. The vidicon is used in many video cameras and operates in a similar fashion to the CRT. The vidicon absorbs light from the surroundings, which charges the plate at the location of the light. This charge causes the cathode-to-plate current to increase when the raster scan points the electron beam at that location. The current increase is converted to a voltage that is proportional to the amount of light absorbed. This results in an electrical signal that represents the pattern of a visual image.

Standard vacuum tubes work well for frequencies up to hundreds of megahertz. At frequencies higher than this, the amount of time that it takes for the electrons to move between the cathode and the plate becomes a limiting factor. There are several special tubes designed to work at microwave frequencies. The *klystron* tube uses the principle of velocity modulation of the electrons to avoid transit time limitations. The beam of electrons travels down a metal drift tube that has interaction gaps along its sides. RF voltages are applied to the gaps and the electric fields that they generate accelerate or decelerate the passing electrons. The relative positions of the electrons shift due to their changing velocities causing the electron density of the beam to vary. The modulation of the electron density is used to perform amplification or oscillation. Klystron tubes tend to be relatively large, with lengths ranging from 10 cm to 2 m and weights ranging from as little as 150 g to over 100 kg. Unfortunately, klystrons have relatively narrow bandwidths, and are not retunable by amateurs for operation on different frequencies.

The *magnetron* tube is an efficient oscillator for microwave frequencies. Magnetrons are most commonly found in microwave ovens and high powered radar equipment. The anode of a magnetron is made up of a number of coupled resonant cavities that surround the cathode. The magnetic field causes the electrons to rotate around the cathode and the energy that they give off as they approach the anode adds to the RF electric field. The RF power is obtained from the anode through a vacuum window. Magnetrons are self oscillating with the frequency determined by the construction of their anodes; however, they can be tuned by coupling either inductance or capacitance to the resonant anode. The range of frequencies depends on how fast the tuning must be accomplished. The tube may be tuned slowly over a range of approximately 10% of the center frequency. If faster tuning is necessary, such as is required for frequency modulation, the range decreases to about 5%.

A third type of tube capable of operating in the microwave range is the *traveling wave tube*. For wide band amplifiers in the microwave range this is the tube of choice. Either permanent magnets or electromagnets are used to focus the beam of electrons that emerges from an electron gun similar to the one described for the CRT tube. The electron beam passes through a helical *slow-wave structure*, in which electrons are accelerated or decelerated, providing density modulation due to the applied RF signal, similar to that in the klystron. The modulated electron beam induces voltages in the helix that provides an amplified tube output whose gain is proportional to the length of the slow-wave structure. After the RF energy is extracted from the electron beam by the helix, the electrons are collected and recycled to the cathode. Traveling wave tubes can often be operated outside their designed frequencies by carefully optimizing the beam voltage.

PHYSICAL ELECTRONICS OF SEMICONDUCTORS

Every atom of matter consists of, among other things, an equal number of protons and electrons. These two subatomic particles must match in number to neutralize the electric charge: one positive charge for a proton and one negative charge for an electron.

Electrons orbit the nucleus, which contains the protons, at different energy levels. The binding of the electrons to the nucleus determines how an atom will behave electrically. Loosely bound electrons are easily liberated from their nuclei; atoms with this property are called *conductors*. In contrast, tightly bound electrons require considerable energy to be dislodged from their atoms; these atoms are called *insulators*. In between these two extremes is a class of elements called *semiconductors*, or partial conductors. As energy is added to a semiconductor atom, electrons are more easily freed. This property leads to many potential applications for this type of material.

In a conductor, such as a metal, the outer, or *valence*, electrons of each atom are shared with the adjacent

atoms so there are many electrons that can move about freely between atoms. The moving free electrons are the constituents of electrical current. In a good conductor, the concentration of these free electrons is very high, on the order of 10^{22} electrons / cm^3. In an insulator, nearly all the electrons are tightly held by their atoms; the concentration of free electrons is very small, on the order of 10 electrons / cm^3.

Semiconductor atoms (germanium—Ge and silicon—Si) share their valence electrons in a chemical bond that holds adjacent atoms together. The electrons are not free to leave their atom in order to move into the sphere of the adjacent atom, as in a conductor. They can be shared by the adjacent atom, however. The sharing of electrons means that the adjacent atoms are attracted to each other, forming a bond that gives the semiconductor its physical structure.

When energy is added to a semiconductor lattice, generally in the form of heat, some electrons are liberated from their bonds and move freely throughout the structure. The bond that loses an electron is then unbalanced and the space that the electron came from is referred to as a *hole*. Electrons from adjacent bonds can leave their positions and fill the holes, thus creating new holes in the adjacent bonds. Two opposite movements can be said to occur: negatively charged electrons move from bond to bond in one direction and positively charged holes move from bond to bond in the opposite direction. Both of these movements represent forms of electrical current, but this is very different from the current in a conductor. While the conductor has *free electrons* that flow regardless of the crystalline structure, the current in a semiconductor is constrained to move only along the crystalline lattice between adjacent bonds.

Crystals formed from pure semiconductor atoms (Ge or Si) are called *intrinsic* semiconductors. In these materials the number of free electrons is equal to the number of holes. Each atom has four valence electrons that form bonds with adjacent atoms. Impurities can be added to the semiconductor material to enhance the formation of electrons or holes. These are *extrinsic* semiconductors. There are two types of impurities that can be added: one kind with five valence electrons *donates* free electrons to the crystalline structure; this is called an *N-type* impurity, for the negative charge that it adds. Some examples are antimony (Sb), phosphorus (P) and arsenic (As). N-type extrinsic semiconductors have more electrons and fewer holes than intrinsic semiconductors. Impurities with three valence electrons accept free electrons from the lattice, adding holes to the overall structure. These are called P-type impurities, for the net positive charge; some examples are boron (B), gallium (Ga) and indium (In).

Intrinsic semiconductor material can be formed by combining equal amounts of N-type and P-type impurity materials. Some examples of this include gallium-arsenide (GaAs), gallium-phosphate (GaP) and indium-phosphide (InP). To make an N-type compound semiconductor, a slightly higher amount of N-type material is used in the mixture. A P-type compound semiconductor has a little more P-type material in the mixture.

The conductivity of an extrinsic semiconductor depends on the charge density (in other words, the concentration of free electrons in N-type, and holes in P-type, semiconductor material). As the energy in the semiconductor increases, the charge density also increases. This is the basis of how all semiconductor devices operate: the major difference is the way in which the energy level is increased. Variations are: The *transistor*, where conductivity is altered by injecting current into the device via a wire; the *thermistor*, where the level of heat in the device is detected by its conductivity, and the *photoconductor*, where light energy that is absorbed by the semiconductor material increases the conductivity.

The PN Semiconductor Junction

If a piece of N-type semiconductor material is placed against a piece of P-type semiconductor material, the location at which they join is called a *PN semiconductor junction*. The junction has characteristics that make it possible to develop diodes and transistors. The action of the junction is best described by a diode operating as a rectifier. Initially, when the two types of semiconductor material are placed in contact, each type of material will have only its majority carriers: P-type will have only holes and N-type will have only free electrons. The net positive charge of the P-type material attracts free electrons from across the junction and

the opposite is true in the N-type material. These attractions lead to diffusion of some of the majority carriers across the junction, which neutralize the carriers immediately on the other side. The region close to the junction is then *depleted* of carriers, and, as such, is named the *depletion region* (or the *space-charge region* or the *transition region*). The width of the depletion region is very small, on the order of 0.5 μm.

If the N-type material is placed at a more negative voltage than the P-type material, current will pass through the junction because electrons are attracted from the lower potential to the higher potential and holes are attracted in the opposite direction. When the polarity is reversed, current does not flow because the electrons that are trying to enter the N-type material are repelled, as are the holes trying to enter the P-type material. This unidirectional current is what allows a semiconductor diode to act as rectifier.

Diodes are commonly made of silicon or germanium. Although they act similarly, they have slightly different characteristics. The *junction threshold voltage*, or *junction barrier voltage,* is the forward bias voltage at which current begins to pass through the device. This voltage is different for the two kinds of diodes. In the diode response curve of **Fig 8.18**, this value corresponds to the voltage at which the positive portion of the curve begins to rise sharply from the x axis. Most silicon diodes have a junction threshold voltage of about 0.7 V, while the value for germanium diodes typically is 0.3 V. The reverse biased leakage current is much lower for silicon diodes than for germanium diodes. The forward resistance of a diode is typically very low and varies with the amount of forward current.

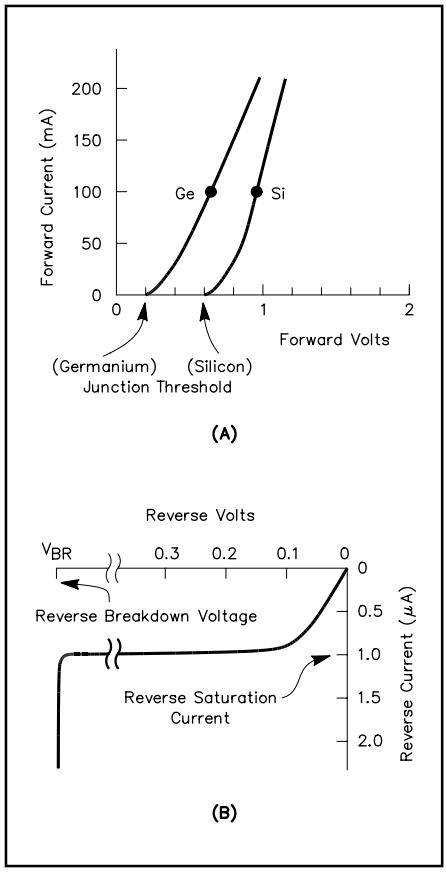

Fig 8.18 — Semiconductor diode (PN junction) response curve. (A) Forward biased (anode voltage higher than cathode) response for Germanium — Ge and Silicon — Si devices. Each curve breaks away from the x-axis at its junction threshold voltage. The slope of each curve is its forward resistance. (B) Reverse biased response. Very small reverse current increases until it reaches the reverse saturation current (I_0). The reverse current increases suddenly and drastically when the reverse voltage reaches the reverse breakdown voltage, V_{BR}.

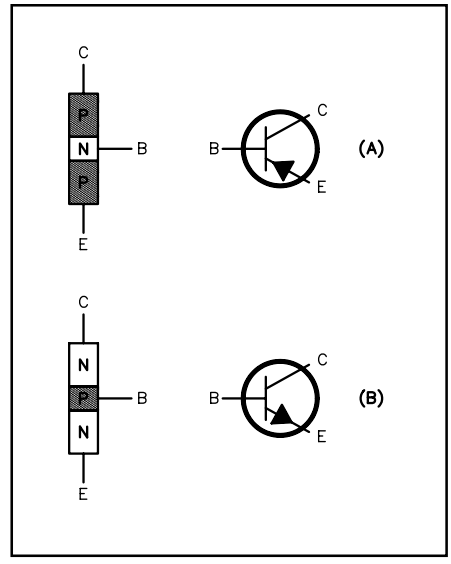

Fig 8.19 — Bipolar transistors. (A) A layer of N-type semiconductor sandwiched between two layers of P-type semiconductor makes a PNP device, the schematic symbol has three leads: collector (C), base (B) and emitter (E), with the arrow pointing in toward the base. (B) A layer of P-type semiconductor sandwiched between two layers of N-type semiconductor makes an NPN device. The schematic symbol has three leads: collector (C), base (B) and emitter (E), with the arrow pointing out away from the base.

Multiple Junctions

A bipolar transistor is formed when two PN junctions are placed next to each other. If N-type material is surrounded by P-type material, the result is a PNP transistor. Alternatively, if P-type material is in the middle of two layers of N-type material, the NPN transistor is formed (**Fig 8.19**).

Physically, we can think of the transistor as two PN junctions back-to-back, such as two diodes connected at their *anodes* (the positive terminal) for an NPN transistor or two diodes connected at their *cathodes* (the negative terminal) for a PNP transistor. The connection point is the base of the

transistor. (You can't actually *make* a transistor this way.) A transistor conducts when the base-emitter junction is forward biased and the base-collector is reverse biased. Under these conditions, the emitter region emits majority carriers into the base region, where they are minority carriers because the materials of the emitter and base regions have opposite polarity. The excess minority carriers in the base are attracted across the base-collector junction, where they are collected and are once again considered majority carriers. The flow of majority carriers from emitter to collector can be modified by the application of a bias current to the base terminal. If the bias current has the same polarity as the base material (for example holes flowing into a P-type base) the emitter-collector current increases. A transistor allows a small base current to control a much larger collector current.

As in a semiconductor diode, the forward biased base-emitter junction has a threshold voltage (V_{BE}) that must be exceeded before the emitter current increases.

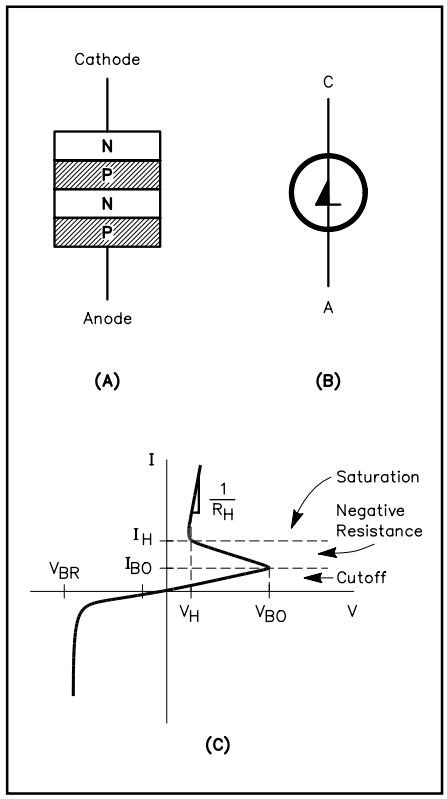

Fig 8.20 — PNPN diode. (A) Alternating layers of P-type and N-type semiconductor. (B) Schematic symbol with cathode (C) and anode (A) leads. (C) Voltage-current response curve. Reverse biased response is the same as normal PN junction diodes. Forward biased response acts as a hysteresis switch. Resistance is very high until the bias voltage reaches V_{B0} and exceeds the cutoff current, I_{B0}. The device exhibits a negative resistance with the current increases as the bias voltage decreases until a voltage of V_H and saturation current of I_H is reached. After this the resistance is very low, with large increases in current for small voltage increases.

PNPN Diode

If four alternate layers of P-type and N-type material are placed together, a PNPN (usually pronounced like *pinpin*) diode with three junctions is obtained (see **Fig 8.20**). This device, when the anode is at a higher potential than the cathode, has its first and third junctions forward biased and its center junction reverse biased. In this state, there is little current, just as in the reverse biased diode. As the forward bias voltage is increased, the current through the device increases slowly until the *breakover (or firing) voltage*, V_{BO}, is reached and the flow of current abruptly increases. The PNPN diode is often considered to be a switch that is off below V_{BO} and on above it.

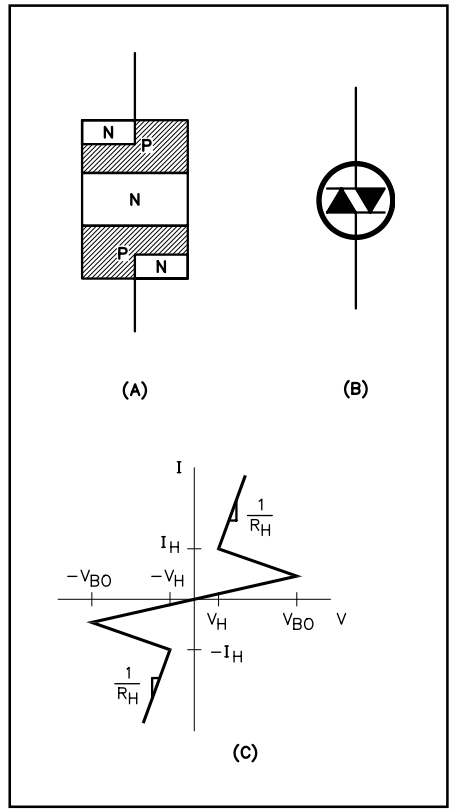

Fig 8.21 — Bilateral switch. (A) Alternating layers of P-type and N-type semiconductor. (B) Schematic symbol. (C) Voltage-current response curve. The right-hand side of the curve is identical to the PNPN diode response in Fig 8.20. The device responds identically for both forward and reverse bias so the left-hand side of the curve is symmetrical to the right-hand

Bilateral Diode Switch

A semiconductor device similar to two PNPN diodes facing in opposite directions and attached in parallel is the *bilateral diode switch* or *diac*. This device has the characteristic curve of the PNPN diode for both positive and negative bias voltages. Its construction, schematic symbol and characteristic curve are shown in **Fig 8.21**.

Silicon Controlled Rectifier

Another device with four alternate layers of P-type and N-type semiconductor is the *silicon controlled rectifier (SCR)*, or *thyristor*. In addition to the connections to the outer two layers, two other terminals can be brought out for the inner two layers. The connection to the P-type material near the cathode is called the *cathode gate* and the N-type material near the anode is called the *anode gate*. In nearly all commercially available SCRs, only the cathode gate is connected (**Fig 8.22**).

Like the PNPN diode switch, the SCR is used to abruptly start conducting when the voltage exceeds a given level. By biasing the gate terminal appropriately, the breakover voltage can be adjusted. The SCR is highly efficient and is used in power control applications. SCRs are available that can handle currents of greater than 100 A and voltage differentials of greater than 1000 V, yet can be switched with gate currents of less than 50 mA.

Triac

A five layered semiconductor whose operation is similar to a bidirectional SCR is the *triac* (**Fig 8.23**). This is also similar to a bidirectional diode switch with a bias control gate. The gate terminal of the triac can control both positive and negative breakover voltages and the devices can pass both polarities of voltage.

SCRs and triacs are often used to modify ac power sources. A sine wave with a given RMS value can be switched on and off at preset points during the cycle to decrease the RMS voltage. When conduction is delayed until after the peak (as **Fig 8.24** shows) the peak-to-peak voltage is reduced. If conduction starts before the peak, the RMS voltage is reduced, but the peak-to-peak value remains the same. This method is used to operate light dimmers and 240 V ac to 120 V ac converters. The sharp switching transients created when

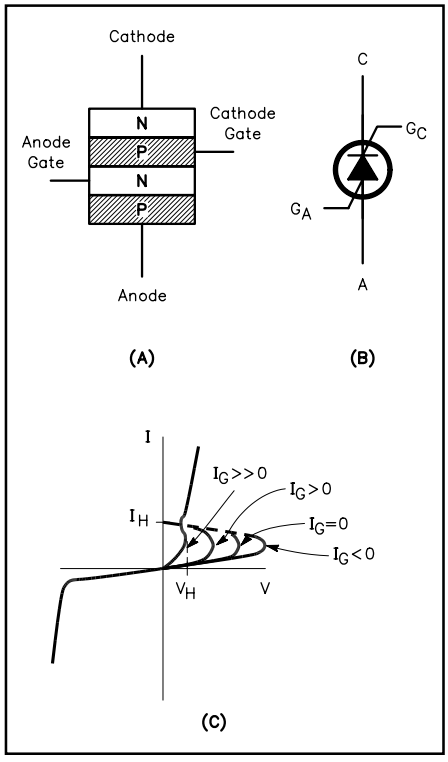

Fig 8.22 — SCR. (A) Alternating layers of P-type and N-type semiconductor. This is similar to a PNPN diode with gate terminals attached to the interior layers. (B) Schematic symbol with anode (A), cathode (C), anode gate (G_A) and cathode gate (G_C). Many devices are constructed without G_A. (C) Voltage-current response curve with different responses for various gate currents. $I_G = 0$ has the same response as the PNPN diode.

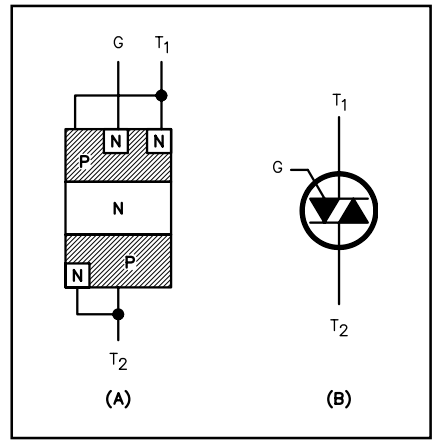

Fig 8.23 — Triac. (A) Alternating layers of P-type and N-type semiconductor. This behaves as two SCR devices facing in opposite directions with the anode of one connected to the cathode of the other and the cathode gates connected together. (B) Schematic symbol.

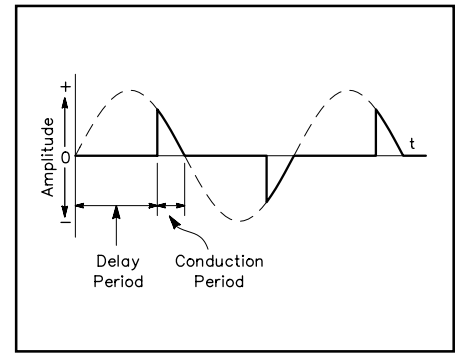

Fig 8.24 — Triac operation on sine wave. The dashed line is the original sine wave and the solid line is the portion that conducts through the triac. The relative delay and conduction period times are controlled by the amount or timing of gate current, I_G. The response of an SCR is the same as this for positive voltages (above the x-axis) and with no conduction for negative voltages.

these devices switch are common sources of RF interference. SCRs are used as "crowbars" in power supply circuits, to short the output to ground and blow a fuse when an overvoltage condition exists.

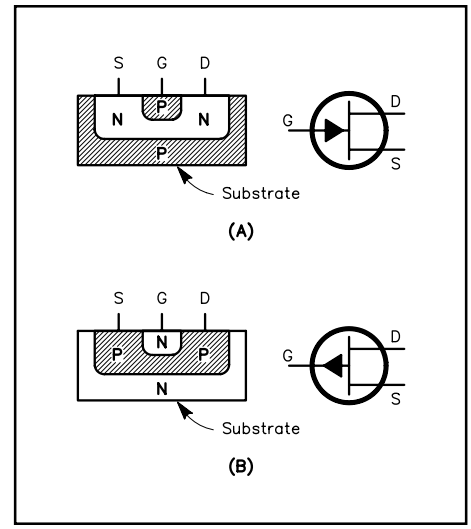

Fig 8.25 — JFET devices with terminals labeled: source (S), gate (G) and drain (D). A) Pictorial of N-type channel embedded in P-type substrate and schematic symbol. B) P-channel embedded in N-type substrate and schematic symbol.

FIELD-EFFECT TRANSISTORS

The *field-effect transistor (FET)* controls the current between two points but does so differently than the bipolar transistor. The FET operates by the effects of an electric field on the flow of electrons through a single type of semiconductor material. This is why the FET is sometimes called a *unipolar* transistor. Also, unlike bipolar semiconductors that can be arranged in many configurations to provide diodes, transistors, photoelectric devices, temperature sensitive devices and so on, the field effect is usually only used to make transistors, although FETs are also available as special-purpose diodes, for use as constant current sources.

Current moves within the FET in a channel, from the source connection to the drain connection. A gate terminal generates an electric field that controls the current (see **Fig 8.25**). The channel is made of either N-type or P-type semiconductor material; an FET is specified as either an N-channel or P-channel device. Majority carriers flow from source to drain. In N-channel devices, electrons flow so the drain potential must be higher than that of the source ($V_{DS} > 0$). In P-channel devices, the flow of holes requires that $V_{DS} < 0$. The polarity of the electric field that controls current in the channel is determined by the majority carriers of the channel, ordinarily positive for P-channel FETs and negative for N-channel FETs.

Variations of FET technology are based on different ways of generating the electric field. In all of these, however, electrons at the gate are used only for their charge in order to create an electric field around the channel, and there is a minimal flow of electrons through the gate. This leads to a very high dc input resistance in devices that use FETs for their input circuitry. There may be quite a bit of capacitance between the gate and the other FET terminals, however. The input impedance may be quite low at RF.

The current through an FET only has to pass through a single type of semiconductor material. There is very little resistance in the absence of an electric field (no bias voltage). The drain-source resistance ($r_{DS\ ON}$) is between a few hundred ohms to less than an ohm. The output impedance of devices made with FETs is generally quite low. If a gate bias voltage is added to operate the transistor near cutoff, the circuit output impedance may be much higher.

FET devices are constructed on a *substrate* of doped semiconductor material. The channel is formed within the substrate and has the opposite polarity (a P-channel FET has N-type substrate). Most FETs are constructed with silicon. In order to achieve a higher gain-bandwidth product, other materials have been used. Gallium Arsenide (GaAs) has electron mobility and drift velocities that are far higher than the standard doped silicon. Amplifiers designed with *GaAs FET* devices have much higher frequency response and lower noise factor at VHF and UHF than those made with standard FETs.

JFET

There are two basic types of FET. In the *junction FET (JFET)*, the gate material is made of the opposite polarity semiconductor to the channel material (for a P-channel FET the gate is made of N-type semiconductor material). The gate-channel junction is similar to a diode's PN junction. As with the diode,

current is high if the junction is forward biased and is extremely small when the junction is reverse biased. The latter case is the way that JFETs are used, since any current in the gate is undesirable. The magnitude of the reverse bias at the junction is proportional to the size of the electric field that "pinches" the channel. Thus, the current in the channel is reduced for higher reverse gate bias voltage.

Because the gate-channel junction in a JFET is similar to a bipolar junction diode, this junction must never be forward biased; otherwise large currents will pass through the gate and into the channel. For an N-channel JFET, the gate must always be at a lower potential than the source ($V_{GS} < 0$). The channel is as fully open as it can get when the gate and source voltages are equal ($V_{GS} = 0$). The prohibited condition is when $V_{GS} > 0$. For P-channel JFETs these conditions are reversed (in normal operation V_{GS} 0 and the prohibited condition is when $V_{GS} < 0$).

MOSFET

Placing an insulating layer between the gate and the channel allows for a wider range of control (gate) voltages and further decreases the gate current (and thus increases the device input resistance). The insulator is typically made of an oxide (such as silicon dioxide, SiO_2). This type of device is called a *metal-oxide-semiconductor FET (MOSFET)* or *insulated-gate FET (IGFET)*. The substrate is often connected to the source internally. The insulated gate is on the opposite side of the channel from the substrate (see **Fig 8.26**). The bias voltage on the gate terminal either attracts or repels the majority carriers of the substrate across the PN junction with the channel. This narrows (depletes) or widens (enhances) the channel, respectively, as V_{GS} changes polarity. For N-channel MOSFETs, positive gate voltages with respect to the substrate and the source ($V_{GS} > 0$) repel holes from the channel into the substrate, thereby widening the channel and decreasing channel resistance. Conversely, $V_{GS} < 0$ causes holes to be attracted from the substrate, narrowing the channel and increasing the channel resistance. Once again, the polarities discussed in this example are reversed for

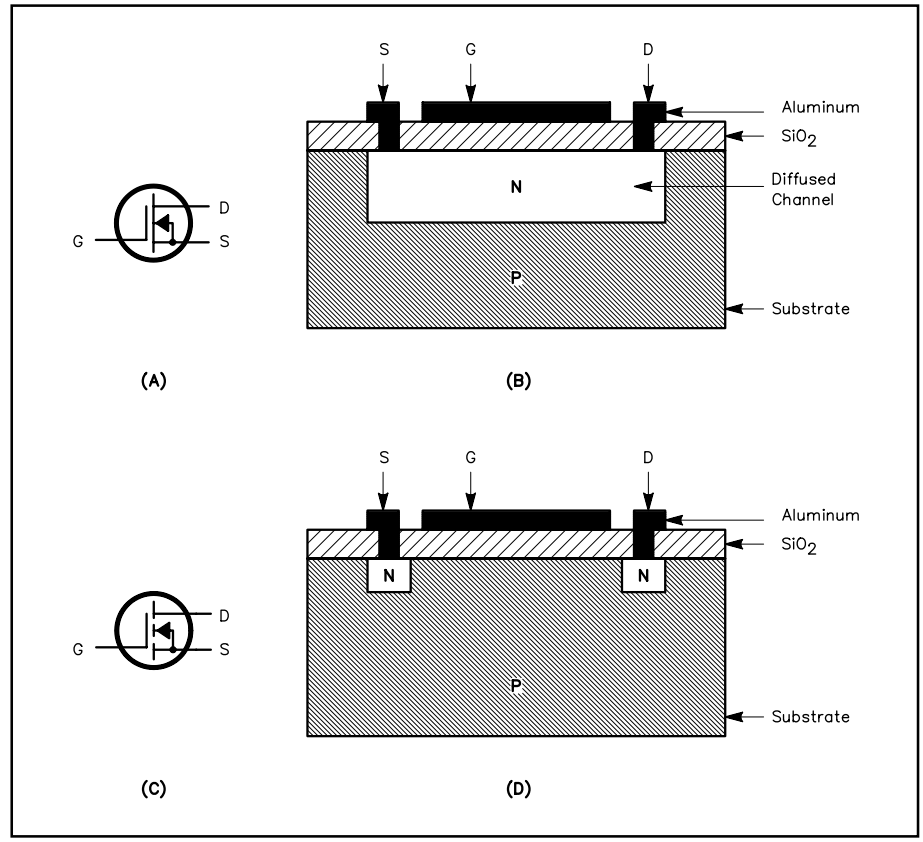

Fig 8.26 — MOSFET devices with terminals labeled: source (S), gate (G) and drain (D). N-channel devices are pictured. P-channel devices have the arrows reversed in the schematic symbols and the opposite type semiconductor material for each of the layers. (A) N-channel depletion mode device schematic symbol and pictorial of P-type substrate, diffused N-type channel, SiO_2 insulating layer and aluminum gate region and source and drain connections. The substrate is connected to the source internally. A negative gate potential narrows the channel. (B) N-channel enhancement mode device schematic and pictorial of P-type substrate, N-type source and drain wells, SiO_2 insulating layer and aluminum gate region and source and drain connections. Positive gate potential forms a channel between the two N-type wells.

P-channel devices. The common abbreviation for an N-channel MOSFET is *NMOS*, and for a P-channel MOSFET, *PMOS*.

Because of the insulating layer next to the gate, input resistance of a MOSFET is usually greater than 10^{12} Ω (a million megohms). Since MOSFETs can both deplete the channel, like the JFET, and also enhance it, the construction of MOSFET devices differs based on the channel size in the resting state, $V_{GS} = 0$. A *depletion mode* device (also called a *normally on MOSFET*) has a channel in resting state that gets smaller as a reverse bias is applied; this device conducts current with no bias applied (see Fig 8.26 A and B). An *enhancement mode* device (also called a *normally off MOSFET*) is built without a channel and does not conduct current when $V_{GS} = 0$; increasing forward bias forms a channel that conducts current (see Fig 8.26 C and D).

Semiconductor Temperature Effects

The number of excess holes and electrons is increased as the temperature of a semiconductor increases. Since the conductivity of a semiconductor is related to the number of excess carriers, this also increases with temperature. With respect to resistance, semiconductors have a negative temperature coefficient. The resistance of silicon *decreases* by about 8% / °C and by about 6% / °C for germanium. Semiconductor temperature properties are the opposite of most metals, which *increase* their resistance by about 0.4% / °C. These opposing temperature characteristics permit the design of circuits with opposite temperature coefficients that cancel each other out, making a temperature insensitive circuit. Left by itself, the semiconductor can experience an effect called *thermal runaway* as the current causes an increase in temperature. The increased temperature decreases resistance and may lead to a further increase in current (depending on the circuit) that leads to an additional temperature increase. This sequence of events can continue until the semiconductor destroys itself.

Semiconductor Failure

There are several common failure modes for semiconductors that are related to heat. The semiconductor material is connected to the outside world through metallic leads. The point at which the metal and the semiconductor are connected is one common place for the semiconductor device to fail. As the device heats up and cools down, the materials expand and contract. The rate of expansion and contraction of semiconductor material is different from that of metal. Over many cycles of heating and cooling the bond between the semiconductor and the metal can break. Some experts have suggested that the lifetime of semiconductor equipment can be extended by leaving the devices on all the time. While this would decrease the type of failure just described, inadequate cooling can lead to another type of semiconductor failure.

Impurities are introduced into intrinsic semiconductors by diffusion, the same physical property that lets you smell cookies baking from several rooms away. Smells diffuse through air much faster than molecules diffuse through solids. Once the impurities diffuse into the semiconductor, they tend to stay in place. Rates of diffusion are proportional to temperature, and semiconductors are doped with impurities at high temperature to save time. Once the doped semiconductor material is cooled, the rate of diffusion of the impurities is so low that they are essentially immobile for many years to come.

A common failure mode of semiconductors is due to the heat generated during semiconductor use. If the temperatures at the junctions rise to high enough levels for long enough periods of time, the impurities start to diffuse across the PN junctions. When enough of these atoms get across the junction, it stops functioning properly and the semiconductor device fails.

Thermistors

The effect of temperature on current in semiconductors is put to use in a controlled fashion in a *thermistor*. The temperature coefficients of silicon and germanium are highly dependent on the amount

of doping. For stability, thermistors are made of oxides such as nickel oxide (NiO), dimanganese trioxide (Mn_2O_3) or dicobalt trioxide (Co_2O_3). If the doping concentration of a semiconductor is high enough, it will start to take on some of the properties of a metal and the temperature coefficient becomes positive. A device made from this type of material is sometimes called a *sensistor*.

Practical Semiconductors

SEMICONDUCTOR DIODES

Although many types of semiconductor diodes are available, there are not many differences between them. The diode is made of a single PN junction that affects current differently depending on its direction. This leads to a large number of applications in electronic circuitry.

The diode symbol is shown in **Fig 8.27**. Current passes most easily from anode to cathode, in the direction of the arrow. This is often referred to as the *forward* direction and the opposite is the *reverse* direction. Remember that *current* refers to the flow of electricity from higher to lower potentials and is in the opposite direction to the flow of electrons (current moves from anode to cathode and electrons flow from cathode to anode, as based on the definitions of the words, *anode* and *cathode*). The anode of a semiconductor junction diode is made of P-type material and the cathode is made of N-type material, as indicated in Fig 8.27. Most diodes are marked with a band on the cathode end (Fig 8.27). The ideal diode would have zero resistance in the forward direction and infinite resistance in the reverse direction. This is not the case for actual devices, which behave as shown in the plot of a diode response in Fig 8.18. Note that the scales of the two parts of the graph are drastically different. The inverse of the slope of the line (the change in voltage between two points on a straight portion of the line divided by the corresponding change in current) on the upper right is the resistance of the diode in the forward direction. The range of voltages is small and the range of currents is large since the forward resistance is very small (in this example, about 2 Ω). The lower left portion of the curve illustrates a much higher resistance that increases from tens of kilohms to thousands of megohms as the reverse voltage gets larger, and then decreases to near zero (a nearly vertical line) very suddenly at the peak inverse voltage (PIV = 100 V in this example).

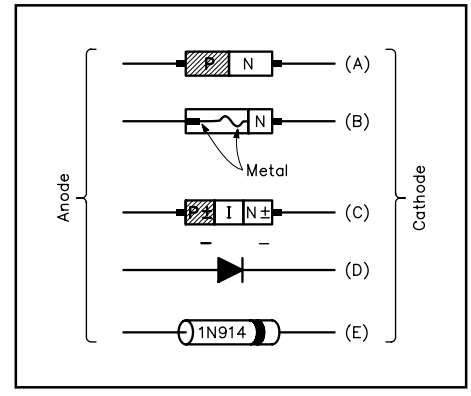

Fig 8.27 — Practical semiconductor diodes. All devices are aligned with anode on the left and cathode on the right. (A) Standard PN junction diode. (B) Point-contact or "cat's whisker" diode. (C) PIN diode formed with heavily doped P-type (P^+), undoped (intrinsic) and heavily doped N-type (N^+) semiconductor material. (D) Diode schematic symbol. (E) Diode package with marking stripe on the cathode end.

There are five major characteristics that distinguish standard junction diodes from one another: the PIV, the current or power handling capacity, the response speed, reverse leakage current and the junction barrier voltage. Each of these characteristics can be manipulated during manufacture to produce special purpose diodes.

The most common application of a diode is to perform rectification; that is, allowing positive voltages to pass and stopping negative voltages. Rectification is used in power supplies that convert ac to dc and in amplitude demodulation. The most important diode parameters to consider for power rectification are the PIV and current ratings. The peak negative voltages that are stopped by the diode must be smaller in magnitude than the PIV and the peak current through the diode when it is forward biased must be less than the maximum amount for which the device was designed. Exceeding the current rating in a diode will cause excessive heating (based on $P = I \times V_F$) that leads to PN junction failure as described earlier.

Fast Diodes

The speed of a diode affects the frequencies that it can act on. The diode response in Fig 8.18 is a steady state response, showing how that diode will act at dc. As the frequency increases, the diode may not be able to keep up with the changing polarity of the signal and its response will not be as expected. Diode speed mainly depends on charge storage in the depletion region. Under reverse bias, excess charges move away from the junction, forming a larger space-charge region that is the equivalent of a dielectric. The diode thus exhibits capacitance, which is inversely proportional to the width of the dielectric and directly proportional to the cross-sectional surface area of the junction.

One way to decrease charge storage time in the depletion region is to form a metal-semiconductor junction. This can be accomplished with a point-contact diode, where a thin piece of aluminum wire, often called a *whisker*, is placed in contact with one face of a piece of lightly doped N-type material. In fact, the original diodes used for detecting radio signals ("cat's whisker diodes") were made this way. A more recent improvement to this technology, the *hot-carrier diode*, is like a point-contact diode with more ideal characteristics attained by using more efficient metals, such as platinum and gold, that act to lower forward resistance and increase PIV. This type of contact is known as a *Schottky barrier*, and diodes made this way are called *Schottky diodes*.

The PIN diode, shown in Fig 8.27C is a *slow response* diode that is capable of passing microwave signals when it is forward biased. This device is constructed with a layer of intrinsic (undoped) semiconductor placed between very highly doped P-type and N-type material (called P^+-type and N^+-type material to indicate the high level of doping), creating a PIN junction. These devices provide very effective switches for RF signals and are often used in TR switches in transceivers. PIN diodes have longer than normal carrier lifetimes, resulting in a slow switching process that causes them to act more like resistors than diodes at high radio frequencies.

Varactors

If the PN junction capacitance is controlled rather than reduced, a diode can be made to act as a variable capacitor. As the reverse bias voltage on a diode increases, the width of the junction increases, which decreases its capacitance. A *varactor* is a diode whose junction is specially formulated to have a relatively large range of capacitance values for a modest range of reverse bias voltages (**Fig 8.28**). Although special forms of varactors are available from manufacturers, other types of diodes may be used as inexpensive varactor diodes, but the relationship between reverse voltage and capacitance is not always reliable. When designing with varactor diodes, the reverse bias voltage must be absolutely free of noise since any variations in the bias voltage will cause changes in capacitance. Unwanted frequency shifts or instability will result if the reverse bias voltage is noisy. It is possible to frequency modulate a signal by adding the audio signal to the reverse bias on a varactor diode used in the carrier oscillator.

Zener Diodes

When the PIV of a reverse biased diode is exceeded, the diode begins to conduct current as it does when it is forward biased. This current does not destroy the diode if it is limited to less than the device's maximum allowable value. When the PIV is controlled during manufacture to be at desired levels, the device is called a *Zener diode*. Zener diodes (named after the American physicist Clarence Zener) provide accurate voltage references and are often used for this purpose in power supply regulators.

When the reverse breakdown voltage is exceeded, the reverse voltage drop across the Zener diode remains constant. With an appropriate current limiting resistor in series with it, the Zener diode provides an accurate voltage reference (**Fig 8.29**). Zener diodes are rated by their reverse breakdown voltage and their power handling capacity. The power is a product of the current passing through the reverse biased Zener diode "in breakdown" (that is, in the breakdown mode of operation) and the breakdown voltage.

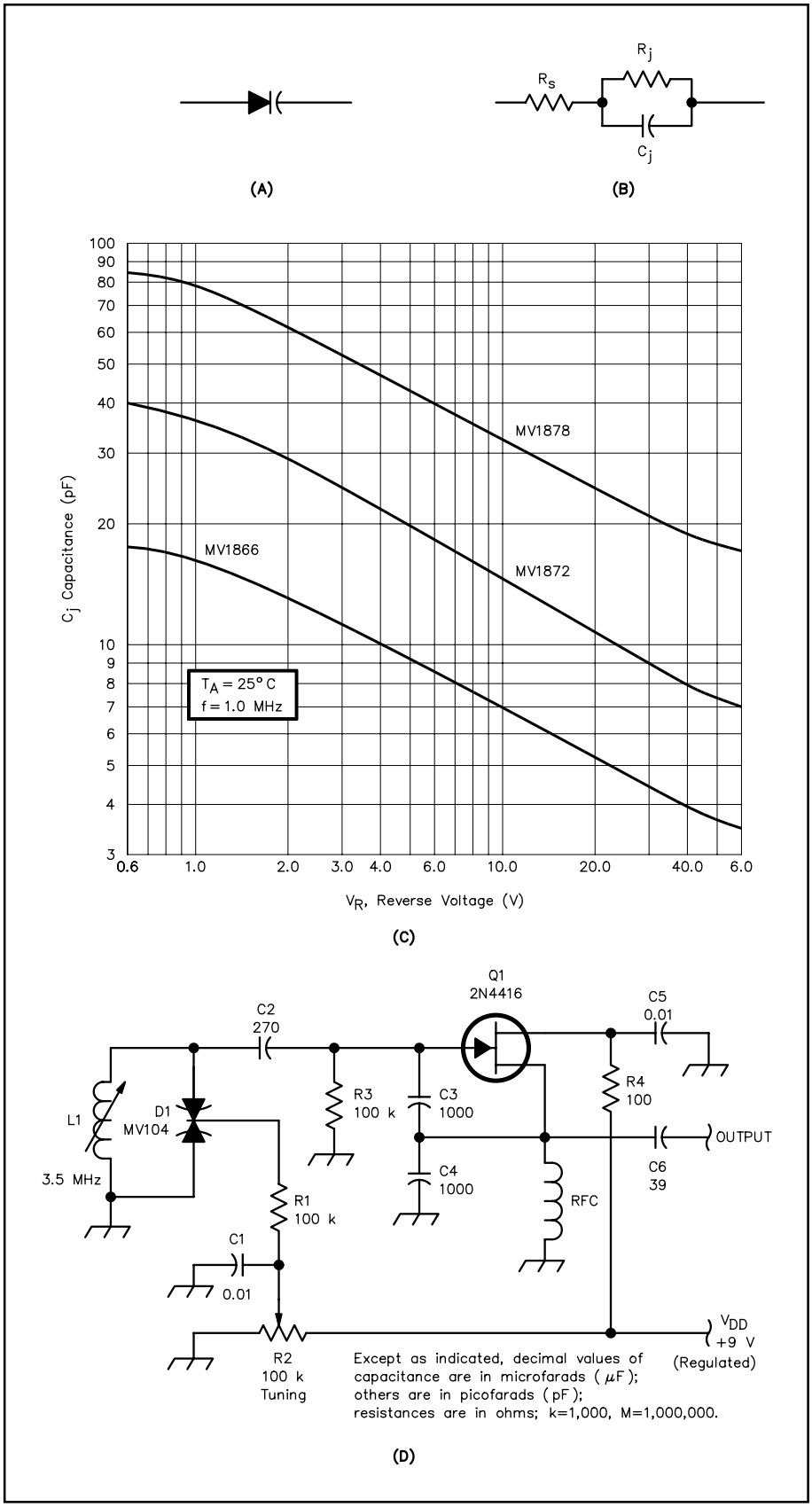

Fig 8.28 — Varactor diode. (A) Schematic symbol. (B) Equivalent circuit of the reverse biased varactor diode. R_S is the junction resistance, R_J is the leakage resistance and C_J is the junction capacitance, which is a function of the magnitude of the reverse bias voltage. (C) Plot of junction capacitance, C_J, as a function of reverse voltage, V_R, for three different varactor devices. Both axes are plotted on a logarithmic scale. (D) Oscillator circuit with varactor tuning. D1-L1 is a tuned circuit with a dual varactor diode that is controlled by the voltage from potentiometer R2. C1 is a filter capacitor to insure that the varactor bias voltage is clean dc. C2 and C6 are dc blocking capacitors. Q1 is an N-channel JFET in common drain configuration with feedback to the gate through C3. R3 is the gate bias resistor. R4 is the drain voltage resistor with filter capacitor C5.

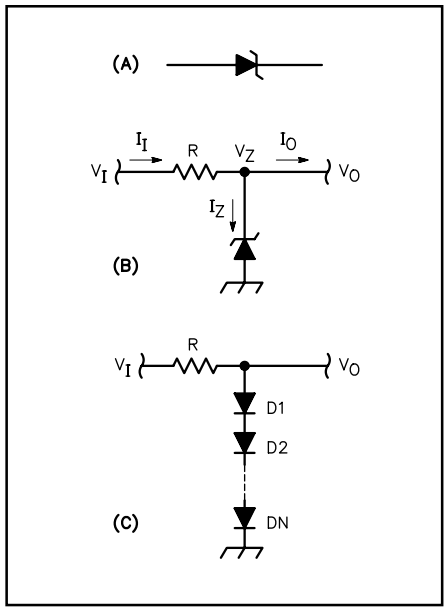

Fig 8.29 — Zener diode. (A) Schematic symbol. (B) Basic voltage regulating circuit. V_Z is the Zener reverse breakdown voltage. The Zener diode draws more current until $V_I - I_iR = V_Z$. The circuit design should select R so that when the maximum current is drawn, $R < (V_I - V_Z) / I_O$. The diode should be capable of passing the same current when there is no output current drawn. (C) For small voltages, several forward biased diodes can be used in place of Zener diodes. Each diode will drop the voltage by about 0.7 V for silicon or 0.3 V for germanium.

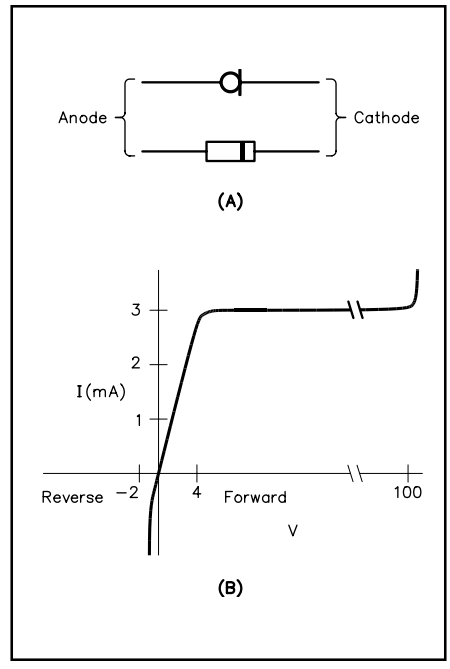

Fig 8.30 — Current regulator diode. (A) Schematic symbol and package with line marking cathode end. (B) Diode characteristic curve (1N5283 device). When forward bias voltage exceeds about 4 V the current passing through the device is held constant regardless of the voltage across the device.

Since the same current must always pass though the resistor to drop the source voltage down to the reference voltage, with that current divided between the Zener diode and the load, this type of power source is very wasteful of current. The Zener diode does make an excellent and efficient voltage reference in a larger voltage regulating circuit where the load current is provided from another device whose voltage is set by the reference. (See the **Power Supplies and Projects** chapter for more information about using Zener diodes as voltage regulators.) The major sources of error in Zener-diode derived voltages are the variation with load current and the variation due to heat. Temperature compensated Zener diodes are available with temperature coefficients as low as 0.0005 % / °C. If this is unacceptable, voltage reference integrated circuits based on Zener diodes have been developed that include additional circuitry to counteract temperature effects.

Constant Current Diodes

A form of diode, called a *field-effect regulator diode*, provides a constant current over a wide range of forward biased voltages. The schematic symbol and characteristic curve for this type of device are shown in **Fig 8.30**. Constant current diodes are very useful in any application where a constant current is desired. Some part numbers are 1N5283 through 1N5314.

Common Diode Applications

Standard semiconductor diodes have many uses in analog circuitry. Several examples of diode circuits are shown in **Fig 8.31**. Rectification has already been described. There are three basic forms of rectification using semiconductor diodes: half wave (1 diode), full-wave center-tapped (2 diodes) and full-wave bridge (4 diodes). These are more fully described in the **Power Supplies and Projects** chapter.

Diodes are commonly used to protect circuits. In battery powered devices a forward biased series diode is often used to protect the circuitry from the user inadvertently inserting the batteries backwards. Likewise, when a circuit is powered from an external dc source, a diode is often placed in series with the power connector in the device to prevent incorrectly wired power supplies from destroying the

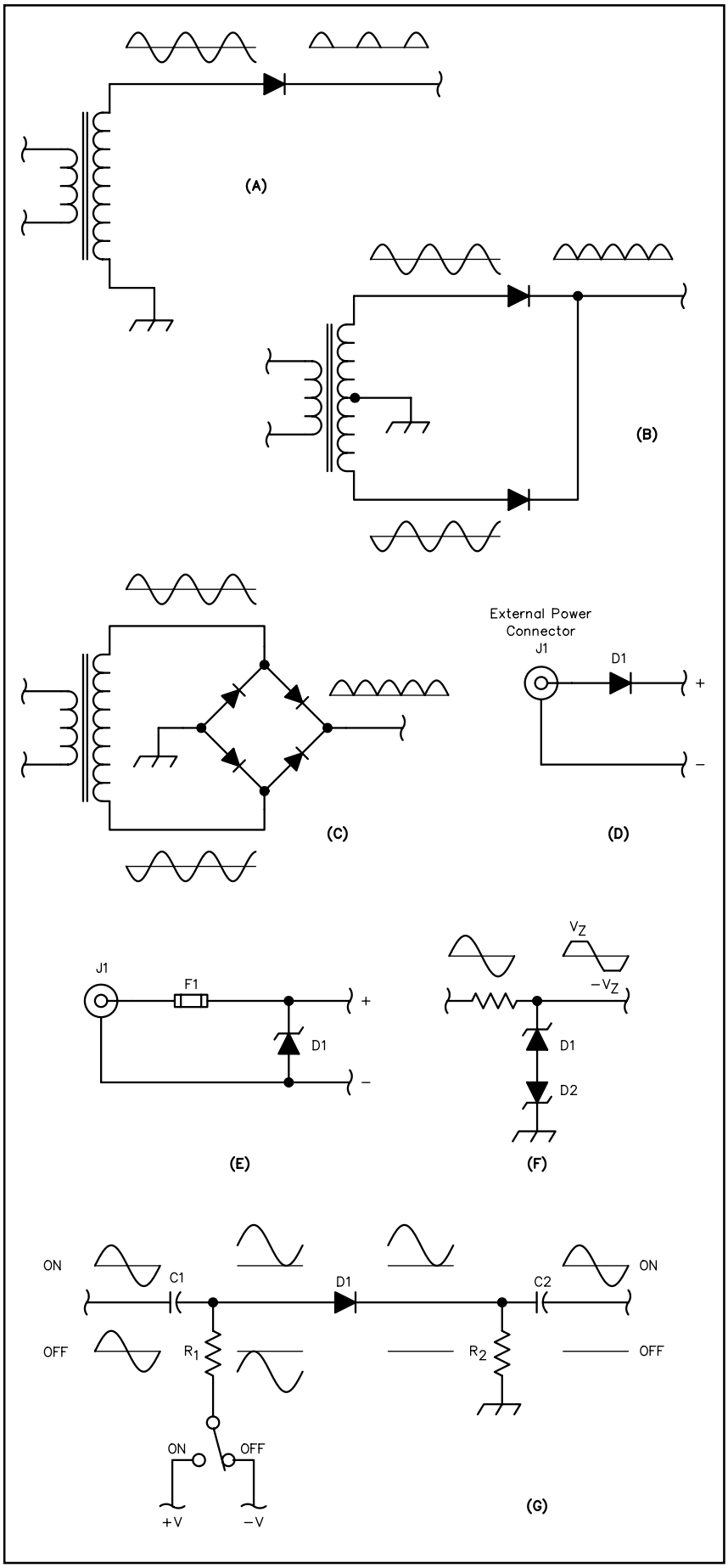

Fig 8.31 — Diode circuits. (A) Half wave rectifier circuit. Only when the ac voltage is positive does current pass through the diode. Current flows only during half of the cycle. (B) Full-wave center-tapped rectifier circuit. Center-tap on the transformer secondary is grounded and the two ends of the secondary are 180° out of phase. During the first half of the cycle the upper diode conducts and during the second half of the cycle the lower diode conducts. There is conduction during the full cycle with only positive voltages appearing at the output. (C) Full-wave bridge rectifier circuit. In each half of the cycle two diodes conduct capacity. (D) Polarity protection for external power connection. J1 is the connector that power is applied to. If polarity is correct, the diode will conduct and if reversed the diode will block current, protecting the circuit that is being powered. (E) Over-voltage protection circuit. If excessive voltage is applied to J1, D1 will conduct current until fuse, F1, is blown. (F) Bipolar voltage clipping circuit. In the positive portion of the cycle, D2 is forward biased but no current is shunted to ground because D1 is reverse biased. D1 starts to conduct when the voltage exceeds the Zener breakdown voltage and the positive peak is clipped. When the negative portion of the cycle is reached, D1 is forward biased but no current is shunted to ground because D2 is reverse biased. When the voltage exceeds the Zener breakdown voltage of D2, it also begins to conduct and the negative peak is clipped. (G) Diode switch. The signal is ac coupled to the diode by C1 at the input and C2 at the output. R2 provides a reference for the bias voltage. When switch S1 is in the ON position, a positive dc voltage is added to the signal so it is forward biased and is passed through the diode. When S1 is in the OFF position, the negative dc voltage added to the signal reverse biases the diode and the signal does not get through.

equipment. Diodes are commonly used to protect analog meters from both reverse voltage and over voltage conditions that would destroy the delicate needle movement.

Zener diodes are sometimes used to protect low-current (a few amps) circuits from over-voltage conditions. A reverse biased Zener diode connected between the positive power lead and ground will conduct excessive current if its breakdown voltage is exceeded. Used in conjunction with a fuse in series with the power lead, the Zener diode will cause the fuse to blow when an over-voltage condition exists.

Very high, short-duration voltage spikes can destroy certain semiconductors, particularly MOS devices. Standard Zener diodes can't handle the high pulse powers found in these voltage spikes. Special Zener diodes are designed for this purpose, such as the *mosorb*. (General Semiconductor Industries, Inc calls these devices *TransZorbs*.) A reverse biased TransZorb with a low-value series resistor can decrease the voltage reaching the sensitive device. Since the polarity of the spike can be positive, negative, or both, over voltage transient suppressor circuits can be designed with two devices wired back-to-back. They protect a circuit over a range of voltages rather than just suppressing positive peaks.

Diodes can be used to clip signals, similar to rectification. If the signal is appropriately biased it can be clipped at any level. Two Zener diodes placed back-to-back can be used to clip both the positive and negative peaks of a signal. Such an arrangement is used to convert a sine wave to an approximate square wave.

Care must be taken when using Zener diodes to process signals. The Zener diode is a relatively noisy device and can add excessive noise to the signals if it operates in breakdown. The Zener diode is often specified for at intentionally generate noise, such as the noise bridge (see the **Test Procedures and Projects** chapter). The reverse biased Zener diode in breakdown generates wide band (nearly white) noise levels as high as 2000 $\mu V / \sqrt{Hz}$. (The noise voltage is determined by multiplying this value by the square root of the circuit bandwidth in Hz.)

Diodes are used as switches for ac coupled signals when a dc bias voltage can be added to the signal to permit or inhibit the signal from passing through the diode. In this case the bias voltage must be added to the ac signal and be of sufficient magnitude so that the entire envelope of the ac signal is above or below the junction barrier voltage, with respect to the cathode, to pass through the diode or inhibit the signal. Special forms of diodes, such as the PIN diode described earlier, which are capable of passing higher frequencies, are used to switch RF signals.

BIPOLAR TRANSISTORS

The bipolar transistor is a *current-controlled device*. The current between the emitter and the collector is governed by the current that enters the base. The convention when discussing transistor operation is that the three currents into the device are positive (I_c into the collector, I_b into the base and I_e into the emitter). Kirchhoff's current law applies to transistors just as it does to passive electrical networks: the total current entering the device must be zero. Thus, the relationship between the currents into a transistor can be generalized as

$$0 = I_c + I_b + I_e \tag{9}$$

which can be rearranged as necessary. For example, if we are interested in the emitter current,

$$I_e = -(I_c + I_b) \tag{10}$$

The back-to-back diode model is appropriate for visualization of transistor construction. In actual transistors, however, the relative sizes of the collector, base and emitter regions differ. A common transistor configuration that spans a distance of 3 mm between the collector and emitter contacts typically has a base region that is only 25 μm across.

Current conduction between collector and emitter is described by regions in the common-base response curves of the transistor device (see **Fig 8.32**). The transistor is in its *active region* when the base-

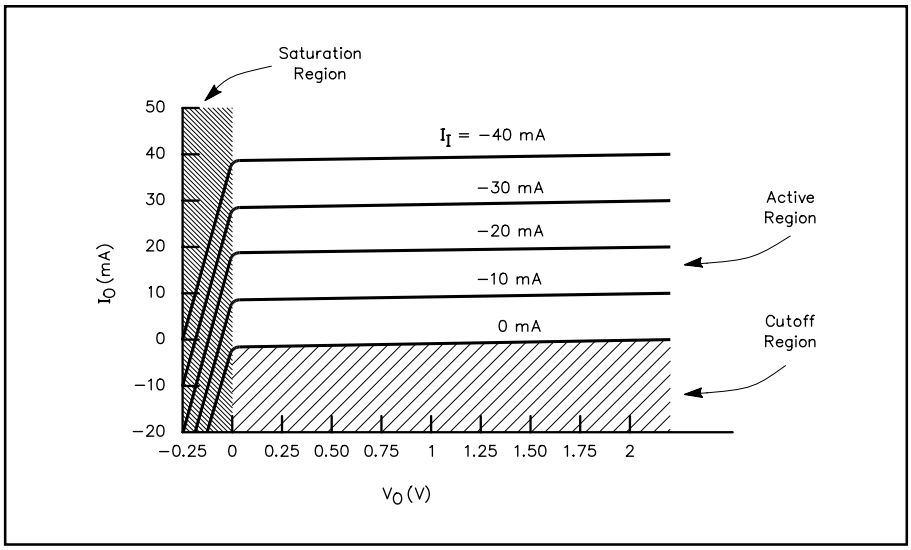

Fig 8.32 — Transistor response curve. The x-axis is the output voltage and the y-axis is the output current. Different curves are plotted for various values of input current. The three regions of the transistor are its cutoff region, where no current flows in any terminal, its active region, where the output current is nearly independent of the output voltage and there is a linear relationship between the input current and the output current, and the saturation region, where the output current has large changes for small changes in output voltage.

collector junction is reverse biased and the base-emitter junction is forward biased. The slope of the output current (I_O) versus the output voltage (V_O) is virtually flat, indicating that the output current is nearly independent of the output voltage. The slight slope that does exist is due to base-width modulation (known as the "Early effect"). Under these conditions, there is a linear relationship between the input current (I_I) and I_O. When both the junctions in the transistor are forward biased, the transistor is said to be in its *saturation region*. In this region, V_O is nearly zero and large changes in I_O occur for very small changes in V_O. The *cutoff region* occurs when both junctions in the transistor are reverse biased. Under this condition, there is very little current in the output, only the nanoamperes or microamperes that result from the very small leakage across the input-to-output junction. These descriptions of junction conditions are the basis for the use of transistors. Various configurations of the transistor in circuitry make use of the properties of the junctions to serve different purposes in analog signal processing.

In the common base configuration, where the input is at the emitter and the output is at the collector, the current gain is defined as

$$\alpha = -\frac{\Delta I_C}{\Delta I_E} \approx 1 \tag{11}$$

In the common emitter configuration, with the input at the base and the output at the collector, the current gain is

$$\beta = \frac{\Delta I_C}{\Delta I_B} \tag{12}$$

and the relationship between α and β is defined as

$$\alpha = \frac{\beta}{1+\beta} \tag{13}$$

Since the common-emitter configuration is the most used transistor-amplifier configuration, another designation for β is often used: h_{FE}, the forward dc current gain. (The "h" refers to "h parameters," a set of parameters for describing a two-port network.) The symbol, h_{fe}, is used for the forward current gain of ac signals. Other transistor transfer function relationships that are measured are h_{ie}, the input impedance, h_{oe}, the output admittance (reciprocal of impedance) and h_{re}, the voltage feedback ratio.

The behavior of a transistor can be defined in many ways, depending on which type of amplifier it is wired to be. A complete description of a transistor must include characteristic curves for each configuration. Typically, two sets of characteristic curves are presented: one describing the input behavior and the other describing the output behavior in each amplifier configuration. Different transistor amplifier configurations have different gains, input and output impedances. At low frequencies, where parasitic capacitances aren't a factor, the common emitter configuration has a high current gain (about – 50, with the negative sign indicating a 180° phase shift), medium to high input impedance (about 50 kΩ) and a medium to low output impedance (about 1 kΩ). The common collector has a high current gain (about 50), a high input impedance (about 150 kΩ) and a low output impedance (about 80 Ω). The common base amplifier has a low current gain (about 1), a low input impedance (about 25 Ω) and a very high output

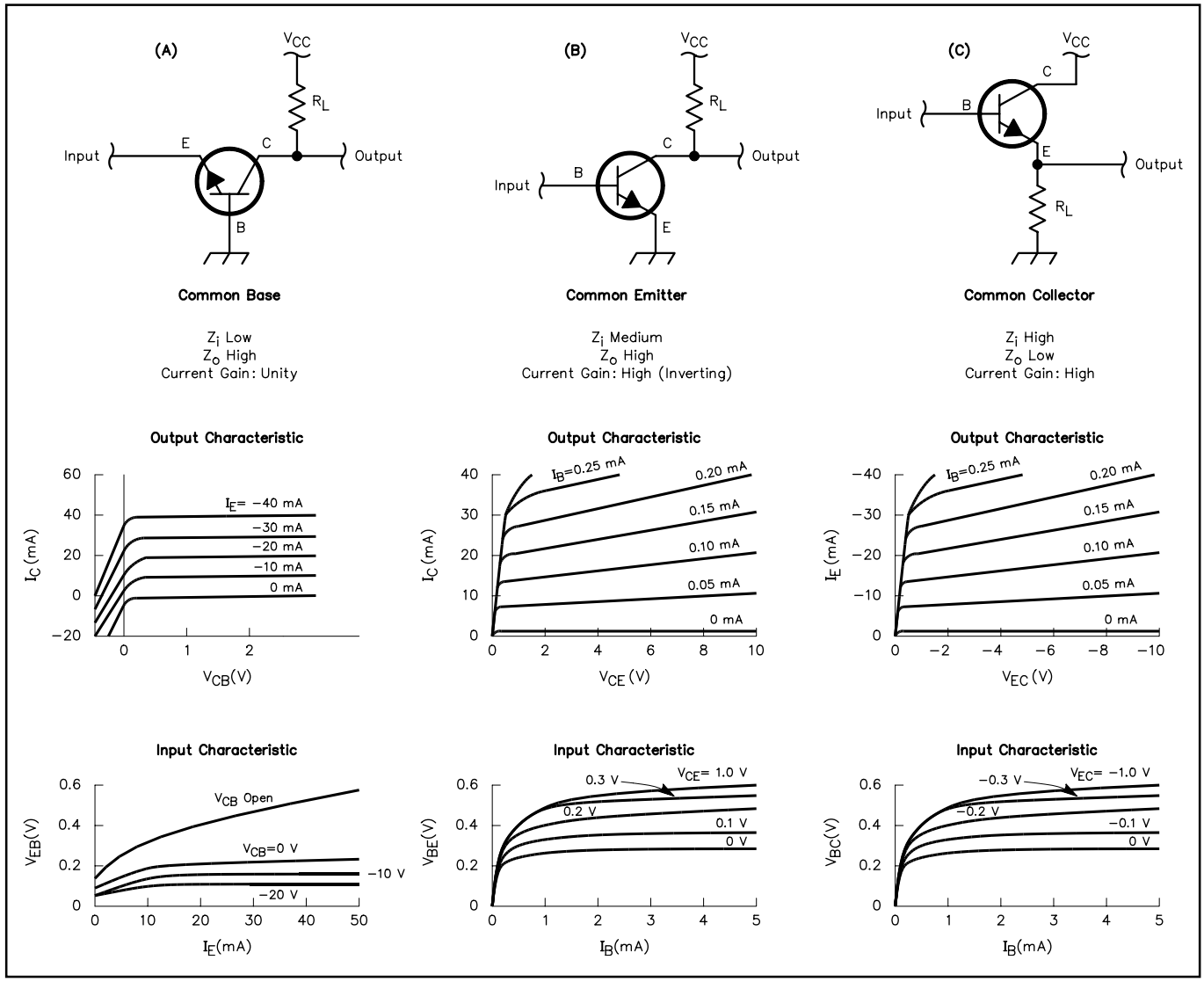

Fig 8.33 — The three configurations of transistor amplifiers. Each has a table of its relative impedance and current gain. The output characteristic curve is plotted for each, with the output voltage along the x-axis, the output current along the y-axis and various curves plotted for different values of input current. The input characteristic curve is plotted for each configuration with input current along the x-axis, input voltage along the y-axis and various curves plotted for different values of output voltage. (A) Common base configuration with input terminal at the emitter and output terminal at the collector. (B) Common emitter configuration with input terminal at the base and output terminal at the collector. (C) Common collector with input terminal at the base and output terminal at the emitter.

impedance (about 2 MΩ). Depending on the intended use of the transistor amplifier in an analog circuit, one configuration will be more appropriate than others. Once the common lead of the transistor amplifier configuration is chosen, the input and output impedance are functions of the device bias levels and circuit loading (**Fig 8.33**). The actual input and output impedances of a transistor amplifier are highly dependent on the input, biasing and load resistors that are used in the circuit.

A typical general-purpose bipolar-transistor data sheet lists important device specifications. Parameters listed in the ABSOLUTE MAXIMUM RATINGS section are the three junction voltages (V_{CEO}, V_{CBO} and V_{EBO}), the continuous collector current (I_C), the total device power dissipation (P_D) and the operating and storage temperature range. In the OPERATING PARAMETERS section, the three guaranteed minimum junction breakdown voltages are listed—$V_{(BR)CEO}$, $V_{(BR)CBO}$ and $V_{(BR)EBO}$—along with the two guaranteed maximum collector cutoff currents—I_{CEO} and I_{CBO}—under OFF CHARACTERISTICS. Under ON CHARACTERISTICS are the guaranteed minimum dc current gain (h_{FE}), guaranteed maximum collector-emitter saturation voltage—$V_{CE(sat)}$—and the guaranteed maximum base-emitter on voltage—$V_{BE(on)}$. The next section is SMALL-SIGNAL CHARACTERISTICS, where the guaranteed minimum current gain-bandwidth product—f_T, the guaranteed maximum output capacitance—C_{obo}, the guaranteed maximum input capacitance—C_{ibo}, the guaranteed range of input impedance—h_{ie}, the small-signal current gain—h_{fe}, the guaranteed maximum voltage feedback ratio—h_{re} and output admittance—h_{oe} are listed. Finally, the SWITCHING CHARACTERISTICS section lists absolute maximum ratings for delay time—t_d, rise time—t_r, storage time—t_s and fall time—t_f.

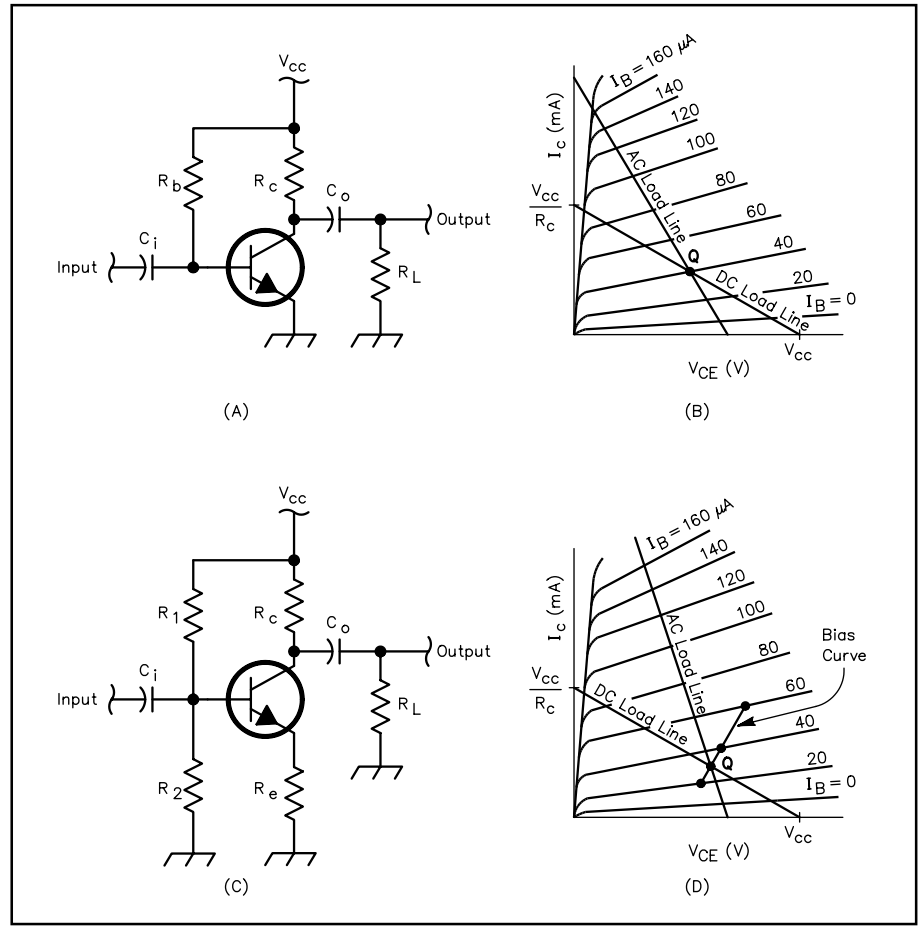

Fig 8.34 — Transistor biasing circuits. (A) Fixed bias. Input signal is ac coupled through C_i. The output has a voltage that is equal to $V_{CC} - I_C \times R_C$. This signal is ac coupled to the load, R_L, through C_o. For dc signals, the entire output voltage is based on the value of R_C. For ac signals, the output voltage is based on the value of R_C in parallel with R_L. (B) Characteristic curve for the transistor amplifier pictured in (A). The slope of the dc load line is equal to $-1 / R_C$. For ac signals, the slope of the ac load line is equal to $-1 / (R_C \parallel R_L)$. The quiescent operating point, Q, is based on the base bias current with no input signal applied and where this characteristic line crosses the dc load line. The ac load line must also pass through point Q. (C) Self-bias. Similar to fixed bias circuit with the base bias resistor split into two: R1 connected to V_{CC} and R2 connected to ground. Also an emitter bias resistor, R_E, is included to compensate for changing device characteristics. (D) This is similar to the characteristic curve plotted in (B) but with an additional "bias curve" that shows how the base bias current varies as the device characteristics change with temperature. The operating point, Q, moves along this line and the load lines continue to intersect it as it changes.

Transistor Biasing

Biasing in a transistor adds or subtracts a fixed amount of current from the signal at the input port. This differs from vacuum tube, FET and operational amplifier biasing where a bias *voltage* is added to the input signal. Fixed bias is the simplest form, as shown in **Fig 8.34A**. The operating point is determined by the intersection between the characteristic curves, the load line and the quiescent current bias line (Fig 8.34B). The problem with fixing the bias current is that if the transistor parameters drift due to heat, the operating point will change. The operating point can be stabilized by self biasing, also called emitter biasing, as pictured in Fig 8.34C. If I_C increases due to temperature changes, the current in R_E increases. The larger current through R_E increases the voltage drop across that resistor, causing a decrease in the base current, I_B. This, in turn, leads to a decreasing I_C, minimizing its variation due to heat. The operating point for this type of biasing is plotted in Fig 8.34D.

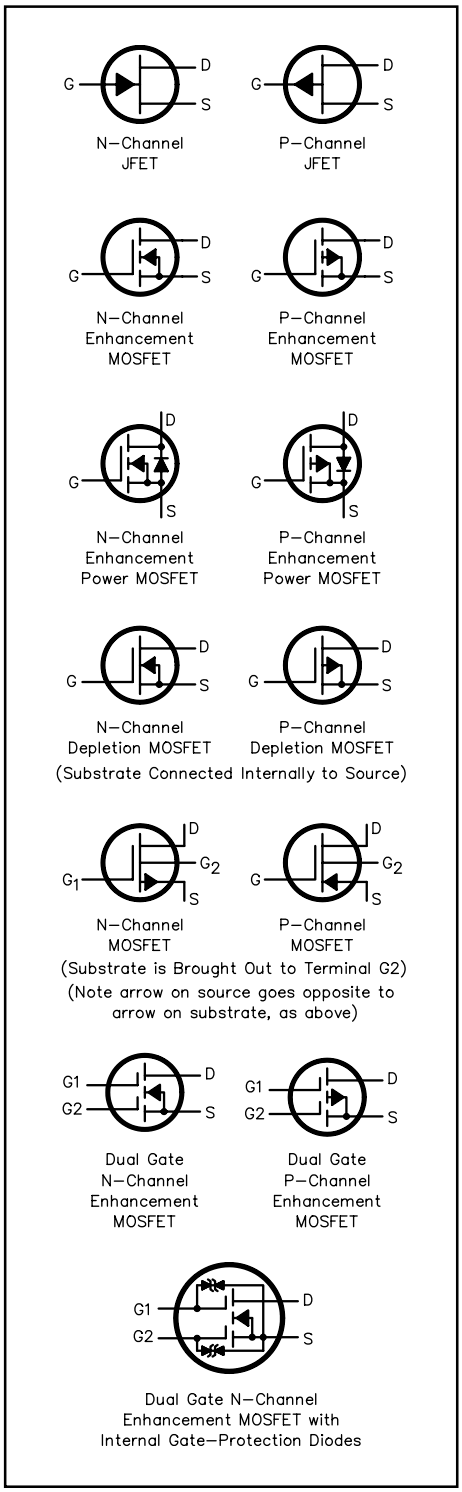

Fig 8.35—FET schematic symbols.

FIELD-EFFECT TRANSISTORS

FET devices are more closely related to vacuum tubes than are bipolar transistors. Both the vacuum tube and the FET are controlled by the voltage level of the input rather than the input current, as in the bipolar transistor. FETs have three basic terminals, the gate, the source and the drain. These are related to both vacuum tube and bipolar transistor terminals: the gate to the grid and the base, the source to the cathode and the emitter, and the drain to the plate and the collector. Different forms of FET devices are pictured in **Fig 8.35**.

The characteristic curves for FETs are similar to those of vacuum tubes. The two most useful relationships are called the transconductance and output curves (**Fig 8.36**). Transconductance curves

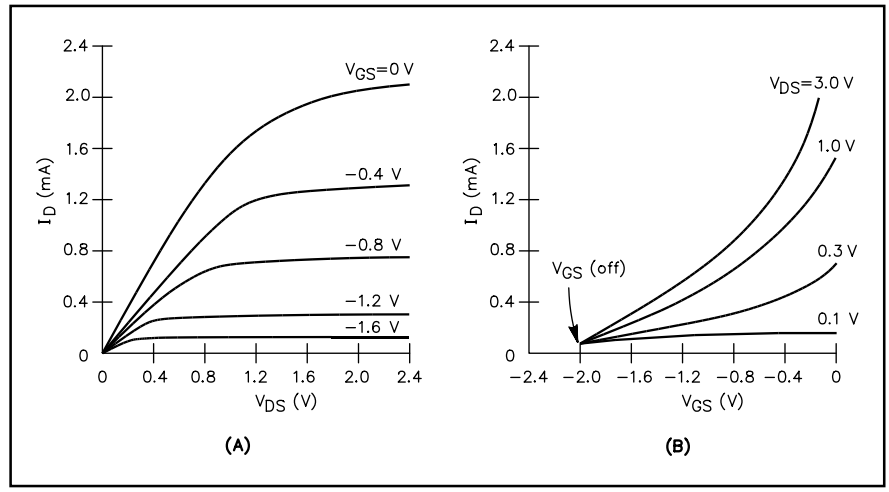

Fig 8.36 — JFET output and transconductance response curves for common source amplifier configuration. (A) Output voltage (V_{DS}) on the x-axis versus output current (I_D) on the y-axis, with different curves plotted for various values of input voltage (V_{GS}). (B) Transconductance curve has the same three variables rearranged, V_{GS} on the x-axis, I_D on the y-axis and curves plotted for different values of V_{DS}.

Transistor Amplifier Design—a Practical Approach

The design of a transistorized amplifier is a straightforward process. Just as you don't need a degree in mechanical engineering to drive an automobile, neither do you need detailed knowledge of semiconductor physics in order to design a transistor amplifier with predictable and repeatable properties.

This sidebar will describe how to design a small-signal "Class A" transistor amplifier, following procedures detailed in one of the best books on the subject—*Solid State Design for the Radio Amateur*, by Wes Hayward, W7ZOI, and Doug DeMaw, W1FB. For many years, both hams and professional engineers have used this classic ARRL book to design untold numbers of working amplifiers.

How Much Gain?

One of the simple, yet profound, observations made in *Solid State Design for the Radio Amateur* is that a designer should *not* attempt to extract every last bit of gain from a single amplifier stage. Trying to do so virtually guarantees that the circuit will be "touchy"—it may end up being more oscillator than amplifier! While engineers might debate the exact number, modern semiconductors circuits are inexpensive enough that you should try for no more than 25 dB of gain in a single stage.

For example, if you are designing a high-gain amplifier system to follow a direct-conversion receiver mixer, you will need a total of about 100 dB of audio amplification. We would recommend a conservative approach where you use four stages, each with 25 dB of gain. But you might risk oscillation and instability by using only two stages, with 50 dB gain each. The component cost will not be greatly different between these approaches, but the headaches and lack of reproducibility of the "simpler" two-stage design will very likely far outweigh any small cost advantages!

Biasing the Transistor Amplifier

The first step in amplifier design is to *bias* the transistor properly. A small-signal linear amplifier is biased properly when there is current at all times. Once you have biased the stage, you can then use several simple rules of thumb to determine all the major properties of the resulting amplifier.

Solid State Design for the Radio Amateur introduces several elegant transistor models. We won't get into that much detail here, except to say that the most fundamental property of a transistor is this: When there is current in the base-emitter junction, a larger current will flow in the collector-emitter junction. When the base-emitter junction is thus *forward biased*, the voltage across the base and emitter leads of a silicon transistor will be relatively constant, at 0.7 V. For most modern transistors, the dc current in the collector-emitter junction will be at least 50 to 100 times greater than the base-emitter current. This dc current gain is called the transistor's *Beta* (β).

See **Fig A**, which shows a simple capacitively coupled low-frequency amplifier suitable for use at 1 MHz. Resistors R1 and R2 form a voltage divider feeding the base of the transistor. The amount of current in the resistive voltage divider is purposely made large enough so the base current is small in comparison, thus creating a "stiff" voltage supply for the base. As stated above, the voltage at the emitter will be 0.7 V less than the base voltage for this NPN transistor. The emitter voltage V_E appears across the series combination of R4 and R5. Note that R5 is bypassed by capacitor C4 for ac current.

By Ohm's Law, the emitter current is equal to the emitter voltage V_E divided by the sum of R4 plus R5. Now, the emitter current is made up of both the base-emitter and the collector-emitter current, but since the base current is much smaller than the collector current, the amount of collector current is essentially equal to the emitter current, at V_E / (R4 + R5).

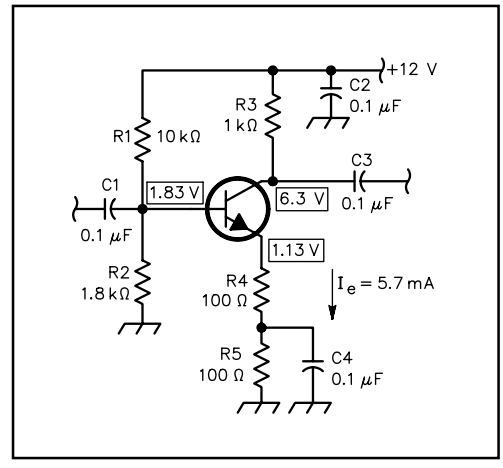

Fig A—Example of a simple low-frequency capacitively coupled transistorized small-signal amplifier. The voltages shown are the preliminary values desired for a collector current of 5 mA. The ac voltage gain is the ratio of the collector load resistor, R3, divided by the unbypassed portion of the emitter resistor, R4.

Our design process starts by specifying the amount of current we want to flow in the collector, with the dc collector voltage equal to half the supply voltage. For good bias stability with temperature variation, the total emitter resistor should be at least 100 Ω for a small-signal amplifier. Let's choose a collector current of 5 mA, and use a total emitter resistance of 200 Ω, with R4 = R5 = 100 Ω each. The voltage across 200 Ω for 5 mA of current is 1.0 V. This means that the voltage at the base must be 1.0 V + 0.7 V = 1.7 V, provided by the voltage divider R1 and R2.

The dc base current requirements for a collector current of 5 mA is approximately 5 mA / 50 = 0.1 mA if the transistor's dc Beta is at least 50, a safe assumption for modern transistors. To provide a "stiff" base voltage, we want the current through the voltage divider to be about five to ten times greater than the base current. For convenience then, we choose the current through R1 to be 1 mA. This is a convenient current value, because the math is simplified—we don't have to worry about decimal points for current or resistance: 1 mA × 1.8 kΩ = 1.8 V. This is very close to the 1.7 V we are seeking. We thus choose a standard value of 1.8 kΩ for R2. The voltage drop across R1 is 12 V – 1.8 V = 10.2 V. With 1 mA in R1, the necessary value is 10.2 kΩ, and we choose the closest standard value, 10 kΩ.

Let's now look at what is happening in the collector part of the circuit. The collector resistor R3 is 1 kΩ, and the 5 mA of collector current creates a 5 V drop across R3. This means that the collector dc voltage must be 12 V – 5 V = 7 V. The dc power dissipated in the transistor will be essentially all in the collector-emitter junction, and will be the collector-emitter voltage (7 V – 1 V = 6 V) times the collector current of 5 mA = 0.030 W, or 30 mW. This dissipation is well within the 0.5 W rating typical of small-signal transistors.

Now, let's calculate more accurately the result from using standard values for R1 and R2. The actual base voltage will be 12 V × [1.8 kΩ / (1.8 kΩ + 10 kΩ)] = 1.83 V, rather than 1.7 V. The resulting emitter voltage is 1.83 V – 0.7 V = 1.13 V, resulting in 1.13 V / 200 Ω = 5.7 mA of collector current, rather than our desired 5 mA. We are close enough—we have finished designing the bias circuitry!

Performance: Voltage Gain

Now we can analyze how our little amplifier will work. The use of the unbypassed emitter resistor R4 results in *emitter degeneration*—a fancy word describing a form of negative feedback. The bottom line for us is that we can use several handy rules of thumb. The first is for the ac voltage gain of an amplifier: A_V = R3 / R4, where A_V is shorthand for *voltage gain.* The ac voltage gain of such an amplifier is simply the ratio of the collector load resistor and the unbypassed emitter resistor. In this case, the gain is 1000 / 100 = 10, which is 20 dB of voltage gain. This expression for gain is true virtually without regard for the exact kind of transistor used in the circuit, provided that we design for moderate gain in a single stage, as we have done.

Performance: Input Resistance

Another useful rule of thumb stemming from use of an unbypassed emitter resistor is the expression for the ac input resistance: R_{in} = Beta × R4. If the ac Beta at low frequencies is about 50, then the input resistance of the transistor is 50 × 100 W = 5000 Ω. The actual input resistance includes the shunt resistance of voltage divider R2 and R1, about 1.5 kΩ. Thus the biasing resistive voltage divider essentially sets the input resistance of the amplifier.

Performance: Overload

We can accurately predict how this amplifier will perform. If we were to supply a peak positive 1 V signal to the base, the voltage at the collector will try to fall by the voltage gain of 10. However, since the dc voltage at the collector is only 7 V, it is clear that the collector voltage cannot fall 10 V. In theory, the collector voltage could fall as low as the 1.13 V dc level at the emitter. This amplifier will "run out of voltage" at a negative collector voltage swing of about 6.3 V – 1.13 V = 5.17 V, when the input voltage is 5.17 divided by the gain of 10 = 0.517 V.

When a negative-going ac voltage is supplied to the base, the collector current falls, and the collector voltage will rise by the voltage gain of 10. The maximum amount of voltage possible is the 12 V supply voltage, where the transistor is cut off with no collector current. The maximum positive collector swing is from the standing collector dc voltage to the supply voltage: 12 V – 6.3 V = 5.7 V positive

swing. This occurs with a peak negative input voltage of 5.7 V / 10 = 0.57 V. Our amplifier will overload rather symmetrically on both negative and positive peaks. This is no accident—we biased it to have a collector voltage halfway between ground and the supply voltage.

When the amplifier "runs out of output voltage" in either direction, another useful rule of thumb is that this is the *1 dB compression point*. This is where the amplifier just begins to depart from linearity, where it can no longer provide any more output for further input. For our amplifier, this is with a peak-to-peak output swing of approximately 5.1 V $\times$ 2 = 10.2 V, or 3.6 V rms. The output power developed in output resistor R3 is $(3.6)^2$ / 1000 = 0.013 W = 13 mW, which is +11.1 dBm (referenced to 1 mW on 50 Ω).

At the 1 dB compression point, the third-order *IMD* (intermodulation distortion) will be roughly 25 dB below the level of each tone. **Fig B** shows a graph of output versus input levels for both the desired signal and for third-order IMD products. The rule of thumb for IMD is that if the input level is decreased by 10 dB, the IMD will decrease by 30 dB. Thus, if input is restricted to be 10 dB below the 1 dB compression point, the IMD will be 25 dB + 30 dB = 55 dB below each output tone.

With very simple math we have thus designed and characterized a simple amplifier. This amplifier will be stable for both dc and ac under almost any thermal and environmental conditions conceivable. That wasn't too difficult, was it?—*R. Dean Straw, N6BV, ARRL Senior Assistant Technical Editor*

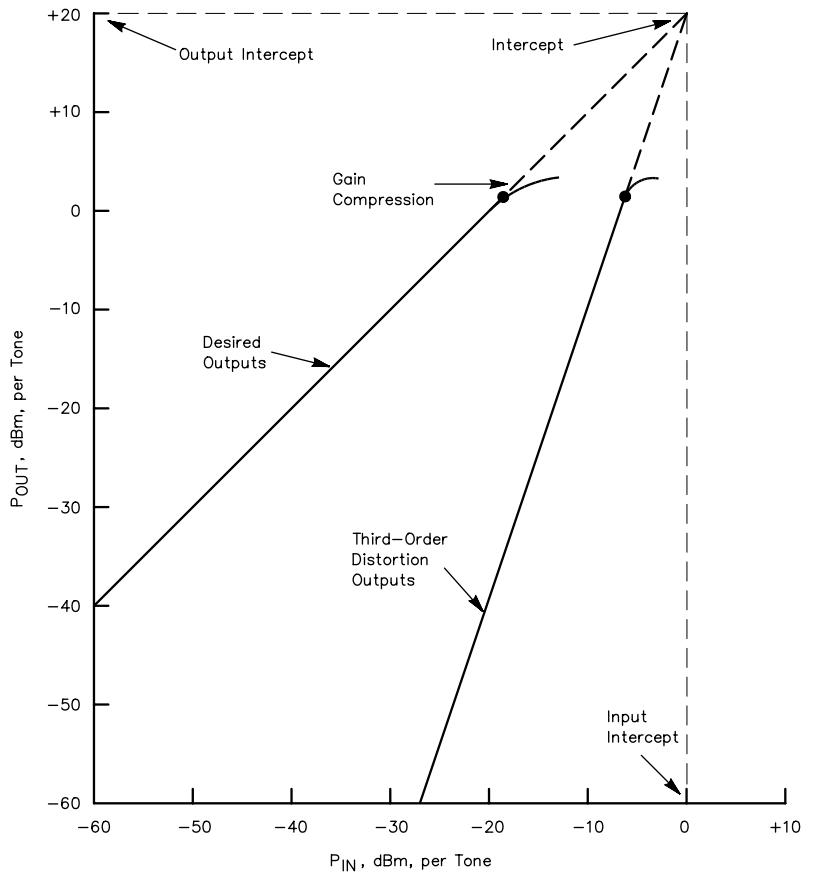

Fig B—Output IMD (intermodulation distortion) as a function of input. In the region below the 1 dB compression point, a decrease in input level of 10 dB results in a drop of IMD products by 30 dB below the level of each output tone in a two-tone signal.

give the drain current, I_D, due to different gate-source voltage differences, V_{GS}, for various drain-source voltages, V_{DS}. The same parameters are interrelated in a different way in the output curve. For different values of V_{GS}, I_D is plotted against V_{DS}. In both of these representations, the device output is the drain current and these curves describe the FET in the common-source configuration. The action of the FET channel is so nearly ideal that, as long as the JFET gate does not become forward biased, the drain and source currents are virtually identical. For JFETs the gate leakage current, I_G, is a function of V_{GS} and this is often expressed with an input curve (**Fig 8.37**). The point at which there is a great increase in I_G is called the *junction break point voltage*. The insulated gates in MOSFET devices do away with any appreciable gate leakage current. MOSFETs do not need input and reverse transconductance curves. Their output curves (**Fig 8.38**) are similar to those of the JFET.

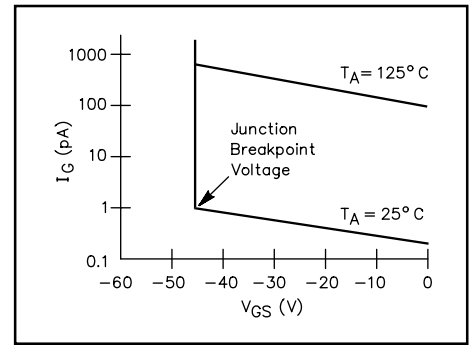

Fig 8.37 — JFET input leakage curves for common source amplifier configuration. Input voltage (V_{GS}) on the x-axis versus input current (I_G) on the y-axis, with two curves plotted for different operating temperatures, 25°C and 125°C. Input current increases greatly when the gate voltage exceeds the junction breakpoint voltage.

The parameters used to describe a FET's performance are also similar to those of vacuum tubes. The dc channel resistance, r_{DS}, is specified in data sheets to be less than a maximum value when the device is biased on ($r_{DS(on)}$). For ac signals, $r_{ds(on)}$ is not necessarily the same as $r_{DS(on)}$, but it is not very different as long as the frequency is not so high that capacitive reactance becomes significant. The common source forward transconductance, g_{fs}, is obtained as the slope of one of the lines in the forward transconductance curve,

$$g_{fs}=\frac{\Delta I_D}{\Delta V_{GS}} \qquad (14)$$

When gate voltage is maximum (V_{GS} = 0 for a JFET) $r_{DS(on)}$ is minimum. This describes the effectiveness of the device as an analog switch.

A typical FET data sheet gives ABSOLUTE MAXIMUM RATINGS for V_{DS}, V_{DG}, V_{GS} and I_D, along with the usual device dissipation (P_D) and storage temperature range. The OFF CHARACTERISTICS listed are the gate-source breakdown voltage, $V_{GS(BR)}$, the reverse gate

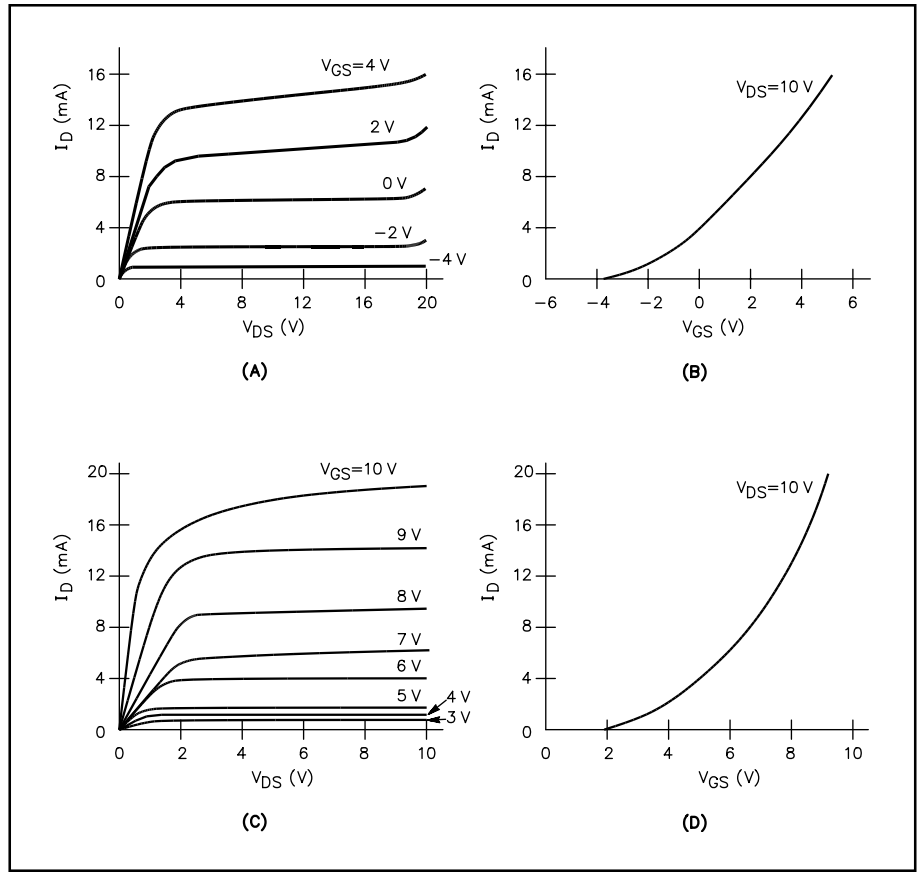

Fig 8.38 — MOSFET output [(A) and (C)] and transconductance [(B) and (D)] response curves. Plots (A) and (B) are for an N-channel depletion mode device. Note that V_{GS} varies from negative to positive values. Plots (C) and (D) are for an N-channel enhancement mode device. V_{GS} has only positive values.

current, I_{GSS} and the gate-source cutoff voltage, $V_{GS(OFF)}$. The ON CHARACTERISTIC is the zero-gate-voltage drain current (I_{DSS}). The SMALL SIGNAL CHARACTERISTICS include the forward transfer admittance, y_{fs}, the output admittance, y_{os}, the static drain-source on resistance, $r_{ds(on)}$ and various capacitances such as input capacitance, C_{iss}, reverse transfer capacitance, C_{rss}, the drain-substrate capacitance, $C_{d(sub)}$. FUNCTIONAL CHARACTERISTICS include the noise figure, NF and the common source power gain G_{ps}.

The relatively flat regions in the MOSFET output curves are often used to provide a constant current source. As is plotted in these curves, the drain current, I_D, changes very little as the drain-source voltage, V_{DS}, varies in this portion of the curve. Thus, for a fixed gate-source voltage, V_{GS}, the drain current can be considered to be constant over a wide range of drain-source voltages.

Multiple gate MOSFETs are also available (MFE130, MPF201, MPF211, MPF521). Due to the insulating layer, the two gates are isolated from each other and allow two signals to control the channel simultaneously with virtually no loading of one signal by the other. A common application of this type of device is an automatic gain control (AGC) amplifier. The signal is applied to one gate and a rectified, low-pass filtered form of the output (the AGC voltage) is fed back to the other gate. Another common application is for mixers.

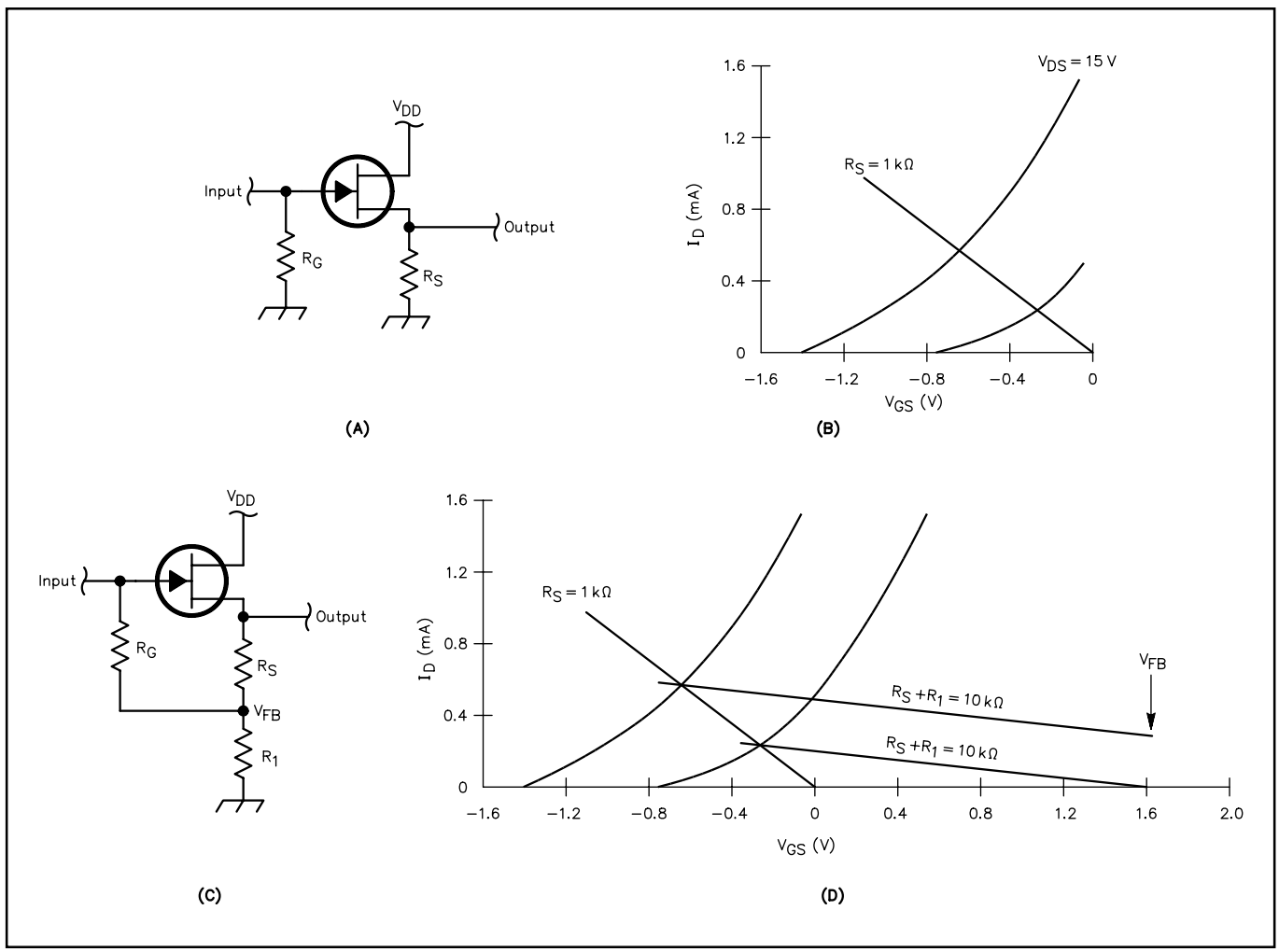

Fig 8.39 — FET biasing circuits. (A) Self biased common drain JFET circuit. (B) Transconductance curve for self biased JFET in (A). Gate bias is determined by current through R_S. Load line has a slope of – 1 / R_S and gate bias voltage can vary between where the load line crosses the characteristic curves. (C) Feedback bias common drain JFET circuit.

FET Biasing

There are two ways to bias an FET, with and without feedback. Source self biasing for an N-channel JFET is pictured in **Fig 8.39A**. In this common-drain amplifier circuit, bias level is determined by the current through R_S, since I_G is very small and there is essentially no voltage drop across R_G. The characteristic curve for this configuration is plotted in Fig 8.39B. The operating points of the amplifier are where the load line intersects the curves. An example of feedback biasing is shown in Fig 8.39C. R_1 is generally much larger than R_S and the load line is determined by the sum of these resistors, as shown in Fig 8.39D. Feedback biasing increases the input impedance of the amplifier, but is rarely required, since input resistance (R_G) can be made very large.

MOSFET Gate Protection

The MOSFET is constructed with a very thin layer of SiO_2 for the gate insulator. This layer is extremely thin in order to improve the gain of the device but this makes it susceptible to damage from high voltage levels. If enough charge accumulates on the gate terminal, it can punch through the gate insulator and destroy it. The insulation of the gate terminal is so good that virtually none of this potential is eased by leakage of the charge into the device. While this condition makes for nearly ideal input impedance (approaching infinity), it puts the device at risk of destruction from even such seemingly innocuous electrical sources as static electricity in the air.

Some MOSFET devices contain an internal Zener diode with its cathode connected to the gate and its anode to the substrate. If the voltage at the gate rises to a damaging level the Zener junction breaks down and bleeds the excess charges off to the substrate. When voltages are within normal operating limits the Zener has little effect on the signal at the gate, although it may decrease the input impedance of the MOSFET. This solution will not work for all MOSFETs. The Zener diode must always be reverse biased to be effective. In the enhancement mode MOSFET, $V_{GS} > 0$ for all valid uses of the part. In depletion mode devices V_{GS} can be both positive and negative; when negative, a protection Zener diode would be forward biased and the MOSFET would not work properly. In some depletion mode MOSFET devices, back-to-back Zener diodes are used to protect the gate.

MOSFET devices are at greatest risk of damage from static electricity when they are out of circuit. Even though static electricity is capable of delivering little current, it can generate thousands of volts. When storing MOSFETs, the leads should be placed into conductive foam. When working with MOSFETs, it is a good idea to minimize static by wearing a grounded wrist strap and working on a grounded table. A humidifier may help to decrease the static electricity in the air. Before inserting a MOSFET into a circuit board it helps to first touch the device leads with your hand and then touch the circuit board. This serves to equalize the excess charge so that when the device is inserted into the circuit board little charge will flow into the gate terminal.

OPTICAL SEMICONDUCTORS

In addition to electrical energy and heat energy, light energy also affects the behavior of semiconductor materials. If a device is made to allow light to fall on the surface of the semiconductor material, the light energy will break covalent bonds and increase the number of electron-hole pairs, decreasing the resistance of the material.

Photoconductors

In commercial *photoconductors* (also called *photoresistors*) the resistance can change by as much as several kilohms for a light intensity change of 100 ft-candles. The most common material used in photoconductors is cadmium sulfide (CdS), with a resistance range of more than 2 MΩ in total darkness to less than 10 Ω in bright light. Other materials used in photoconductors respond best at specific colors.

Lead sulfide (PbS) is most sensitive to infrared light and selenium (Se) works best in the blue end of the visible spectrum.

A similar effect is used in some diodes and transistors so that their operation can be controlled by light instead of electrical current biasing. These devices are called *photodiodes* and *phototransistors*. The flow of minority carriers across the reverse biased PN junction is increased by light falling on the doped semiconductor material. In the dark, the junction acts the same as any reverse biased PN junction, with a very low current (on the order of 10 μA) that is nearly independent of reverse voltage. The presence of light not only increases the current but also provides a resistance-like relationship (reverse current increases as reverse voltage increases). See **Fig 8.40** for the characteristic response of a photodiode. Even with no reverse voltage applied, the presence of light causes a small reverse current, as indicated by the points at which the lines in Fig 8.40 intersect the left side of the graph. Photoconductors and photodiodes are generally used to produce light-related analog signals that require further processing. The phototransistor can often be used to serve both purposes, acting as an amplifier whose gain varies with the amount of light present. It is also more sensitive to light than the other devices. Phototransistors have lots of gain, but photodiodes normally have less noise, so they make sensitive detectors.

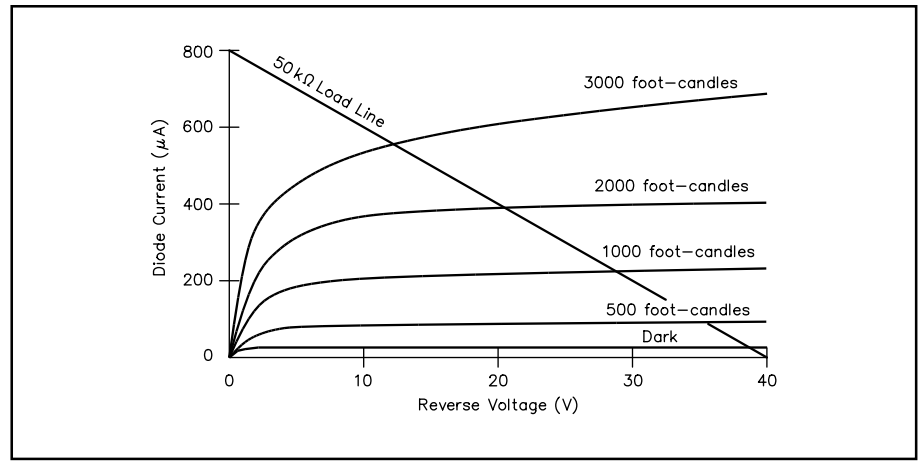

Fig 8.40 — Photodiode response curve. Reverse voltage is plotted on the x-axis and current through diode is plotted on the y-axis. Various response lines are plotted for different illumination. Except for the zero illumination line, the response does not pass through the origin since there is current generated at the PN junction by the light energy. A load line is shown for a 50 kΩ resistor in series with the photodiode.

Photovoltaic Effect

When illuminated, the reverse biased photodiode has a reverse current due to excess minority carriers. As the reverse voltage is reduced, the potential barrier to the forward flow of majority carriers is also reduced. Since light energy leads to the generation of both majority and minority carriers, when the resistance to the flow of majority carriers is decreased these carriers form a forward current. The voltage at which the forward current equals the reverse current is called the *photovoltaic potential* of the junction. If the illuminated PN junction is not connected to a load, a voltage equal to the photovoltaic potential can be measured across it. Devices that use light from the sun to produce electricity in this way are called *solar cells* or *solar batteries*. Common operating characteristics of silicon photovoltaic cells are an open circuit voltage of about 0.6 V and a conversion efficiency of about 10 to 15%.

Light Emitting Diodes

In the photodiode, energy from light falling on the semiconductor material is absorbed to make additional electron-hole pairs. When the electrons and holes recombine, the same amount of energy is given off. In normal diodes the energy from recombination of carriers is given off as heat. In certain forms of semiconductor material, the recombination energy is given off as light with a mechanism called *electroluminescence*. Unlike the incandescent light bulb, electroluminescence is a cold light source that typically operates with low voltages and currents (such as 1.5 V and 10 mA). Devices made for this

purpose are called *light emitting diodes (LEDs)*. They have the advantages of low power requirements, fast switching times (on the order of 10 ns) and narrow spectra (relatively pure color). The LED emits light when it is forward biased and excess carriers are present. As the carriers recombine, light is produced with a color that depends on the properties of the semiconductor material used. Gallium arsenide (GaAs) generates light in the infrared region, gallium phosphide (GaP) gives off red light when doped with oxygen or green light when doped with nitrogen. Orange light is attained with a mixture of GaAs and GaP (GaAsP). Silicon doped with carbon gives off yellow light but does not produce much illumination. Other colors are also possible with different types and concentrations of dopants but usually have lower illumination efficiencies.

The LED is very simple to use. It is connected across a voltage source with a series resistor that limits the current to the desired level for the amount of light to be generated.The cathode lead is connected to the lower potential, and is usually specially marked (flattening of the lead near the package, a dot of paint next to the lead, and a flat portion of the round device located next to the lead are all common methods).

Optoisolators

An interesting combination of optoelectronic components proves very useful in many analog signal processing applications. An *optoisolator* consists of an LED optically coupled to a phototransistor, usually in an enclosed package. The optoisolator, as its name suggests, isolates different circuits from each other. Typically, isolation resistance is on the order of $10^{11}\ \Omega$ and isolation capacitance is less than 1 pF. Maximum voltage isolation varies from 1,000 to 10,000 V ac. The most common optoisolators are available in 6 pin DIP packages.

Optoisolators are used for voltage level shifting and signal isolation. The isolation has two purposes: to protect circuitry from excessive voltage spikes and to isolate noisy circuitry from noise sensitive circuitry. A disadvantage of an optoisolator is that it adds a finite amount of noise and is not appropriate for use in many applications with low level signals. Optoisolators also cannot transfer signals with high power levels. The power rating of the LED in a 4N25 device is 120 mW. Optoisolators have a limited frequency response due to the high capacitance of the LED. A typical bandwidth for the 4N25 series is 300 kHz.

As an example of voltage level shifting, the input to an optoisolator can be derived from a tube amplifier that has a signal varying between 0 and 150 V by using a series current limiting resistor. In order to drive a semiconductor circuit that operates in the –1 to 0 V range, the output of the optoisolator can be biased to operate in that range. This conversion of voltage levels, without a common ground connection between the circuits, is not easily performed in any other way.

A 1000 V spike that is high enough to destroy a semiconductor circuit will only saturate the LED in the optoisolator and will not propagate to the next stage. The worst that will happen is the LED will be destroyed, but very often it is capable of surviving even very high voltage spikes.

Optoisolators are also useful for isolating different ground systems. The input and output signals are totally isolated from each other, even with respect to the references for each signal. A common application for optoisolation is when a computer is used to control radio equipment. The computer signal, and even its ground reference, typically contains considerable wide band noise due to the digital circuitry. The best way to keep this noise out of the radio is to isolate both the signal and its reference; this is easily done with an optoisolator.

The design of circuits with optoisolators is not different from the design of circuits with LEDs and with transistors. The LED is forward biased and usually driven with a series current limiting resistor whose value is set so that the forward current will be less than the maximum value for the device (such as 60 mA in a 4N25). Signals must be appropriately dc shifted so that the LED is always forward biased. The phototransistor typically has all three leads available for connection. The base lead is used for biasing, since the signal is usually derived from the optics, and the collector and emitter leads are used as they would be in any transistor amplifier circuit.

Fiber Optics

An interesting variation of the optoisolator is the *fiber optic* connection. Like the optoisolator, the signal is introduced to an LED device that modulates light. The signal is recovered by a photodetecting device (photoresistor, photodiode, or phototransistor). Instead of locating the input and output devices next to each other, the light is transmitted in a fiber optic cable, an extruded glass fiber that efficiently carries light over long distances and around fairly sharp bends. The fiber optic cable isolates the two circuits and provides an interesting transmission line. Fiber optics generally have far less loss than coaxial cable transmission lines. They do not leak RF energy, nor do they pick up electrical noise. Special forms of LEDs and phototransistors are available with the appropriate optical couplers for connecting to fiber optic cables. These devices are typically designed for higher frequency operation with bandwidths in the tens and hundreds of megahertz.

LINEAR INTEGRATED CIRCUITS

If you look into a transistor, the actual size of the semiconductor is quite small compared to the size of the packaging. For most semiconductors, the packaging takes considerably more space than the actual semiconductor device. Thus, an obvious way to reduce the physical size of circuitry is to combine more of the circuit inside a single package.

Hybrid Integrated Circuits

It is easy to imagine placing several small semiconductor chips in the same package. This is known as *hybrid circuitry*, a technology in which several semiconductor chips are placed in the same package and miniature wires are connected between them to make complete circuits.

Hybrid circuits miniaturize analog electronic circuits by replacing much of the packaging that is inherent in discrete electronics. The term *discrete* refers to the use of individual components to make a circuit, each in its own package. One application that still exists for hybrid circuitry is microwave amplifiers. The components of the amplifier are placed in a standard TO-39 package that is only 1 cm in diameter. The small dimensions of these circuits permit operation at VHF. For example, the Motorola MWA5157 can provide over 23 dB of gain at 1 GHz.

Both discrete and hybrid circuitry require that connections be made between the leads of the components. This takes space, is relatively expensive to construct and is the source of most failures in electronic circuitry. If multiple components could be placed on a single piece of semiconductor with the connections between them as part of the semiconductor chip, these three disadvantages would be overcome.

Monolithic Integrated Circuits

In order to build entire circuits on a single piece of semiconductor, it must be possible to fabricate other devices, such as resistors and capacitors, as well as transistors and diodes. The entire circuit is combined into a single unit, or chip, that is called a *monolithic integrated circuit*.

An integrated circuit (IC) is fabricated in layers. An example of a semiconductor circuit schematic and its implementation in an IC is pictured in **Fig 8.41**. The base layer of the circuit, the *substrate*, is made of P-type semiconductor material. Although less common, the polarity of the substrate can also be N-type material. Since the mobility of electrons is about three times higher than that of holes, bipolar transistors made with N-type collectors and FETs made with N-type channels are capable of higher speeds and power handling. Thus, P-type substrates are far more common. For devices with N-type substrates, all polarities in the ensuing discussion would be reversed. Other substrates have been used, one of the most successful of which is the silicon-on-sapphire (SOS) construction that has been used to increase the bandwidth of integrated circuitry. Its relatively high manufacturing cost has impeded its use, however.

On top of the P-type substrate is a thin layer of N-type material in which the active and passive

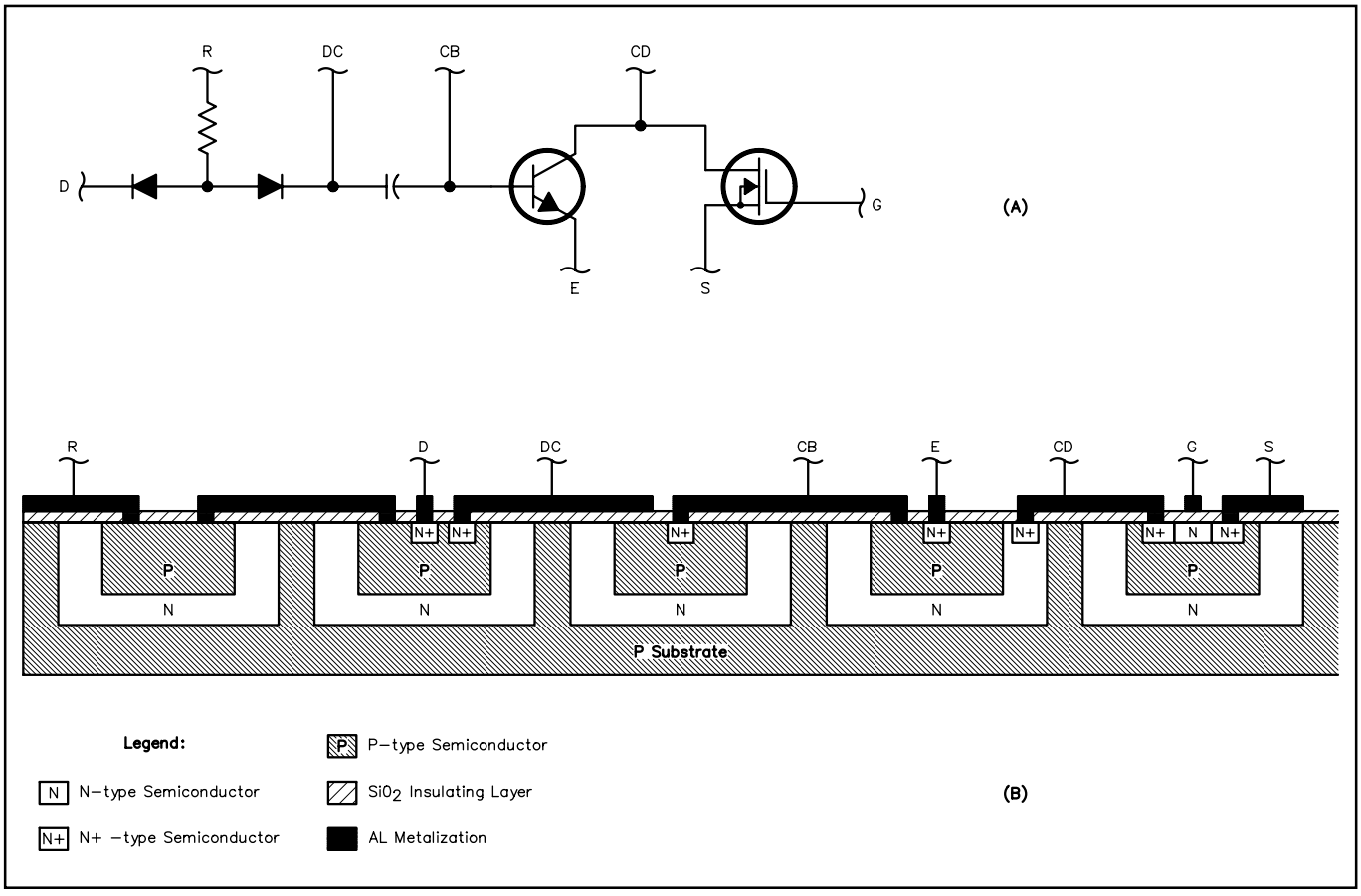

Fig 8.41 — Integrated circuit layout. (A) Circuit containing two diodes, a resistor, a capacitor, an NPN transistor and an N-channel MOSFET. Labeled leads are D for diode, R for resistor, DC for diode-capacitor, E for emitter, S for source, CD for collector-drain and G for gate. (B) Integrated circuit that is identical to circuit in (A). Same leads are labeled for comparison. Circuit is built on a P-type semiconductor substrate with N-type wells diffused into it. An insulating layer of SiO_2 is above the semiconductor and is etched away where aluminum metal contacts are made with the semiconductor. Most metal-to-semiconductor contacts are made with heavily doped N-type material (N^+-type semiconductor).

components are built. Impurities are diffused into this layer to form the appropriate component at each location. To prevent random diffusion of impurities into the N-layer, its upper surface must be protected. This is done by covering the N-layer with a layer of silicon dioxide (SiO_2). Wherever diffusion of impurities is desired, the SiO_2 is etched away. The precision of placing the components on the semiconductor material depends mainly on the fineness of the etching. The fourth layer of an IC is made of metal (usually aluminum) and is used to make the interconnections between the components.

Different components are made in a single piece of semiconductor material by first diffusing a high concentration of acceptor impurities into the layer of N-type material. This process creates P-type semiconductor—often referred to as P^+-type semiconductor because of its high concentration of acceptor atoms—that isolates regions of N-type material. Each of these regions is then further processed to form single components. A component is produced by the diffusion of a lesser concentration of acceptor atoms into the middle of each isolation region. This results in an N-type *isolation well* that contains P-type material, is surrounded on its sides by P^+-type material and has P-type material (substrate) below it. The cross sectional view in Fig 8.41B illustrates the various layers. Contacts to the metal layer are often made by diffusing high concentrations of donor atoms into small regions of the N-type well and

the P-type material in the well. The material in these small regions is N^+-type and facilitates electron flow between the metal contact and the semiconductor. In some configurations, it is necessary to connect the metal directly to the P-type material in the well.

An isolation well can be made into a resistor by making two contacts into the P-type semiconductor in the well. Resistance is inversely proportional to the cross-sectional area of the well. An alternate type of resistor that can be integrated in a semiconductor circuit is a *thin film resistor*, where a metallic film is deposited on the SiO_2 layer, masked on its upper surface by more SiO_2 and then etched to make the desired geometry, thus adjusting the resistance.

There are two ways to form capacitors in a semiconductor. One is to make use of the PN junction between the N-type well and the P-type material that fills it. Much like a varactor diode, when this junction is reverse biased a capacitance results. Since a bias voltage is required, this type of capacitor is polarized, like an electrolytic capacitor. Nonpolarized capacitors can also be formed in an integrated circuit by using thin film technology. In this case, a very high concentration of donor ions is diffused into the well, creating an N^+-type region. A thin metallic film is deposited over the SiO_2 layer covering the well and the capacitance is created between the metallic film and the well. The value of the capacitance is adjusted by varying the thickness of the SiO_2 layer and the cross-sectional size of the well. This type of thin film capacitor is also known as a metal oxide semiconductor (MOS) capacitor.

Unlike resistors and capacitors, it is very difficult to create inductors in integrated circuits. Generally, RF circuits that need inductance require external inductors to be connected to the IC. In some cases, particularly at lower frequencies, the behavior of an inductor can be mimicked by an amplifier circuit. In many cases the appropriate design of IC amplifiers can obviate the need for external inductors.

Transistors are created in integrated circuitry in much the same way that they are fabricated in their discrete forms. The NPN transistor is the easiest to make since the wall of the well, made of N-type semiconductor, forms the collector, the P-type material in the well forms the base and a small region of N^+-type material formed in the center of the well becomes the emitter. A PNP transistor is made by diffusing donor ions into the P-type semiconductor in the well to make a pattern with P-type material in the center (emitter) surrounded by a ring of N-type material that connects all the way down to the well material (base), and this is surrounded by another ring of P-type material (collector). This configuration results in a large base width separating the emitter and collector, causing these devices to have much lower current gain than the NPN form. This is one reason why integrated circuitry is designed to use many more NPN transistors than PNP transistors.

The simplest form of diode is generated by connecting to an N^+-type connection point in the well for the cathode and to the P-type well material for the anode. Diodes are often converted from NPN transistor configurations. Integrated circuit diodes made this way can either short the collector to the base or leave the collector unconnected. The base contact is the anode and the emitter contact is the cathode.

FETs can also be fabricated in IC form. Due to its many functional advantages, the MOSFET is the most common form used for digital ICs. MOSFETs are made in a semiconductor chip much the same way as MOS capacitors, described earlier. In addition to the signal processing advantages offered by MOSFETs over other transistors, the MOSFET device can be fabricated in 5% of the physical space required for bipolar transistors. MOSFET ICs can contain 20 times more circuitry than bipolar ICs with the same chip size. Just as discrete MOSFETs are at risk of gate destruction, IC chips made with MOSFET devices have a similar risk. They should be treated with the same care to protect them from static electricity as discrete MOSFETs. Integrated circuits need not be made exclusively with MOSFETs or bipolar transistors. It is common to find IC chips designed with both technologies, taking advantage of the strengths of each.

Complementary Metal Oxide Semiconductors

Power dissipation in a circuit can be reduced to very small levels (on the order of a few nW) by using

the MOSFET devices in complementary pairs (CMOS). Each amplifier is constructed of a series circuit of MOSFET devices, as in **Fig 8.42**. The gates are tied together for the input signal, as are the drains for the output signal. In saturation and cutoff, only one of the devices conducts. The current drawn by the circuit under no load is equal to the OFF leakage current of either device and the voltage drop across the pair is equal to V_{DD}, so the steady state power used by the circuit is always equal to $V_{DD} \times I_{D(off)}$. For ac signals, power consumption is proportional to frequency.

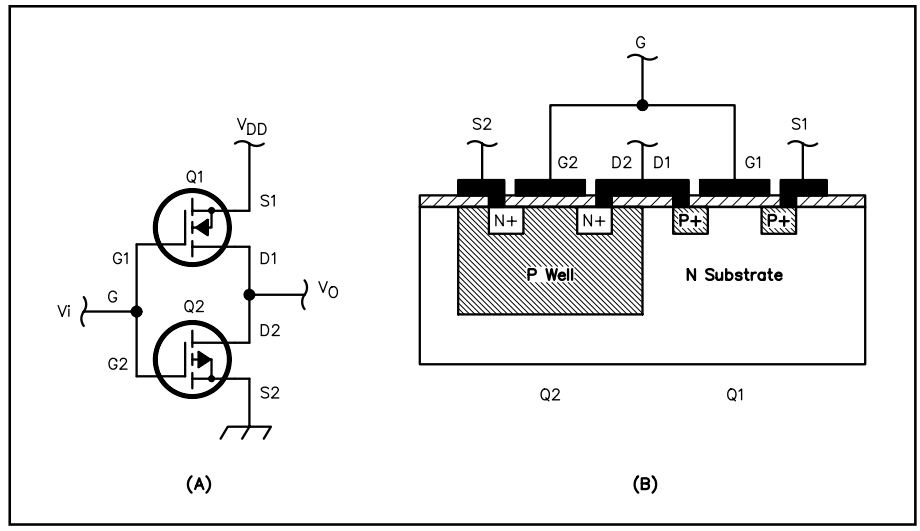

Fig 8.42 — Complementary metal oxide semiconductor (CMOS). (A) CMOS device is made from a pair of enhancement mode MOS transistors, the upper is an N-channel device and the lower is a P-channel device. When one transistor is biased on, the other is biased off so there is minimal current from V_{DD} to ground. (B) Implementation of a CMOS pair as an integrated circuit.

CMOS circuitry could be built with discrete components; however, the number of extra parts and the need for the complementary components to be matched has made it an unusual design technique. Although CMOS is most commonly used in digital integrated circuitry, its low power consumption has been put to advantage by several manufacturers of analog ICs.

Integrated Circuit Advantages

There are many advantages of monolithic integrated circuitry over similar circuitry implemented with discrete components. The integration of the interconnections is one that has already been mentioned. This procedure alone serves to greatly decrease the physical size of the circuit and to improve its reliability. In fact, in one study performed on failures of electronic circuitry, it was found that the failure rate is not necessarily related to the complexity of the circuit, as had been previously thought, but is more closely a function of the number of interconnections between packages. Thus, the more circuitry that can be integrated onto a single piece of semiconductor material, the more reliable the circuit should be.

The amount of circuitry that can be placed onto a single semiconductor chip is a function of two factors: the size of the chip and how closely the various components are spaced. A revolution in IC manufacture occurred when semiconductor material was created in the laboratory rather than found in nature. The man-made semiconductor wafers are more pure and allow for larger wafer sizes. This, along with the steady improvement of the etching resolution on the chips, has caused an exponential increase over the past two decades in the amount of circuitry that can be placed in a single IC package. Currently, it is not unusual to find chips with more than one million transistors on them.

Decreased circuit size and improved reliability are only two of the advantages of monolithic integrated circuitry. The uncertainty of the exact behavior of the integrated components is the same as it is for discrete components, as discussed earlier. The relative properties of the devices on a single chip are very predictable, however. Since adjacent components on a semiconductor chip are made simultaneously (the entire N-type layer is grown at once, a single diffusion pass isolates all the wells and another pass fills them), the characteristics of identically formed components on a single chip of silicon should be identical. Even if the exact characteristics of the components are unknown, very often in analog circuit design

the major concern is how components interact. For instance, push-pull amplifiers require perfectly matched transistors, and the gain of many amplifier configurations is governed by the ratio between two resistors and not their absolute values of resistance.

Integrated circuits often have an advantage over discrete circuits in their temperature behavior. The variation of performance of the components on an integrated circuit due to heat is no better than that of discrete components. While a discrete circuit may be exposed to a wide range of temperature changes, the entire semiconductor chip generally changes temperature by the same amount; there are fewer "hot spots" and "cold spots." Thus, integrated circuits can be designed to better compensate for temperature changes.

A designer of analog devices implemented with integrated circuitry has more freedom to include additional components that could improve the stability and performance of the implementation. The inclusion of components that could cause a prohibitive increase in the size, cost or complexity of a discrete circuit would have very little effect on any of these factors in an integrated circuit.

Once an integrated circuit is designed and laid out, the cost of making copies of it is very small, often only pennies per chip. Integrated circuitry is responsible for the incredible increase in performance with a corresponding decrease in price of electronics over the last 20 years. While this trend is most obvious in digital computers, analog circuitry has also benefited from this technology.

The advent of integrated circuitry has also improved the design of high frequency circuitry. One problem in the design and layout of RF equipment is the radiation and reception of spurious signals. As frequencies increase and wavelengths approach the dimensions of the wires in a circuit board, the interconnections act as efficient antennas. The dimensions of the circuitry within an IC are orders of magnitude smaller than in discrete circuitry, thus greatly decreasing this problem and permitting the processing of much higher frequencies with fewer problems of interstage interference. Another related advantage of the smaller interconnections in an IC is the lower inherent inductance of the wires, and lower stray capacitance between components and traces.

Integrated Circuit Disadvantages

Despite the many advantages of integrated circuitry, disadvantages also exist. ICs have not replaced discrete components, even tubes, in some applications. There are some tasks that ICs cannot perform, even though the list of these continues to decrease over time as IC technology improves.

Although the high concentration of components on an IC chip is considered to be an advantage of that technology, it also leads to a major limitation. Heat generated in the individual components on the IC chip is often difficult to dissipate. Since there are so many heat generating components so close together, the heat can build up and destroy the circuitry. It is this limitation that currently causes many power amplifiers to be designed with discrete components.

Integrated circuits, despite their short interconnection lengths and lower stray inductance, do not have as high a frequency response as similar circuits built with appropriate discrete components. (There are exceptions to this generalization, of course. Monolithic microwave integrated circuits—MMICs—are available for operation on frequencies up through 10 GHz.) The physical architecture of an integrated circuit is the cause of this limitation. Since the substrate and the walls of the isolation wells are made of opposite types of semiconductor material, the PN junction between them must be reverse biased to prevent current from passing into the substrate. Like any other reverse biased PN junction, a capacitance is created at the junction and this limits the frequency response of the devices on the IC. This situation has improved over the years as isolation wells have gotten smaller, thus decreasing the capacitance between the well and the substrate, and techniques have been developed to decrease the PN junction capacitance at the substrate. One such technique has been to create an N^+-type layer between the well and the substrate, which decreases the capacitance of the PN junction as seen by the well. As an example, in the 1970s the LM324 operational amplifier IC package was developed by National Semiconductor and

claimed a gain-bandwidth product of 1 MHz. In the 1990s the HFA1102 operational amplifier IC, developed by Harris Semiconductor, was introduced with a gain-bandwidth product of 600 MHz.

A major impediment to the introduction of new integrated circuits, particularly with special applications, is the very high cost of development of new designs. The masking cost alone for a designed and tested integrated circuit can exceed $100,000. Adding the design, layout and debugging costs motivates IC manufacturers to produce devices that will be widely used so that they can recoup the development costs by volume of sales. While a particular application would benefit from customization of circuitry on an IC, the popularity of that application may not be wide enough to compel an IC manufacturer to develop that design. A designer who wishes to use IC chips must often settle for circuits that do not behave exactly as desired for the specific application. This trade-off between the advantages afforded by the use of integrated circuitry and the loss of performance if the available IC products do not exactly meet the desired specifications must be considered by equipment designers. It often leads to the use of discrete circuitry in sensitive applications. Once again, the improvements afforded by technology have mitigated this problem somewhat. The design and layout of ICs has been made more affordable by computer-based aids. Interaction between the computer aided design (CAD) software and modern chip masking hardware has also decreased the masking costs. As these development costs decrease, we are seeing an increase in the number of specialty chips that are being marketed and also of small companies that are created to fill the needs of the niche markets.

Common Types of Linear Integrated Circuits

The three main advantages of designing a circuit into an IC are to take advantage of the matched characteristics of like components, to make highly complex circuitry more economical, and to miniaturize the circuit. As a particular technology becomes popular, a rash of integrated circuitry is developed to service that technology. A recent example is the cellular telephone industry. Cellular phones have become so pervasive that IC manufacturers have developed a large number of devices targeted toward this technology. Space limitations prohibit a comprehensive listing of all analog special function ICs but a sampling of those that are more useful in the radio field is presented.

Component Arrays

The most basic form of linear integrated circuit is the component array. The most common of these are the resistor, diode and transistor arrays. Though capacitor arrays are also possible, they are used less often. Component arrays usually provide space saving but this is not the major advantage of these devices. They are the least densely packed of the integrated circuits, limited mainly by the number of off-chip connections needed. While it may be possible to place over a million transistors on a single semiconductor chip, individual access to these would require a total of three million pins and this is beyond the limits of practicability. More commonly, resistor and diode arrays contain from five to 16 individual devices and transistor arrays contain from three to six individual transistors. The advantage of these arrays is the very close matching of component values within the array. In a circuit that needs matched components, the component array is often a good method of obtaining this feature. The components within an array can be internally combined for special functions, such as termination resistors, diode bridges and Darlington pair transistors. A nearly infinite number of possibilities exists for these combinations of components and many of these are available in arrays.

Multivibrators

A *multivibrator* is a circuit that oscillates, usually with a square wave output in the audio frequency range. The frequency of oscillation is accurately controlled with the addition of appropriate values of external resistance and capacitance. The most common multivibrator in use today is the 555 (NE555 by Signetics [now Philips] or LM555 by National Semiconductor). This very simple eight-pin DIP device

has a frequency range from less than one hertz to several hundred kilohertz. Such a device can also be used in *monostable* operation, where an input pulse generates an output pulse of a different duration, or in *astable* operation, where the device freely oscillates. Some other applications of a multivibrator are as a frequency divider, a delay line, a pulse width modulator and a pulse position modulator.

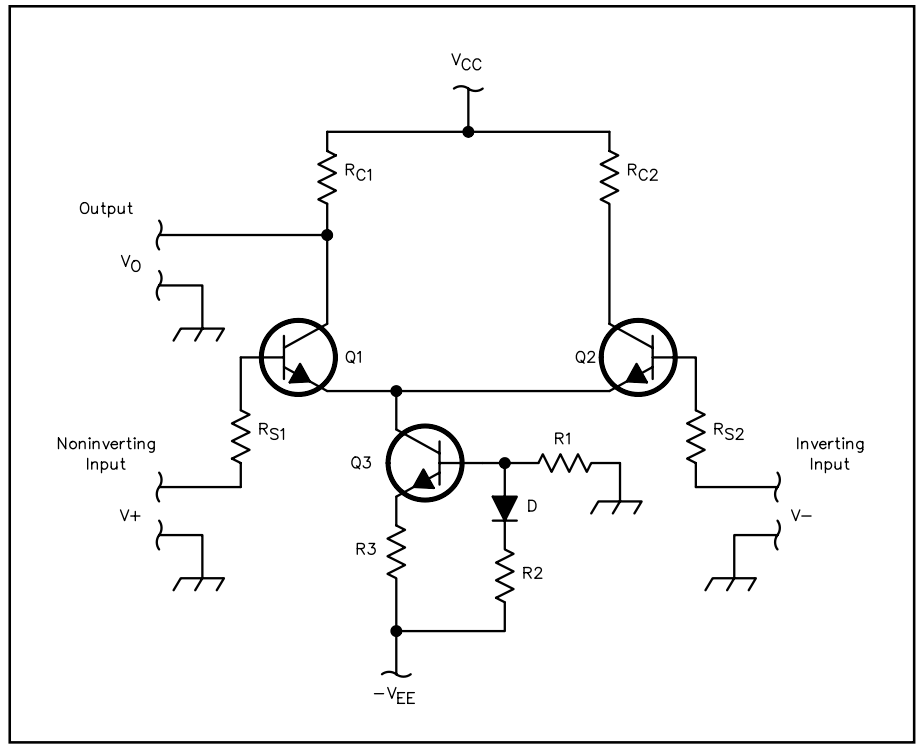

Fig 8.43 — Schematic of the components that make up an operational amplifier. Q1 and Q2 are matched emitter-coupled amplifiers. Q3 provides a constant current source. The symmetry of this device makes the matching of the components critical to its operation. This is why this circuit is usually implemented only in integrated circuitry. This simple op amp design has a large dc offset voltage at the output. Most practical designs include a level-shifting circuit, so the output voltage can exist near ground potential.

Operational Amplifiers

An *operational amplifier*, or *op amp*, is one of the most useful linear devices that has been developed in integrated circuitry. While it is possible to build an op amp with discrete components, the symmetry of this circuit requires a close match of many components and is more effective, and much easier, to implement in integrated circuitry. **Fig 8.43** shows a basic op-amp circuit. The op amp approaches a perfect analog circuit building block.

Ideally, an op amp has an infinite input impedance (Z_i), a zero output impedance (Z_o) and an open loop voltage gain (A_v) of infinity. Obviously, practical op amps do not meet these specifications, but they do come closer than most other types of amplifiers. An older op amp that is based on bipolar transistor technology, the LM324, has the following characteristics: guaranteed minimum CMRR of 65 dB, guaranteed minimum A_v of 25000, an input bias current (related to Z_i) guaranteed to be below 250 nA (2.5×10^{-7} A), output current capability (which determines Z_o) guaranteed to be above 10 mA and a gain-bandwidth product of 1 MHz. The TL084, which is a pin compatible replacement for the LM324 but is made with both JFET and bipolar transistors, has a guaranteed minimum CMRR of 80 dB, an input bias current guaranteed to be below 200 pA (2.0×10^{-10} A, almost 1000 times smaller than the LM324) and a gain-bandwidth product of 3 MHz. Philips has recently introduced the LMC6001 op amp with an input bias current of 25 fA (2.5×10^{-14} A, almost 10,000 times smaller than the TL084). This is equivalent to 156

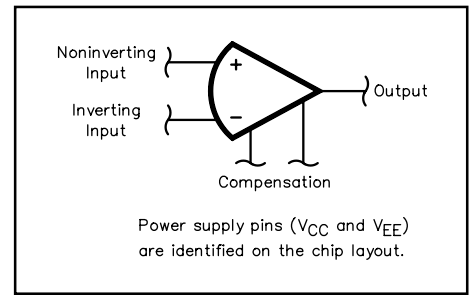

Fig 8.44 — Operational amplifier schematic symbol. The terminal marked with a + sign is the noninverting input. The terminal marked with a – sign is the inverting input. The output is to the right. On some op amps, external compensation is needed and leads are provided, pictured here below the device. Usually, the power supply leads are not shown on the op amp itself but are specified in the data sheet.

electrons entering the device every millisecond and corresponds to nearly infinite input impedance. Op amps can be customized to perform a large variety of functions by the addition of external components.

The typical op amp has three signal terminals (see). There are two input terminals, the noninverting terminal marked with a + sign and the inverting terminal marked with a – sign. The output of the amplifier has a single terminal and all signal levels within the op amp float, which means they are not tied to a specific reference. Rather, the reference of the input signals becomes the reference for the output signal. In many circuits this reference level is ground. Older operational amplifiers have an additional two connections for *compensation*. To keep the amplifier from going into oscillation at very high gains (increase its stability) it is often necessary to place a capacitor across the compensation terminals. This also decreases the frequency response of the op amp. Most modern op amps are internally compensated and do not have separate pins to add compensation capacitance. Additional compensation can be attained by connecting a capacitor between the op amp output and the inverting input.

One of the major advantages of using an op amp is its very high common mode rejection ratio (CMRR). Since there are two input terminals to an op amp, anything that is common to both terminals will be subtracted from the signal during amplification. The CMRR is a measure of the effectiveness of this removal. High CMRR results from the symmetry between the circuit halves. The rejection of power-supply noise is also an important parameter of an op amp. This is attained similarly, since the power supply is connected equally to both symmetrical halves of the op amp circuit. Thus, the power supply rejection ratio (PSRR) is similar to the CMRR and is often specified on the device data sheets.

Just as the symmetry of the transistors making up an op amp leads to a device with high values of Z_i, A_v and CMRR and a low value of Z_o, a symmetric combination of op amps is used to further improve these parameters. This circuit, shown in **Fig 8.45**, is called an instrumentation amplifier.

The op amp is capable of amplifying signals to levels limited mainly by the power supplies. Two power supplies are required, thus defining the range of signal voltages that can be processed. In most op amps the signal levels that can be handled are less than the power supply limits (rails), usually one or two diode drops (0.7 V or 1.4 V) away from each rail. Thus, if an op amp has 15 V connected as its upper rail (usually denoted V^+) and ground connected as its lower rail (V^-), input signals can be amplified to be as high as 13.6 V and as low as 1.4 V in most amplifiers. Any values that would be amplified beyond those limits are clamped (output voltages that should be 1.4 V or less appear as 1.4 V and those that should be 13.6 V or more appear as 13.6 V). This clamping action is illustrated in Fig 8.1. Recently, op amps have been developed to handle signals all the way out to the power supply rails (for example, the MAX406, from Maxim Integrated Products).

If a signal is connected to the input terminals of an op amp, it will be amplified as much as the device is able (up to A_v), and will probably grow so large that it clamps, as described above. Even if such large gains are desired, A_v varies from one device to the next and cannot be guaranteed. In most applications the op amp gain is limited to a more reasonable value and this is usually realized by providing a negative feedback path from the

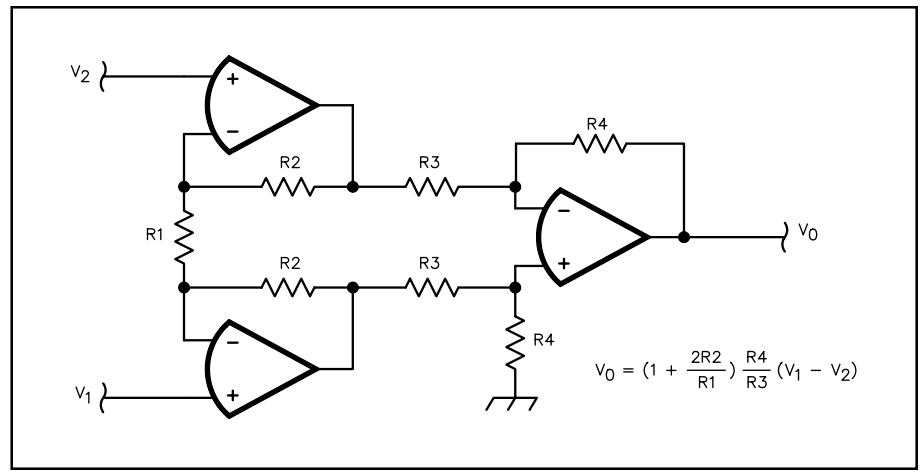

Fig 8.45 — Operational amplifiers arranged as an instrumentation amplifier. The balanced and cascaded series of op amps work together to perform differential amplification with good common-mode rejection and very high input impedance (no load resistor required) on both the inverting (V_1) and noninverting (V_2) inputs.

output terminal to the inverting input terminal. The *closed loop gain* of an op amp depends solely on the values of the passive components used to form the loop (usually resistors and, for frequency-selective circuits, capacitors). Some examples of different circuit configurations that manipulate the loop gain follow.

The op amp is often used as either an inverting or a noninverting amplifier. Accurate amplification can be achieved with just two resistors: the feedback resistor, R_f, and the input resistor, R_i (see **Fig 8.46**). If connected in the noninverting configuration, the input signal is connected to the noninverting terminal. The feedback resistor is connected between the output and the inverting terminal. The inverting terminal is connected to R_i, which is connected to ground. The gain of this configuration is:

$$\frac{V_o}{V_n} = \left(1 + \frac{R_f}{R_i}\right) \qquad (15)$$

where:

V_o is the output voltage, and

V_n is the input voltage to the noninverting terminal.

In the inverting configuration, the input signal (V_i) is connected through R_i to the inverting terminal. The feedback resistor is again connected between the inverting terminal and the output. The noninverting terminal can be connected to ground or to a dc offset voltage. The gain of this circuit is:

$$\frac{V_o}{V_i} = -\frac{R_f}{R_i} \qquad (16)$$

where V_i represents the voltage input to R_i.

The negative sign in equation 16 indicates that the signal is inverted. For ac signals, inversion represents a 180° phase shift. The gain of the noninverting op amp can vary from a minimum of $\times 1$ to the maximum of which the device is capable. The gain of the inverting op amp configuration can vary from a minimum of $\times 0$ (gains from $\times 0$ to $\times 1$ attenuate the signal while gains of $\times 1$ and higher amplify the signal) to the maximum that the device is capable of, as indicated by A_v for dc signals, or the gain-bandwidth product for ac signals. Both parameters are usually specified in the manufacturer's data sheet.

A voltage follower is a type of op amp that is commonly used as a buffer stage. The voltage follower has the input connected directly to the noninverting terminal and the output connected directly to the inverting terminal (**Fig 8.47**). This configuration has unity gain and provides the maximum possible input impedance and the minimum possible output impedance of which the device is capable.

A *differential amplifier* is a special application of an operational amplifier (see **Fig 8.48**). It amplifies the difference between two analog signals and is very useful to cancel noise under certain

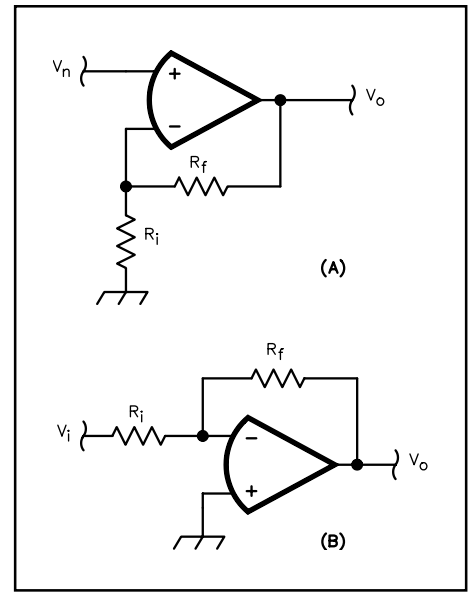

Fig 8.46 — Operational amplifier circuits. (A) Noninverting configuration. (B) Inverting configuration.

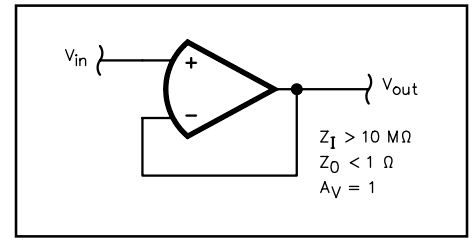

Fig 8.47 — Voltage follower. This operational amplifier circuit makes a nearly ideal buffer with a voltage gain of about one, extremely high input impedance and extremely low output impedance.

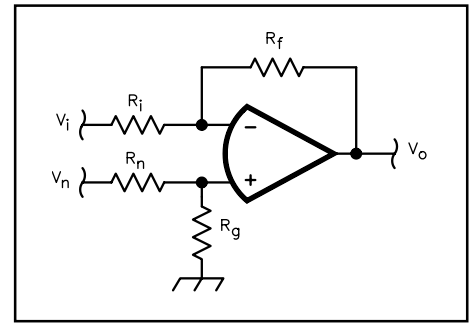

Fig 8.48 — Differential amplifier. This operational amplifier circuit amplifies the difference between the two input signals.

conditions. For instance, if an analog signal and a reference signal travel over the same cable they may pick up noise, and it is likely that both signals will have the same amount of noise. When the differential amplifier subtracts them, the signal will be unchanged but the noise will be completely removed, within the limits of the CMRR. The equation for differential amplifier operation is

$$V_o = \frac{R_f}{R_i}\left[\frac{1}{\frac{R_n}{R_g}+1}\left(\frac{R_i}{R_f}+1\right)V_n - V_i\right] \tag{17}$$

which, if the ratios $\frac{R_i}{R_f}$ and $\frac{R_n}{R_g}$ are equal, simplifies to

$$V_o = \frac{R_f}{R_i}\left(V_n - V_i\right) \tag{18}$$

Note that the differential amplifier response is identical to the inverting op amp response (equation 16) if the voltage source to the noninverting terminal is equal to zero. If the voltage source to the inverting terminal (V_i) is set to zero, the analysis is a little more complicated but it is possible to derive the noninverting op amp response (equation 15) from the differential amplifier response by taking into account the influence of R_n and R_g.

DC offset is an important consideration in op amps for two reasons. Actual op amps have a slight mismatch between the inverting and noninverting terminals that can become a substantial dc offset in the output, depending on the amplifier gain. The op amp output must not be too close to the clamping limits or distortion will occur. Introduction of a small dc correction voltage to the noninverting terminal is sometimes used to apply an offset voltage that counteracts the internal mismatch and centers the signal in the rail-to-rail range.

The high input impedance of an op amp makes it ideal for use as a *summing amplifier*. In either the inverting or noninverting configuration, the single input signal can be replaced by multiple input signals that are connected together through series resistors, as shown in **Fig 8.49**. For the inverting summing amplifier, the gain of each input signal can be calculated individually using equation 16 and, because of the superposition property, the output becomes the sum of each input signal multiplied by its gain. In the noninverting configuration, the output is the gain times the weighted sum of the m different input signals:

$$V_n = V_{n1}\frac{R_{p1}}{R_1 + R_{p1}} + V_{n2}\frac{R_{p2}}{R_2 + R_{p2}} + \ldots + V_{nm}\frac{R_{pm}}{R_m + R_{pm}} \tag{19}$$

where R_{pm} is the parallel resistance of all m resistors excluding R_m. For example, with three signals being summed, R_{p1} is the parallel combination of R_2 and R_3.

Other combinations of summing and difference amplification

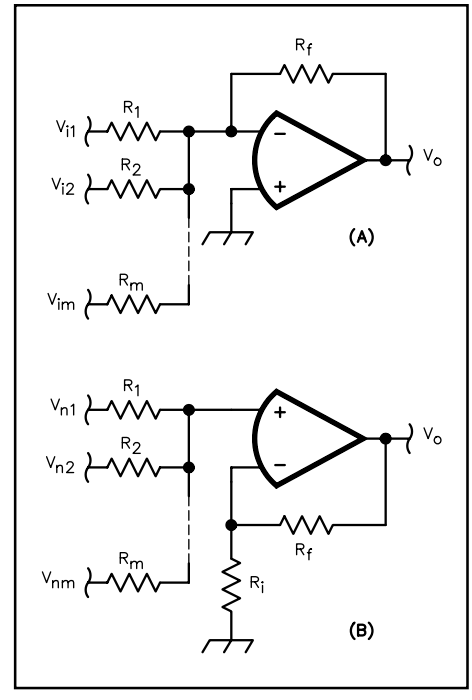

Fig 8.49 — Summing operational amplifier circuits. (A) Inverting configuration. (B) Noninverting configuration.

can be realized with a single op amp. The analyses of such circuits use the standard op amp equations coupled with the principle of superposition.

A *voltage comparator* is another special form of an operational amplifier. It takes in two analog signals and provides a binary output that is true if the voltage of one signal is bigger than that of the other, and false if not. A standard operational amplifier can be made to act as a comparator by connecting the two voltages to the noninverting and inverting inputs with no input or feedback resistors. If the voltage of the noninverting input is higher than that of the inverting input, the output voltage will be clamped to the positive clamping limit. If the inverting input is at a higher potential than the noninverting input, the output voltage will be clamped to the negative clamping limit (although this is not necessarily a negative voltage, depending on the value of the lower rail). Some applications of a voltage comparator are a zero crossing detector, a signal squarer (which turns other cyclical wave forms into square waves) and a peak detector.

Charge Coupled Devices

As the speed of integrated circuitry increases, it becomes possible to process some of the signals digitally while other processing occurs in analog form, all of this on the same IC chip. Such a chip is often called a *mixed modality* or *hybrid* chip (not to be confused with the hybrid circuitry discussed earlier). An example of this is the *charge coupled device (CCD)*. Pure digital analysis of signals requires digitization in two domains, namely the time sampling of a signal into individual packets and the amplitude sampling of each time packet into digital levels. CCDs perform time sampling but the time packets remain in analog form; they can take on any voltage value rather than a fixed number of discrete values. The CCD is often used to produce a delay filter. While most analog filters introduce some phase shift or delay into the signal, the relationship between the phase shift and the frequency is not always linear; different frequencies are delayed by different amounts of time. The goal of an ideal delay filter is to delay all parts of the signal by the same time. The CCD is used to realize this by sampling the signal, shifting the time packets through a series of capacitors and then reconstructing the continuous signal at the other end. The rate of shifting the time packets and the number of stages determines the amount of the delay. When originally introduced in the late 1970s, CCDs were described as bucket brigade devices (after the old fire fighting technique), where the buckets filled with signal packets are passed along the line until they are dumped at the end and recombined into an analog signal. These devices are simply constructed in an IC where each bucket is a MOS capacitor that is surrounded by two MOSFETs. When the transistors are biased to conduct, the charge moves from one bucket to the next and, while biased off, the charges are held in their capacitors. Very accurate filters, called *switched capacitor filters*, can be made with CCDs (see the **Filters** chapter).

A special form of CCD has also become quite popular in recent years, replacing the vidicon in modern camera circuitry. A two dimensional array of CCD elements has been developed with light sensitive semiconductor material; the charge that enters the capacitors is proportional to the amount of light incident on that location of the chip. The charges are held in their array of capacitors until shifted out, one horizontal line at a time, in a raster format. The CCD array mimics the operation of the vidicon camera and has many advantages. CCD response linearity across the field is superior to that of the vidicon. Very bright light at one location saturates the CCD elements only at that location rather than the blooming effect in vidicons where bright light spreads radially from the original location. CCD imaging elements do not suffer from image retention, which is another disadvantage of vidicon tubes.

Balanced Mixers

The *balanced mixer* is a device with many applications in modern radio transceivers (see the **Mixers, Modulation and Demodulation** chapter). Audio signals can be modulated onto a carrier or demodulated from the carrier with a balanced mixer. RF signals can be downconverted to intermediate frequency (IF) or IF can be upconverted to RF with a balanced mixer. This device is made with a bridge of four

matched Schottky diodes and the necessary transformers packaged in a small metal, plastic or ceramic container. The consequence of unmatched diodes is poor isolation between the local oscillator (LO) and the two signals. IC mixers often usrt cell" to provide LO isolation as high as -30 dB at 500 MHz. The isolation improves with decreasing frequency.

Receiver Subsystems

High performance ICs have been designed that make up complete receivers with the addition of only a few external components. Two examples that are very similar are the Motorola MC3363 and the Philips NE627. Both of these chips have all the active RF stages necessary for a double conversion FM receiver. The MC3363 has an internal local oscillator (LO) with varactor diodes that can generate frequencies up to 200 MHz, although the rest of the circuit is capable of operating at frequencies up to 450 MHz with an external oscillator. The RF amplifier has a low noise factor and gives this chip a 0.3 μV sensitivity. The intermediate frequency stages contain limiter amplifiers and quadrature detection. The necessary circuitry to implement receiver squelch and zero crossing detection of FSK modulation is also present. The circuit also contains received signal strength ("S-meter") circuitry (RSSI). The input and output of each stage are also brought out of the chip for versatility. The audio signal out of this chip must be appropriately amplified to drive a low-impedance speaker. This chip can be driven with a dc power source from 2 to 7 V and it draws only 3 mA with a 2 V supply.

The Philips NE627 is a newer chip than the MC3363 and has better performance characteristics even though it has essentially the same architecture. Its LO can generate frequencies up to 150 MHz and external oscillator frequencies up to 1 GHz can be used. The chip has a 4.6 dB noise figure and 0.22 μV sensitivity. The circuit can be powered with a dc voltage between 4.5 and 8 V and it draws between 5.1 and 6.7 mA. This chip is also ESD hardened so it resists damage from electrostatic discharges, such as from nearby lightning strikes.

The various stages in the receiver subsystem ICs are made available by connections on the package. There are two reasons that this is done. Filtering that is added between stages can be performed more effectively with inductors and crystal or ceramic filters, which are difficult to fabricate in integrated circuitry, so the output of one stage can be filtered externally before being fed to the next stage. It also adds to the versatility of the device. Filter frequencies can be customized for different intermediate frequencies. Stages can be used individually as well, so these devices can be made to perform direct conversion or single conversion reception or other forms of demodulation instead of FM.

Older integrated circuits that are sub-sets of the receiver subsystems are popular. The NE602 contains one double balanced mixer and a local oscillator, along with voltage regulation and buffering (**Fig 8.50**). It contains almost everything required to construct a direct conversion receiver. Its small size, an 8 pin DIP, makes it more desirable for this purpose than using part of an MC3363, which is in a 24 pin DIP and is more expensive. The NE604 contains the IF amplifiers and quadrature detector that, together with two NE602s and an RF amplifier, could almost duplicate the functions of the MC3363 or the NE367.

Transmitter Subsystems

Single chips are available to implement FM transmitters. One implementation is the Motorola MC2831A. This chip contains a mike preamplifier with limiting, a tone generator for CTCSS or AFSK, and a frequency modulator. It has an internal voltage controlled oscillator that can be controlled with a crystal or an LC circuit. This chip also contains circuitry to check the power supply voltage and produce a warning if it falls too low. Together with an FM receiver IC, an entire transceiver can be fabricated with very few parts.

Monolithic Microwave Integrated Circuit

A class of bipolar IC that is capable of higher frequency responses is the *monolithic microwave*

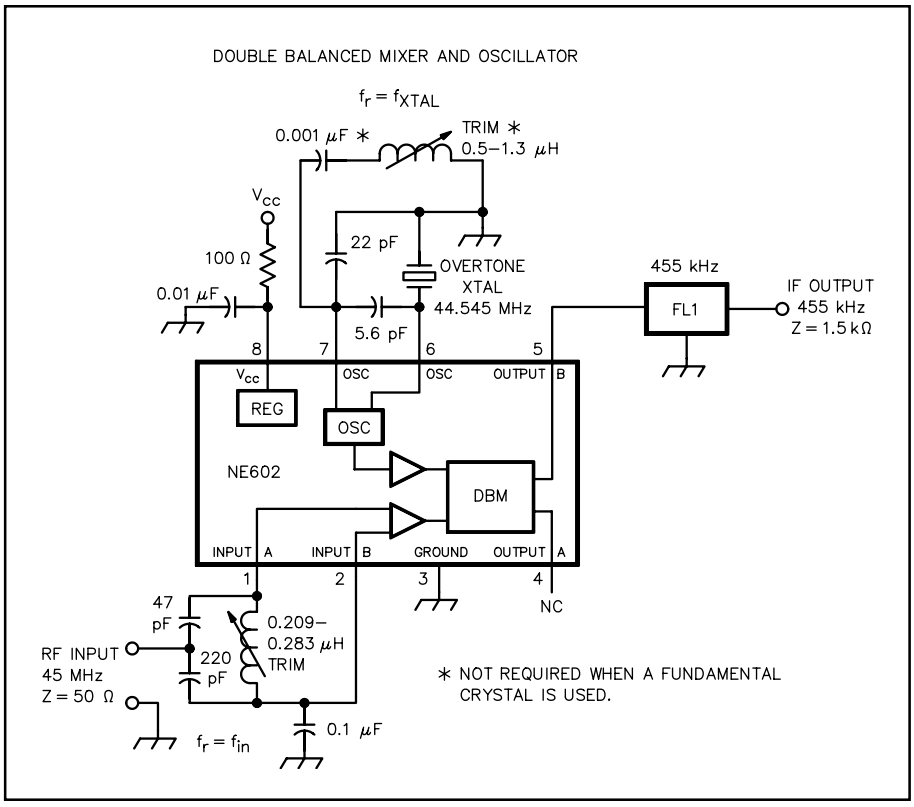

Fig 8.50 — The NE602 functional block diagram in circuit. This device contains a doubly balanced mixer, a local oscillator, buffers and a voltage regulator. This application uses the NE602 to convert an RF signal in a receiver to IF.

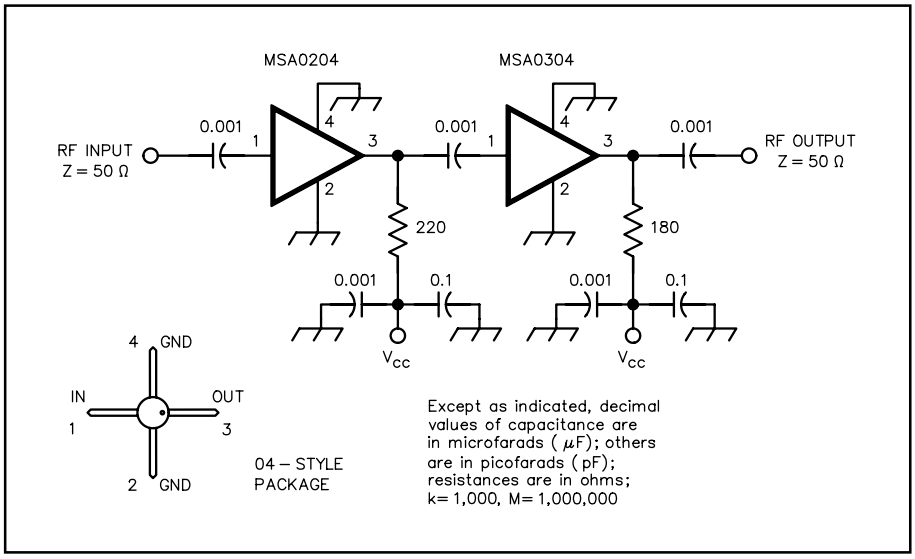

Fig 8.51 — The MSA0204 and MSA0304 MMICs in circuit. Both amplifiers have both input and output impedance of 50 Ω and a bandwidth of more than 2.5 GHz.

integrated circuit (MMIC). There is no formal definition of when an IC amplifier becomes an MMIC and, as the performance of IC devices improves, particularly MOS based devices, the distinction is becoming blurred. MMIC devices typically have predefined operating characteristics and require few external components. An example of an MMIC is a fixed gain amplifier, the MSA0204 (**Fig 8.51**), which can deliver 12 dB of gain up to 1 GHz. More modern MMIC devices are being developed with bandwidths in the tens of GHz.

Comparison of Analog Signal Processing Components

Analog signal processing deals with changing a signal to a desired form. Vacuum tubes, bipolar transistors, field-effect transistors and integrated circuitry perform similar functions, each with specific advantages and disadvantages. These are summarized here.

Of the four component types, vacuum tubes are physically the largest and require the most operating power. They have more limited life spans, usually because the heater filament burns out just as a light bulb does. Regardless of its use, a vacuum tube always generates heat. Miniaturization is difficult with vacuum tubes both because of their size and because of the need for air space around them for cooling. Vacuum tubes do have advantages, however. They are electrically robust. You need not be as concerned about static charges destroying vacuum tubes. A transmitter with vacuum tube finals usually has a variable matching network built in, and can be loaded into a higher SWR than

one with semiconductor finals. Tubes are generally able to withstand the high voltages generated by reflections under high SWR conditions. They are not as easily damaged by short-term overloads or the electromagnetic pulses generated by lightning. The relatively high plate voltages mean that the plate current is lower for a given power output; thus power supplies do not need as high a current handling capability. Vacuum tubes are capable of considerable heat dissipation and many high power applications still use them. Special forms of vacuum tubes are also still used. Most video displays use CRTs, and microwave transmitting tubes are still common.

Bipolar transistors have many advantages over vacuum tubes. When treated properly they can have virtually unlimited life spans. They are relatively small and, if they do not handle high currents, do not generate much heat, improving miniaturization. They make excellent high-frequency amplifiers. Compared to MOSFET devices they are less susceptible to damage from electrostatic discharge. RF amplifiers designed with bipolar transistors in their finals generally include circuitry to protect the transistors from the high voltages generated by reflections under high SWR conditions. Lightning strikes in the area (not direct hits) have been known to destroy all kinds of semiconductors, including bipolar transistors. Semiconductors have replaced almost all small-signal applications of tubes.

There are many performance advantages to FET devices, particularly MOSFETs. The extremely low gate currents allow the design of analog stages with nearly infinite input resistance. Signal distortion due to loading is minimized in this way. As these characteristics are improved by technology, we are seeing an increase in FET design at the expense of bipolar transistors.

The current trend in electronics is portability. Transceivers are decreasing in size and in their power requirements. Integrated circuitry has played a large part in this trend. Extremely large circuits have been designed with microscopic proportions. It is more feasible to use MOSFETs within an IC chip than as discrete components since the devices at risk are usually those that are connected to the outside world. It is not necessary to use electrostatic discharge protection circuitry on the gate of every MOSFET in an IC; only the ones that connect to the pins on the chip need this protection. This arrangement both improves the performance of the internal MOSFETs and decreases the circuit size even further. Semiconductors are slowly replacing the last tube applications. CCD chips have been so successful in video cameras that it is difficult to find an application for vidicon tubes. The liquid crystal displays (LCDs) in laptop computers have given considerable competition to the CRT tube.

An important consideration in the use of analog components is the future availability of parts. At an ever increasing rate, as new components are developed to replace older technology, the older components are discontinued by the manufacturers and become unavailable for future use. This tends to be a fairly long term process but it is not unusual for a manufacturer to stop offering a component when demand for it falls. This has become evident with vacuum tubes, which are becoming more difficult to find and more expensive as fewer manufacturers produce them.

The major disadvantages of IC technology have been power handling capability, frequency response and noncustomized circuitry. These characteristics have improved at an amazing pace over recent years; it is a process that feeds itself. As ICs are improved they are used to make more powerful tools (such as computers and electronic test equipment) that are used in the design of further IC improvements. Entire transceivers are designed with just a few IC chips and the appropriate transistors for power amplification. The quiescent current draw of these devices has been reduced to the microampere level so they can operate effectively from small battery packs. The improved noise performance of circuitry has also decreased the need for high transmitter power, further decreasing the current requirements for these devices. If this trend continues, we should eventually see a near total switch to IC components with few discrete semiconductors and no vacuum tubes.