Contents

Chapter 3 — CD-ROM Content

Supplemental Articles

- "Hands-On Radio: The Common Emitter Amplifier" by Ward Silver, NØAX
- "Hands-On Radio: The Emitter-Follower Amplifier" by Ward Silver, NØAX
- "Hands-On Radio: The Common Base Amplifier" by Ward Silver, NØAX
- "Hands-On Radio: Field Effect Transistors" by Ward Silver, NØAX
- "Hands-On Radio: Basic Operational Amplifiers" by Ward Silver, NØAX
- Cathode Ray Tubes
- Large Signal Transistor Operation

Tools and Data

- *LTSpice* Simulation Files for Chapter 3
- Frequency Response Spreadsheet

Chapter 3

Analog Basics

The first section of this chapter discusses a variety of techniques for working with analog signals, called signal processing. These basic functions are combined to produce useful systems that become radios, instruments, audio recorders and so on.

Subsequent sections present several types of active components and circuits that can be used to manipulate analog signals. (An active electronic component is one that requires a power source to function; passive components such as resistors, capacitors and inductors are described in the **Electrical Fundamentals** chapter.) The chapter concludes with a discussion of analog-digital conversion by which analog signals are converted into digital signals and vice versa.

Material in this chapter was updated by Ward Silver, NØAX, and Courtney Duncan, N5BF, building on material from previous editions by Greg Lapin, N9GL, and Leonard Kay, K1NU. The section "Poles and Zeroes" was adapted from material contributed by David Stockton, GM4ZNX and Fred Telewski, WA7TZY.

3.1 Analog Signal Processing

The term *analog signal* refers to voltages, currents and waves that make up ac radio and audio signals, dc measurements, even power. The essential characteristic of an analog signal is that the information or energy it carries is continuously variable. Even small variations of an analog signal affect its value or the information it carries. This stands in contrast to *digital signals* (described in the **Digital Basics** chapter) that have values only within well-defined and separate ranges called *states*. To be sure, at the fundamental level all circuits and signals are analog: Digital signals are created by designing circuits that restrict the values of analog signals to those discrete states.

Analog signal processing involves various electronic stages to perform functions on analog signals such as amplifying, filtering, modulation and demodulation. A piece of electronic equipment, such as a radio, is constructed by combining 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 sections later in the chapter.

3.1.1 Linearity

The premier properties of analog signals are *superposition* and *scaling*. Superposition is the property by which signals are combined, whether in a circuit, in a piece of wire, or even in air, as 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 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. (Mathematically, this is a *linear combination*.) For this reason, analog signals and components are often referred to as *linear signals* or *linear components*. A linear system whose characteristics do not change, such as a resistive voltage divider, is called *time-invariant*. If the system changes with time, it is *time-varying*. The variations may be random, intermittent (such as being adjusted by an operator) or periodic.

One of the more important features of superposition, for the purposes of signal processing, is that signals that have been combined by superposition can be separated back into the original signals. This is what allows multiple signals that have been received by an antenna to be separated back into individual signals by a receiver.

3.1.2 Linear Operations

Any operation that modifies a signal and obeys the rules of superposition and scaling is a *linear operation*. The following sections explain the basic linear operations from which linear systems are made.

AMPLIFICATION AND ATTENUATION

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.

An *amplifier* is 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 3.1**). The linear amplifier multiplies every value of a signal by a constant value. Amplifier gain is often expressed as a multiplication factor (× 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. (Gain is often expressed in decibels (dB) as explained in the chapter on **Electrical Fundamentals**.)

Certain types of amplifiers for which currents are the input and output signals have *current gain*. Most amplifiers have both voltage and current gain. *Power gain* is the ratio of output power to input power.

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. (Gain can also be less than one.) The ability of an amplifier to change a signal's level is limited by the amplifier's *dynamic range*, however. An amplifier's dynamic range is the range of signal levels over which the amplifier produces the required gain without distortion. Dynamic range is limited for small signals by noise, distortion and other nonlinearities.

Dynamic range is limited for large signals because an amplifier can only produce output voltages (and currents) that are within the range of its power supply. (Power-supply voltages are also called the *rails* of a circuit.) As the amplified output approaches one of the rails, the output can not exceed a given voltage near the rail and the operation of the amplifier becomes nonlinear as described below in the section on Clipping and Rectification.

Another similar limitation on amplifier linearity is called *slew rate*. Applied to an amplifier, this term describes the maximum rate at which a signal can change levels and still be accurately amplified in a particular device. *Input slew rate* is the maximum rate of change to which the amplifier can react linearly. *Output slew rate* refers to the maximum rate at which the amplifier's output circuit can change. Slew rate is an important concept, because there is

Obtaining a Frequency Response

With the computer tools such as spreadsheets, it's easy to do the calculations and make a graph of frequency response. If you don't have a spreadsheet program, then use semi-log graph paper with the linear axis used for dB or phase and the logarithmic axis for frequency. An Excel spreadsheet set up to calculate and display frequency response is available on the CD that comes with this book. You can modify it to meet your specific needs.

Follow these rules whether using a spreadsheet or graph paper:

- Measure input and output in the same units, such as volts, and use the same conventions, such as RMS or peak-to-peak.
- Measure phase shift from the input to the output. (The **Test Equipment and Measurements** chapter discusses how to make measurements of amplitude and phase.)
- Use 10 log (P_O/P_I) for power ratios and 20 log (V_O/V_I) for voltage or current.

To make measurements that are roughly equally spaced along the logarithmic frequency axis, follow the "1-2-5 rule." Dividing a range this way, for example 1-2-5-10-20-50-100-200-500 Hz, creates steps in approximately equal ratios that then appear equally spaced on a logarithmic axis.

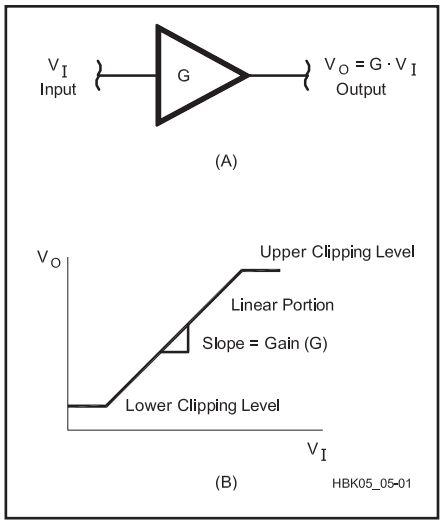

Fig 3.1 — Generic amplifier. (A) Symbol. For the linear amplifier, gain is the constant value, G, and the output voltage is equal to the input voltage times G; (B) Transfer function, input voltage along the x-axis is converted to the output voltage along the y-axis. The linear portion of the response 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.

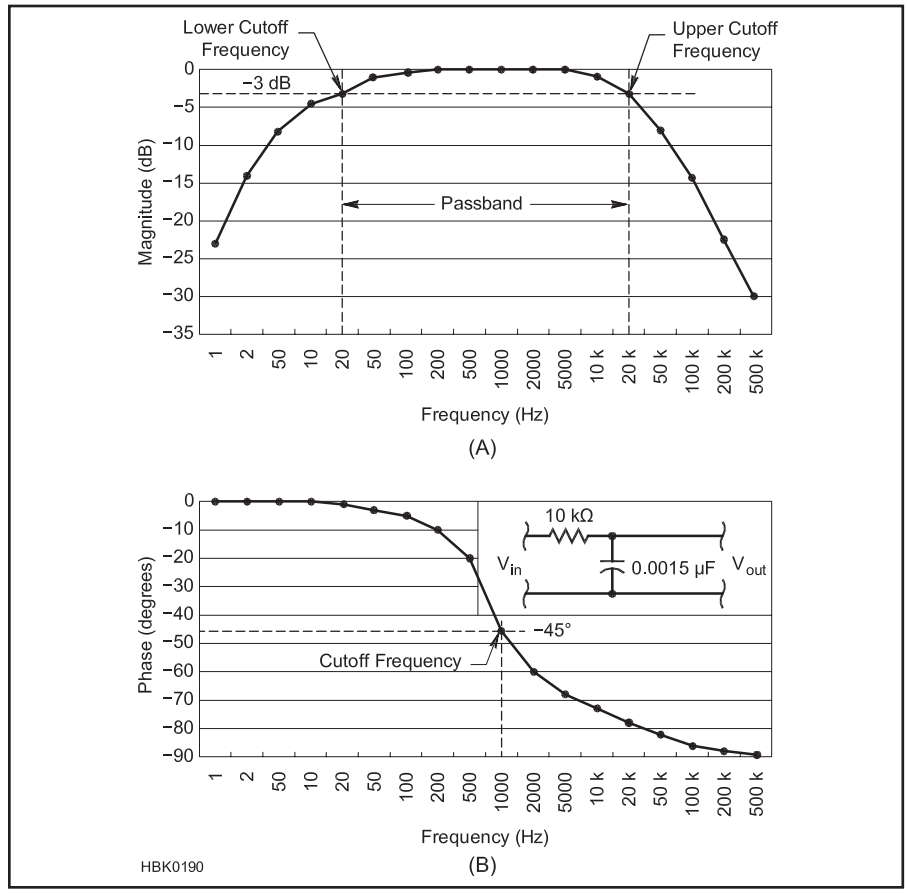

Fig 3.2 — Bode plot of (A) band-pass filter magnitude response and (B) an RC low-pass filter phase response.

a direct correlation between a signal level's rate of change and the frequency content of that signal. The amplifier's ability to react to or reproduce that rate of change affects its frequency response and dynamic range.

An *attenuator* is a circuit that reduces the amplitude of a signal. Attenuators can be constructed from passive circuits, such as the attenuators built using resistors, described in the chapter on **Test Equipment and Measurements**. Active attenuator circuits include amplifiers whose gain is less than one or circuits with adjustable resistance in the signal path, such as a PIN diode attenuator or amplifier with gain is controlled by an external voltage.

FREQUENCY RESPONSE AND BODE PLOTS

Another important characteristic of a circuit is its *frequency response*, a description of how it modifies a signal of any frequency. Frequency response can be stated in the form of a mathematical equation called a *transfer function*, but it is more conveniently presented as a graph of gain vs frequency. The ratio of output amplitude to input amplitude is often called the circuit's *magnitude* or *amplitude response*. Plotting the circuit's magnitude response in dB versus frequency on a logarithmic scale, such as in **Fig 3.2A**, is called a *Bode plot* (after Henrik Wade Bode). The combination of decibel and log-frequency scales is used because the behavior of most circuits depends on ratios of amplitude and frequency and thus appears linear on a graph in which both the vertical and horizontal scales are logarithmic.

Most circuits also affect a signal's phase along with its amplitude. This is called *phase shift*. A plot of phase shift from the circuit's input to its output is called the *phase response*, seen in Fig 3.2B. Positive phase greater than 0° indicates that the output signal *leads* the input signal, while *lagging* phase shift has a negative phase. The combination of an amplitude and phase response plot gives a good picture of what effect the circuit has on signals passing through it.

TRANSFER CHARACTERISTICS

Transfer characteristics are the ratio of an output parameter to an input parameter, such as output current divided by input current, h_{FE}. There are different families of transfer characteristics, designated by letters such as h, s, y or z. Each family compares parameters in specific ways that are useful in certain design or analysis methods. The most common transfer characteristics used in radio are the h-parameter family (used in transistor models) and the s-parameter family (used in RF design, particularly at VHF and above). See the **RF Techniques** chapter for more discussion of transfer characteristics.

COMPLEX FREQUENCY

We are accustomed to thinking of frequency as a real number — so many cycles per second — but frequency can also be a complex number, s, with both a real part, designated by σ, and an imaginary part, designated by $j\omega$. (ω is also equal to $2\pi f$.) The resulting complex frequency is written as $s = \sigma + j\omega$. At the lower left of **Fig 3.3** a pair of real and imaginary axes are used to plot values of s. This is called the *s-plane*. Complex frequency is used in Laplace transforms, a mathematical technique used for circuit and signal analysis. (Thorough treatments of the application of complex frequency can be found in college-level textbooks on circuit and signal analysis.)

When complex frequency is used, a sinusoidal signal is described by Ae^{st}, where A is the amplitude of the signal and t is time. Because s is complex, $Ae^{st} = Ae^{(\sigma+j\omega)t} = A(e^{\sigma t})(e^{j\omega t})$. The two exponential terms describe independent characteristics of the signal. The second term, $e^{j\omega t}$, is the sine wave with which we are familiar and that has frequency f, where $f = \omega/2\pi$. The first term, $e^{\sigma t}$, represents the rate at which the signal increases or decreases. If σ is negative, the exponential term decreases with time and the signal gets smaller and smaller. If σ is positive, the signal gets larger and larger. If $\sigma = 0$, the exponential term equals 1, a constant, and the signal amplitude stays the same.

Complex frequency is very useful in describing a circuit's stability. If the response to an input signal is at a frequency on the right-hand side of the s-plane for which $\sigma > 0$, the system is *unstable* and the output signal will get larger until it is limited by the circuit's power supply or some other mechanism. If the response is on the left-hand side of the s-plane, the system is *stable* and the response to the input signal will eventually die out. The larger the absolute value of σ, the faster the response changes. If the response is precisely on the $j\omega$ axis where $\sigma = 0$, the response will persist indefinitely.

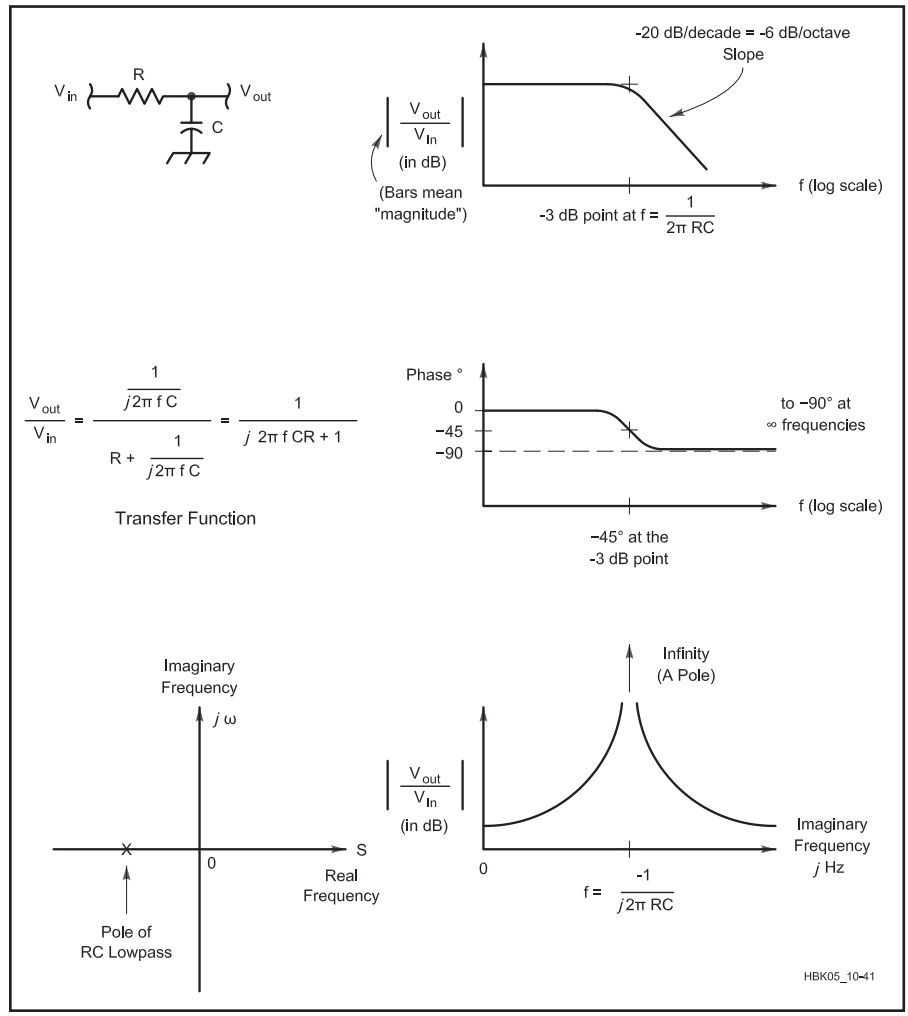

Fig 3.3 — The transfer function for a circuit describes both the magnitude and phase response of a circuit. The RC circuit shown at the upper left has a pole at $f = 2\pi RC$, the filter's –3 dB or cutoff frequency, at which the phase response is a 45° lagging phase shift. Poles cause an infinite response on the imaginary frequency axis.

In Fig 3.3 the equation for the simple RC-circuit's transfer function is shown at the left of the figure. It describes the circuit's behavior at real-world frequencies as well as imaginary frequencies whose values contain *j*. Because complex numbers are used for f, the transfer function describes the circuit's phase response, as well as amplitude. At one such frequency, $f = -j/2\pi RC$, the denominator of the transfer function is zero, and the gain is infinite! Infinite gain is a pretty amazing thing to achieve with a passive circuit — but because this can only happen at an imaginary frequency, it does not happen in the real world.

The practical effects of complex frequency can be experienced in a narrow CW crystal or LC filter. The poles of such a filter are just to the left of the $j\omega$ axis, so the input signal causes the filters to "ring", or output a damped sine wave along with the desired signal. Similarly, the complex frequency of an oscillator's output at power-up must have $\sigma > 0$ or the oscillation would never start! The output amplitude continues to grow until limiting takes place, reducing gain until $\sigma = 1$ for a steady output.

POLES AND ZEROES

Frequencies that cause the transfer function to become infinite are called *poles*. This is shown at the bottom right of Fig 3.3 in the graph of the circuit's amplitude response for imaginary frequencies shown on the horizontal axis. (The pole causes the graph to extend up "as a pole under a tent," thus the name.) Similarly, circuits can have *zeroes* which occur at imaginary frequencies that cause the transfer function to be zero, a less imaginative name, but quite descriptive.

A circuit can also have poles and zeroes at frequencies of zero and infinity. For example, the circuit in Fig 3.3 has a zero at infinity because the capacitor's reactance is zero at infinity and the transfer function is zero, as well. If the resistor and capacitor were exchanged, so that the capacitor was in series with the output, then at zero frequency (dc), the output would be zero because the capacitor's reactance was infinite, creating a zero.

Complex circuits can have multiple poles or multiple zeroes at the same frequency. Poles and zeroes can also occur at frequencies that are combinations of real and imaginary numbers. The poles and zeroes of a circuit form a pattern in the complex plane that corresponds to certain types of circuit behavior. (The relationships between the pole-zero pattern and circuit behavior is beyond the scope of this book, but are covered in textbooks on circuit theory.)

What is a Pole?

Poles cause a specific change in the circuit's amplitude and phase response for real-world frequencies, even though we can't experience imaginary frequencies directly. A pole is associated with a bend in a magnitude response plot that changes the slope of the response downward with increasing frequency by 6 dB per octave (20 dB per decade; an octave is a 2:1 frequency ratio, a decade is a 10:1 frequency ratio).

There are four ways to identify the existence and frequency of a pole as shown in Fig 3.3:

1. For a downward bend in the magnitude versus frequency plot, the pole is at the –3 dB frequency for a single pole. If the bend causes a change in slope of more than 6 dB/octave, there must be multiple poles at this frequency.

2. A 90° lagging change on a phase versus frequency plot, where the lag increases with frequency. The pole is at the point of 45° added lag on the S-shaped transition. Multiple poles will add their phase lags, as above.

3. On a circuit diagram, a single pole looks like a simple RC low-pass filter. The pole is at the –3 dB frequency ($f = 1/2\pi RC$ Hz). Any other circuit with the same response has a pole at the same frequency.

4. In an equation for the transfer function of a circuit, a pole is a theoretical value of frequency that would result in infinite gain. This is clearly impossible, but as the value of frequency will either be absolute zero, or will have an imaginary component, it is impossible to make an actual real-world signal at a pole frequency.

For example, comparing the amplitude responses at top and bottom of Fig 3.3 shows that the frequency of the pole is equal to the circuit's –3 dB cutoff frequency ($1/2\pi fC$) multiplied by *j*, which is also the frequency at which the circuit causes a –45° (lagging) phase shift from input to output.

What Is a Zero?

A zero is the complement of a pole. In math, it is a frequency at which the transfer function equation of a circuit is zero. This is not impossible in the real world (unlike the pole), so zeroes can be found at real-number frequencies as well as complex-number frequencies.

Each zero is associated with an *upward* bend of 6 dB per octave in a magnitude response. Similarly to a pole, the frequency of the zero is at the +3 dB point. Each zero is associated with a transition on a phase-versus-frequency plot that reduces the lag by 90°. The zero is at the 45° leading phase point. Multiple zeroes add their phase shifts just as poles do.

In a circuit, a zero creates gain that increases with frequency forever above the zero frequency. This requires active circuitry that would inevitably run out of gain at some frequency, which implies one or more poles up there. In real-world circuits, zeroes are usually not found by themselves, making the magnitude response go up, but rather paired with a pole of a different frequency, resulting in the magnitude response having a slope between two frequencies but flat above and below them.

Real-world circuit zeroes are only found accompanied by a greater or equal number of poles. Consider a classic RC high-pass filter, such as if the resistor and capacitor in Fig 3.3 were exchanged. The response of such a circuit increases at 6 dB per octave from 0 Hz (so there must be a zero at 0 Hz) and then levels off at $1/2\pi RC$ Hz. This leveling off is due to the presence of a pole adding its 6 dB-per-octave roll-off to cancel the 6 dB-per-octave roll-up of the zero. The transfer function for such as circuit would equal zero at zero frequency and infinity at the imaginary pole frequency.

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*. (The chapter on **Oscillators and Synthesizers** includes a discussion of positive feedback.)

Most practical amplifiers have some intrinsic and unavoidable feedback either as part of the circuit or in the amplifying device(s) itself. 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.

Although negative feedback reduces amplifier or stage gain, the advantages of *stable* gain, freedom from unwanted oscillations and the reduction of distortion are often key design objectives and advantages of using negative feedback.

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.

SUMMING

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 according to the principle of superposition. When processing signals, we would like to control the summing operation so the signals do not distort or combine in a nonlinear way. If two signals come from separate stages and they are connected together directly, the circuitry of the stages may interact, causing distortion of either or both signals.

Summing amplifiers generally use a resistor in series with each stage, so the resistors connect to the common input of the following stage. This provides some *isolation* between the output circuits of each stage. **Fig 3.4** illustrates the resistors connecting to a summing amplifier. Ideally, any time we wanted to combine signals (for example, combining an audio signal with a sub-audible tone in a 2 meter FM transmitter prior to modulating the RF signal) we could use a summing amplifier.

FILTERING

A *filter* is a common linear stage in radio equipment. Filters are characterized by their ability to selectively attenuate certain frequencies in the filter's *stop band*, while passing or amplifying other frequencies in the *passband*. If the filter's passband extends to or near dc, it is a *low-pass* filter, and if to infinity (or at least very high frequencies for the circuitry involved), it is a *high-pass* filter. Filters that pass a range of frequencies are *band-pass* filters. *All-pass* filters are designed to affect only the phase of a signal without changing the signal amplitude. The range of frequencies between a band-pass circuit's low-pass and high-pass regions is its *mid-band*.

Fig 3.2A is the amplitude response for a typical band-pass audio filter. It shows that the input signal is passed to the output with no loss (0 dB) between 200 Hz and 5 kHz. This is the filter's *mid-band response*. Above and below those frequencies the response of the filter begins to drop. By 20 Hz and 20 kHz, the amplitude response has been reduced to one-half of (–3 dB) the mid-band response. These points are called the circuit's *cutoff* or *corner* or *half-power frequencies*. The range between the cutoff frequencies is the filter's passband. Outside the filter's passband, the amplitude response drops to 1/200th of (–23 dB) mid-band response at 1 Hz and only 1/1000th (–30 dB) at 500 kHz. The steepness of the change in response with frequency is the filter's *roll-off* and it is usually specified in dB/octave (an octave is a doubling or halving of frequency) or dB/decade (a decade is a change to 10 times or 1/10th frequency).

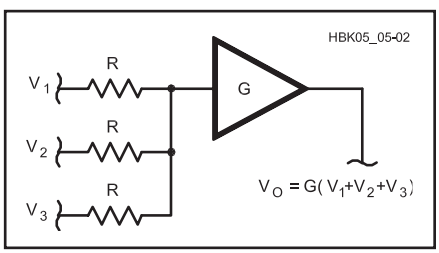

Fig 3.4 — Summing amplifier. The output voltage is equal to the sum of the input voltages times the amplifier gain, G. As long as the resistance values, R, are equal and the amplifier input impedance is much higher, the actual value of R does not affect the output signal.

Fig 3.2B represents the phase response of a different filter — the simple RC low-pass filter shown at the upper right. As frequency increases, the reactance of the capacitor becomes smaller, causing most of the input signal to appear across the fixed-value resistor instead. At low frequencies, the capacitor has little effect on phase shift. As the signal frequency rises, however, there is more and more phase shift until at the cutoff frequency, there is 45° of lagging phase shift, plotted as a negative number. Phase shift then gradually approaches 90°.

Practical passive and active filters are described in the **RF and AF Filters** chapter. Filters implemented by digital computation (*digital filters*) are discussed in the chapter on **DSP and Software Radio Design**. Filters at RF may also be created by using transmission lines as described in the **Transmission Lines** chapter. All practical amplifiers are in effect either low-pass filters or band-pass filters, because their magnitude response decreases as the frequency increases beyond their gain-bandwidth products.

AMPLITUDE MODULATION/ DEMODULATION

Voice and data signals can be transmitted over the air by using *amplitude modulation* (AM) to combine them with higher frequency *carrier* signals (see the **Modulation** 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 — amplitude modulation by the sum of two audio signals produces the same signal as the sum of amplitude modulation by each audio signal individually. Another aspect of the linear behavior of amplitude modulation is that amplitude-modulated signals can be demodulated exactly to their original form. Amplitude demodulation is the converse of amplitude modulation, and is represented as division, also a linear operation.

In the linear model of amplitude modulation, the signal that performs the modulation (such as the audio signal in 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 (and demodulation) is that it can be performed nonlinearly, as well. Accurate analog multipliers (and dividers) are difficult and expensive to fabricate, so less-expensive nonlinear methods are often employed. Each nonlinear form of amplitude modulation generates the desired linear combination of signals, called *products*, in addition to other unwanted products that must then be removed. Both linear and nonlinear modulators and demodulators are discussed in the chapter on **Mixers, Modulators, and Demodulators**.

3.1.3 Nonlinear Operations

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 the amount of attenuation at different frequencies is always the same, regardless of the amplitude of the signals passing through the filter.

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 from the signals). There are a number of reasons why we would like to do this. As shown in Fig 3.1, clipping is the process of limiting the positive and negative peaks of a signal. (Clipping is also called *clamping*.)

Clamping might be used to prevent a large audio signal from causing excessive deviation in an FM transmitter that would interfere with communications on adjacent channels. Clipping circuits are also used to protect sensitive inputs from excessive voltages. Clipping distorts the signal, changing it so that the original signal waveform is lost.

Another kind of clipping results in *rectification*. A *rectifier* circuit clips off all voltages of one polarity (positive or negative) and passes only voltages of the other polarity, thus changing ac to pulsating dc (see the **Power Sources** chapter). Another use of rectification is in a *peak detection* circuit that measures the peak value of a waveform. Only one polarity of the ac voltage needs to be measured and so a rectifier clips the unwanted polarity.

LIMITING

Another type of clipping occurs when an amplifier is intentionally operated with so much gain that the input signals result in an output that is clipped at the limits of its power supply voltages (or some other designated voltages) . The amplifier is said to be driven into *limiting* and an amplifier designed for this behavior is called a *limiter*. Limiters are used in FM receivers to amplify the signal until all amplitude variations in the signal are removed and the only characteristic of the original signal that remains is the frequency.

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** chapter).

3.2 Analog Devices

There are several different kinds of components that can be used to build circuits for analog signal processing. Bipolar semiconductors, field-effect semiconductors and integrated circuits comprise a wide spectrum of active devices used in analog signal processing. (Vacuum tubes are discussed in the chapter on **RF Power Amplifiers**, their primary application in Amateur Radio.) Several different devices can perform the same function, each with its own advantages and disadvantages based on the physical characteristics of each type of 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.

3.2.1 Terminology

A similar terminology is used when describing active electronic devices. The letter V or v stands for voltages and I or i for currents. Capital letters are often used to denote dc or bias values (bias is discussed later in this chapter). Lower-case often denotes instantaneous or ac values.

Voltages generally have two subscripts indicating the terminals between which the voltage is measured (V_{BE} is the dc voltage between the base and the emitter of a bipolar transistor). Currents have a single subscript indicating the terminal into which the current flows (I_C is the dc current into the collector of a bipolar transistor). If the current flows out of the device, it is generally treated as a negative value.

Resistance is designated with the letter R or r, and impedance with the letter Z or z. For example, r_{DS} is resistance between drain and source of an FET and Z_i is input impedance. For some parameters, 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. (See the section on transistor amplifiers later in this chapter for a discussion of the common-emitter circuit.) 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.

Power supply voltages have two subscripts that are the same, indicating the terminal to which the voltage is applied. V_{DD} would represent the power supply voltage applied to the drain of a field-effect transistor.

Since integrated circuits are collections of semiconductor components, the abbreviations for the type of semiconductor used also apply to the integrated circuit. For example, V_{CC} is a power supply voltage for an integrated circuit made with bipolar transistor technology in which voltage is applied to transistor collectors.

3.2.2 Gain and Transconductance

The operation of an amplifier is specified by its *gain*. 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*. *Power gain* is often used, as well. Gain is technically unit-less, but is often given in V/V. Decibels are often used to specify gain, particularly power gain.

If an amplifier's input is a voltage and the output is a current, the ratio of the change in output current to the change in input voltage is called *transconductance, g_m*.

$$g_m = \frac{\Delta I_o}{\Delta V_i} \qquad (2)$$

Transconductance has the same units as conductance and admittance, Siemens (S), but is only used to describe the operation of active devices, such as transistors or vacuum tubes.

3.2.3 Characteristic Curves

Analog devices are described most completely with their *characteristic curves*. The characteristic curve is a plot of the interrelationships between two or three variables. The vertical (y) axis parameter is the output, or

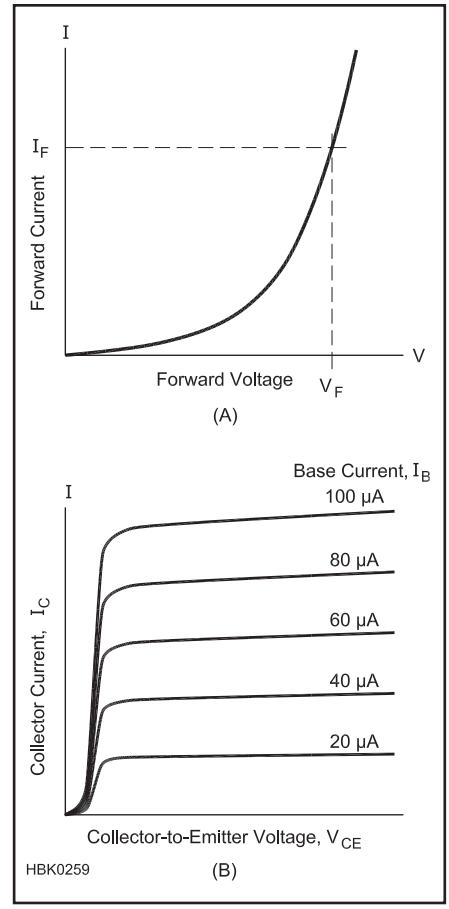

Fig 3.5 — Characteristic curves. A forward voltage vs forward current characteristic curve for a semiconductor diode is shown at (A). (B) shows a set of characteristic curves for a bipolar transistor in which the collector current vs collector-to-emitter voltage curve is plotted for five different values of base current.

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.

Almost all devices of concern 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. Characteristic curves are used to graphically describe a device's operation in both its linear and nonlinear regions.

Fig 3.5A shows the characteristic curve for a semiconductor diode with the y-axis showing the forward current, I_F, flowing through the diode and the x-axis showing forward voltage, V_F, across the diode. This curve shows the relationship between current and voltage in the diode when it is conducting current. Characteristic curves showing voltage and current in two-terminal devices such as diodes are often called *I-V curves*. Characteristic curves may include all four quadrants of operation in which both axes include positive and negative values. It is also common for different scales to be used in the different quadrants, so inspect the legend for the curves carefully.

The parameters plotted in a characteristic curve depend on how the device will be used so that the applicable design values can be obtained from the characteristic curve. The slope of the curve is often important because it relates changes in output to changes in input. To determine the slope of the curve, two closely-spaced 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, for example) is calculated as:

$$m = \frac{\Delta y}{\Delta x} = \frac{y_1 - y_2}{x_1 - x_2} \qquad (3)$$

It is important to pick points that are close together or the slope will not reflect the actual behavior of the device. A device whose characteristic curve is not a straight line will not have a linear response to inputs because the slope changes with the value of the input parameter.

For a device in which three parameters interact, such as a transistor, sets of characteristic curves can be drawn. Fig 3.5B shows a set of characteristic curves for a bipolar transistor where collector current, I_C, is shown on the y axis and collector-to-emitter voltage, V_{CE}, is shown on the x axis. Because the amount of collector current also depends on base current, I_B, the curve is repeated several times for different values of I_B. From this set of curves, an amplifier circuit using this transistor can be designed to have specific values of gain.

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. The device's *bias point* is its set of operating parameters when no input signal is applied. The bias point is also known as the *quiescent point* or *Q-point*. By changing the bias point, the circuit designer can affect the relationship between the input and output signal. The bias point can also be considered as a dc offset of the input signal. Devices that perform analog signal processing require appropriate input signal biasing.

As an example, consider the characteristic curve shown in **Fig 3.6**. (The exact types of device and circuit are unimportant.) The characteristic curve shows the relationship between an input voltage and an output current. Increasing input voltage results in an increase in output current so that an input signal is reproduced at the output. The characteristic curve is linear in the middle, but is quite nonlinear in its upper and lower regions.

In the circuit described by the figure, bias points are established by adding one of the three voltages, V_1, V_2 or V_3 to the input signal. Bias voltage V_1 results in an output current of I_1 when no input signal is present. This is shown as Bias Point 1 on the characteristic curve. When an input signal is applied, the input voltage varies around V_1 and the output current varies around I_1 as shown. If the dc value of the output current is subtracted, a reproduction of the input signal is the result.

If Bias Point 2 is chosen, we can see that the input voltage is reproduced as a changing output current with the same shape. In this case, the device is operating linearly. If either Bias Point 1 or Bias Point 3 is chosen, however, the shape of the output signal is distorted because the characteristic curve of the device is nonlinear in this region. Either the increasing portion of the input signal results in more variation than the decreasing portion (Bias Point 1) or vice versa (Bias Point 3). Proper biasing is crucial to ensure that a device operates linearly.

3.2.4 Manufacturer's Data Sheets

Manufacturer's data sheets list device characteristics, along with the specifics of the part type (polarity, semiconductor type), identity of the pins and leads (*pinouts*), 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

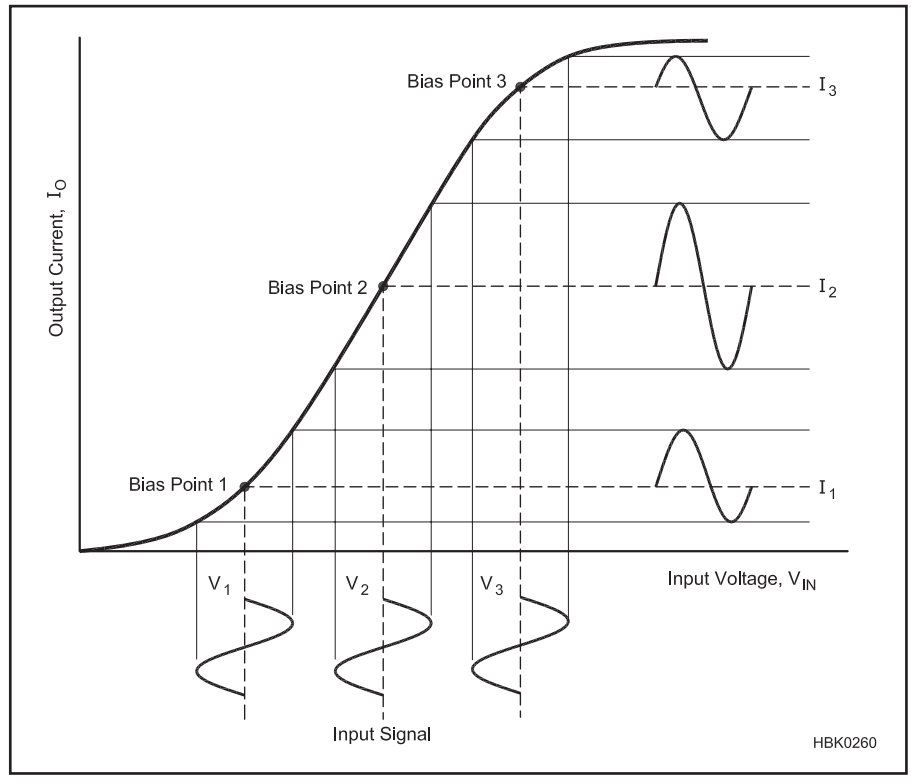

Fig 3.6 — Effect of biasing. An input signal may be reproduced linearly or nonlinearly depending on the choice of bias points.

usually three values listed in the data sheet for each parameter: 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 might contain a minimum Gain-Bandwidth Product (f_T or GBW), maximum output capacitance, maximum input capacitance, and the range of the transfer parameters applicable to a given device. Finally, the SWITCHING CHARACTERISTICS section might list 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 *cross-references* and *substitution guides* 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.

3.2.5 Physical Electronics of Semiconductors

In a conductor, such as a metal, some of the outer, or *valence*, electrons of each atom are free to move about between atoms. These *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 and the concentration of free electrons is very small — on the order of 10 electrons/cm^3.

Between the classes of materials considered to be conductors and insulators is a class of elements called *semiconductors*, materials with conductivity much poorer than metals and much better than insulators. (In electronics, "semiconductor" means a device made from semiconductor elements that have been chemically manipulated as described below, leading to interesting properties that create useful applications.)

Semiconductor atoms (silicon, Si, is the most widely used) share their valence electrons in a chemical bond that holds adjacent atoms together, forming a three-dimensional *lattice* that gives the material its physical characteristics. A lattice of pure semiconductor material (one type of atom or molecule) can form a crystal, in which the lattice structure and orientation is preserved throughout the material. *Monocrystalline* or "single-crystal" is the type of material used in electronic semiconductor devices. *Polycrystalline* material is made of up many smaller crystals with their own individual lattice orientations.

Crystals of pure semiconductor material are called *intrinsic* semiconductors. When energy, generally in the form of heat, is added to a semiconductor crystal lattice, some electrons are liberated from their bonds and move freely throughout the lattice. The bond that loses an electron is then unbalanced and the space that the electron came from is referred to as a *hole*. In these materials the number of free electrons is equal to the number of holes.

Electrons from adjacent bonds can leave their positions to fill the holes, thus leaving behind a hole in their old location. As a consequence of the electron moving, 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 a conductor has free electrons that flow independently from the bonds of the crystalline lattice, the current in a pure semiconductor is constrained to move from bond to bond.

Impurities can be added to intrinsic semiconductors (by a process called *doping*) to enhance the formation of electrons or holes and thus improve conductivity. These materials are *extrinsic* semiconductors. Since the additional electrons and holes can move, their movement is current and they are called *carriers*. The type of carrier that predominates in the material is called the *majority carrier*. In N-type material the majority carriers are electrons and in P-type material, holes.

There are two types of impurities that can be added: a *donor impurity* with five valence electrons donates free electrons to the crystalline structure; this is called an *N-type* impurity, for the negative charge of the majority carriers. Some examples of donor impurities are antimony (Sb), phosphorus (P) and arsenic (As). N-type extrinsic semiconductors have more electrons and fewer holes than intrinsic semiconductors. *Acceptor 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 positive charge of the majority carriers; some examples are boron (B), gallium (Ga) and indium (In).

It is important to note that even though N-type and P-type material have different numbers of holes and free electrons than intrinsic material, they are still electrically neutral. When an electron leaves an atom, the positively-charged atom that remains in place in the crystal lattice electrically balances the roaming free electron. Similarly, an atom gaining an electron acquires a negative charge that balances the positively-charged atom it left. At no time does the material acquire a net electrical charge, positive or negative.

Compound 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.

Impurities are introduced into intrinsic semiconductors by diffusion, the same physical process that lets you smell cookies baking from several rooms away. (Molecules diffuse through air much faster than through solids.) Rates of diffusion are proportional to temperature, so 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. If an electronic device made from a structure of N- and P-type materials is raised to a high temperature, such as by excessive current, the impurities can again migrate and the internal structure of the device may be destroyed. The maximum operating temperature for semiconductor devices is specified at a level low enough to limit additional impurity diffusion.

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 include: 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.

3.2.6 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 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 presence of the positive charges (holes) in the P-type material attracts free electrons from the N-type material immediately across the junction. The opposite is true in the N-type material.

These attractions lead to diffusion of some of the majority carriers across the junction, which combine with and neutralize the majority carriers immediately on the other side (a process called *recombination*). As distance from the junction increases, the attraction quickly becomes too small to cause the carriers to move. The region close to the junction is then *depleted* of carriers, and so is named the *depletion region* (also 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 (the *cathode*) is placed at a more negative voltage than the P-type material (the *anode*), 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. This *forward bias* forces the majority carriers toward the junction where recombination occurs with the opposite type of majority carrier. The source of voltage supplies replacement electrons to the N-type material and removes electrons from the P-type material so that the majority carriers are continually replenished. Thus, the net effect is a *forward current* flowing through the semiconductor, across the PN junction. The *forward resistance* of a diode conducting current is typically very low and varies with the amount of forward current.

When the polarity is reversed, majority carriers are attracted away from the junction, not toward it. Very little current flows across the PN junction — called *reverse leakage current* — in this case. Allowing only unidirectional current flow is what allows a semiconductor diode to act as rectifier.

3.2.7 Junction Semiconductors

Semiconductor devices that operate using the principles of a PN junction are called *junction semiconductors*. These devices can have one or several junctions. The properties of junction semiconductors can be tightly controlled by the characteristics of the materials used and the size and shape of the junctions.

SEMICONDUCTOR DIODES

Diodes are commonly made of silicon and occasionally germanium. Although they act similarly, they have slightly different characteristics. The *junction threshold voltage*, or *junction barrier voltage,* is the forward bias voltage (V_F) 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 3.7**, V_F 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 voltage for germanium diodes is typically 0.3 V. Reverse leakage current is much lower for silicon diodes than for germanium diodes.

The characteristic curve for a semiconductor diode junction is given by the following equation (slightly simplified) called the *Fundamental Diode Equation* because it describes the behavior of all semiconductor PN junctions.

$$I = I_S \left(e^{\frac{V}{\eta V_t}} - 1 \right) \qquad (4)$$

where

I = diode current
V = diode voltage
I_s = reverse-bias saturation current
V_t = kT/q, the thermal equivalent of voltage (about 25 mV at room temperature)
η = emission coefficient.

The value of I_s varies with the type of semiconductor material. η also varies from 1 to 2 with the type of material and the method of fabrication. (η is close to 2 for silicon at normal current levels, decreasing to 1 at high currents.) This curve is shown in **Fig 3.8B**.

The obvious differences between Fig 3.8A and B are that the semiconductor diode has a finite *turn-on* voltage — it requires a small but nonzero forward bias voltage before it begins conducting. Furthermore, once conducting, the diode voltage continues to increase very slowly with increasing current, unlike a true short circuit. Finally, when the applied voltage is negative, the reverse current is not exactly zero but very small (microamperes). The reverse current flow rapidly reaches a level that varies little with the reverse bias voltage. This is the *reverse-bias saturation current*, I_s.

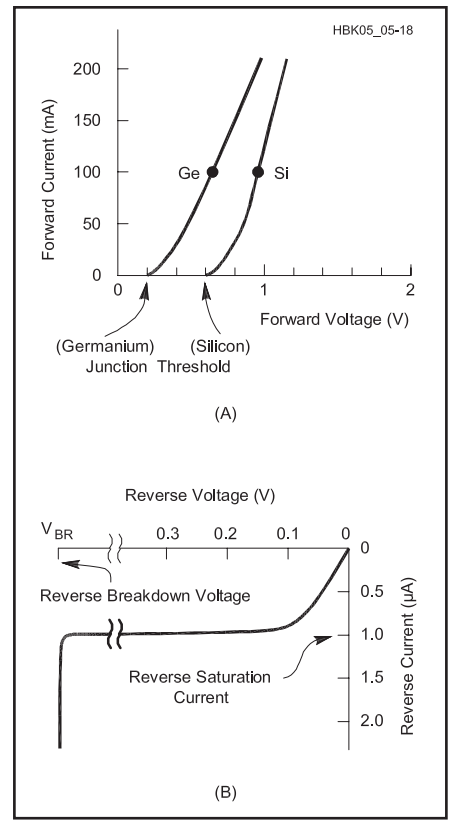

Fig 3.7 — Semiconductor diode (PN junction) characteristic curve. (A) Forward- biased (anode voltage higher than cathode) response for Germanium (Ge) and Silicon (Si) devices. Each curve breaks away from the x-axis at its junction threshold voltage. The slope of each curve is its forward resistance. (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}.

For bias (dc) circuit calculations, a useful model for the diode that takes these two effects into account is shown by the artificial I-V curve in Fig 3.8C. This model neglects the negligible reverse bias current I_s.

When converted into an equivalent circuit, the model in Fig 3.8C yields the circuit in Fig 3.8D. The ideal voltage source V_a represents the turn-on voltage and R_f represents the effective resistance caused by the small increase in diode voltage as the diode current increases. The turn-on voltage is material-dependent: approximately 0.3 V for germanium diodes and 0.7 for silicon. R_f is typically on the order of 10 Ω, but it can vary according to the specific component. R_f can often be completely neglected in comparison to the other

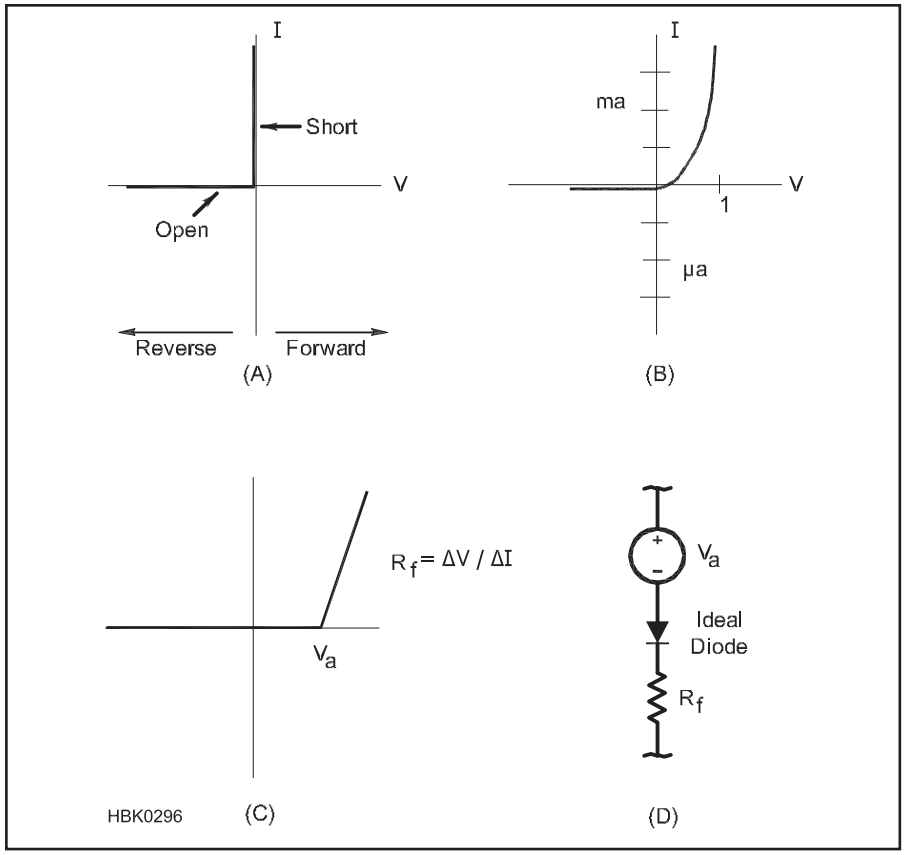

Figure 3.8 — Circuit models for rectifying switches (diodes). A: I-V curve of the ideal rectifier. B: I-V curve of a typical semiconductor diode showing the typical small leakage current in the reverse direction. Note the different scales for forward and reverse current. C shows a simplified diode I-V curve for dc-circuit calculations (at a much larger scale than B). D is an equivalent circuit for C.

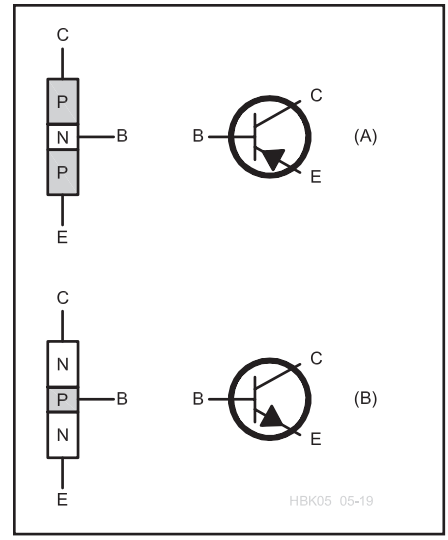

Fig 3.9 — Bipolar transistors. (A) A layer of N-type semiconductor sandwiched between two layers of P-type semiconductor makes a PNP device. The schematic symbol has three leads: collector (C), base (B) and emitter (E), with the arrow pointing in toward the base. (B) A layer of P-type semiconductor sandwiched 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.

resistances in the circuit. This very common simplification leaves only a pure *voltage drop* for the diode model.

BIPOLAR TRANSISTOR

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 3.9**).

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 — this is a representation for illustration only.)

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 become minority carriers because the materials of the emitter and base regions have opposite polarity. The excess minority carriers in the base are then attracted across the very thin base to the base-collector junction, where they are collected and are once again considered majority carriers before they can flow to the base terminal.

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 causes majority carriers to be injected into the base material (electrons flowing into an N-type base or out of a P-type base) the emitter-collector current increases. In this way, 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. As the base-emitter current continues to increase, the point is reached at which further increases in base-emitter current cause no additional change in collector current. This is the condition of *saturation*. Conversely, when base-emitter current is reduced to the point at which collector current ceases to flow, that is the situation of *cutoff*.

THYRISTORS

Thyristors are semiconductors made with four or more alternating layers of P- and N-type semiconductor material. In a four-layer thyristor, when the anode is at a higher potential than the cathode, the first and third junctions are forward biased and the center junction reverse biased. In this state, there is little current, just as in the reverse-biased diode. The different types of thyristor have different ways in which they turn on to conduct current and in how they turn off to interrupt current flow.

PNPN Diode

The simplest thyristor is a PNPN (usually pronounced like *pinpin*) diode with three junctions (see **Fig 3.10**). 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.

Bilateral Diode Switch (Diac)

A semiconductor device similar to two

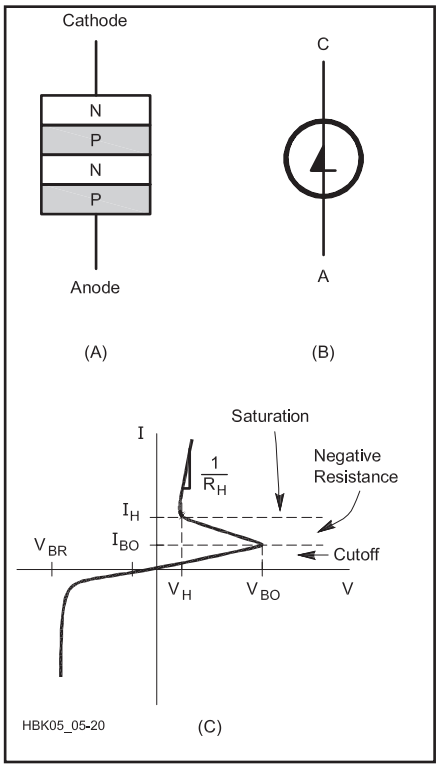

Fig 3.10 — PNPN diode. (A) Alternating layers of P-type and N-type semiconductor. (B) Schematic symbol with cathode (C) and anode (A) leads. (C) I-V curve. Reverse-biased response is the same as normal PN junction diodes. Forward biased response acts as a hysteresis switch. Resistance is very high until the bias voltage reaches V_{BO} (where the center junction breaks over) and exceeds the cutoff current, I_{BO}. The device exhibits a negative resistance when 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.

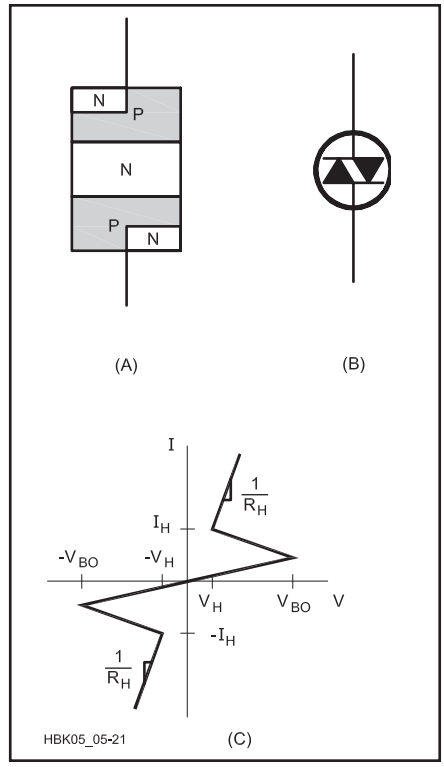

Fig 3.11 — Bilateral switch. (A) Alternating layers of P-type and N-type semiconductor. (B) Schematic symbol. (C) I-V curve. The right-hand side of the curve is identical to the PNPN diode response in Fig 3.10. The device responds identically for both forward and reverse bias so the left-hand side of the curve is symmetrical with the right-hand side.

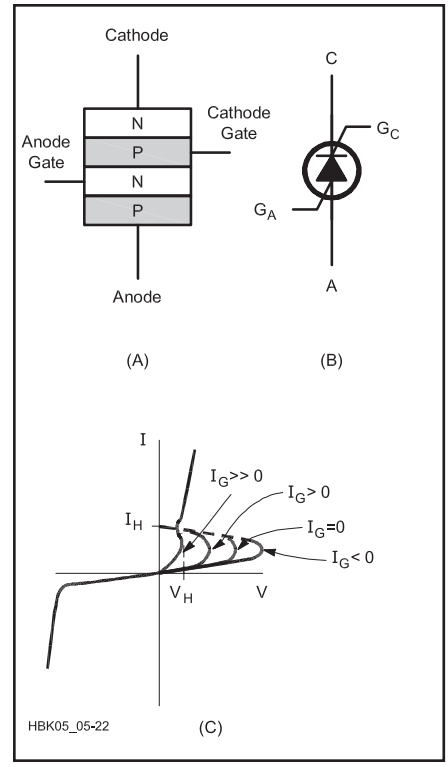

Fig 3.12 — SCR. (A) Alternating layers of P-type and N-type semiconductor. This is similar to a PNPN diode with gate terminals attached to the interior layers. (B) Schematic symbol with anode (A), cathode (C), anode gate (G_A) and cathode gate (G_C). Many devices are constructed without G_A. (C) I-V curve with different responses for various gate currents. I_G = 0 has a similar response to the PNPN diode.

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 3.11**.

Silicon Controlled Rectifier (SCR)

Another device with four alternate layers of P-type and N-type semiconductor is the *silicon controlled rectifier (SCR)*. (Some sources refer to an SCR as a thyristor, as well.) 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 3.12**).

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.

Triac

A five-layered semiconductor whose operation is similar to a bidirectional SCR is the *triac* (**Fig 3.13**). 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.

Thyristor Applications

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. Because of their high current-handling capability, SCRs are used as "crowbars" in power supply circuits, to short the output to ground and blow a fuse when an overvoltage condition exists.

SCRs and triacs are often used to control 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 3.14** 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 these devices turn on are common sources of RF interference. (See the chapter on **RF Interference** for information on dealing with interference from thyristors.)

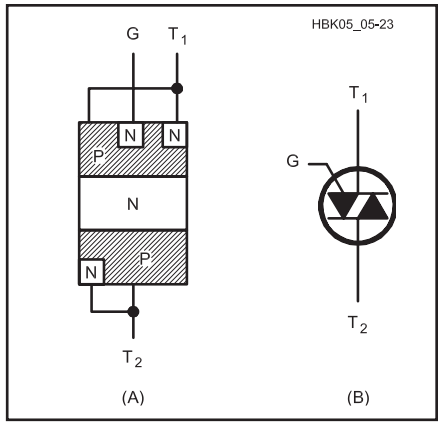

Fig 3.13 — Triac. (A) Alternating layers of P-type and N-type semiconductor. This behaves as two SCR devices facing in opposite directions with the anode of one connected to the cathode of the other and the cathode gates connected together. (B) Schematic symbol.

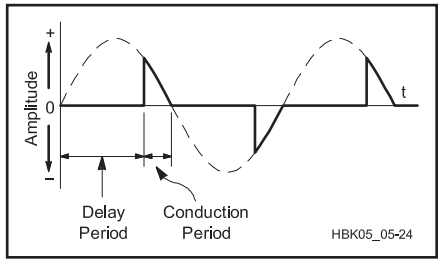

Fig 3.14 — Triac operation on sine wave. The dashed line is the original sine wave and the solid line is the portion that conducts through the triac. The relative delay and conduction period times are controlled by the amount or timing of gate current, I_G. The response of an SCR is the same as this for positive voltages (above the x-axis) and with no conduction for negative voltages.

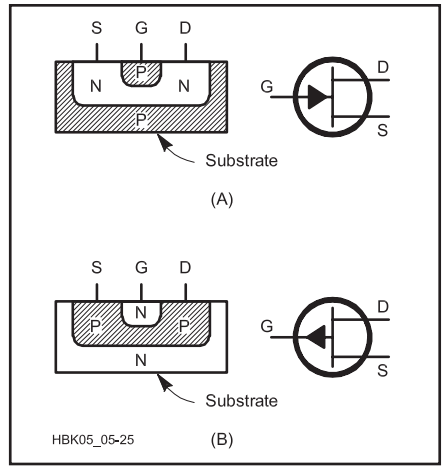

Fig 3.15 — JFET devices with terminals labeled: source (S), gate (G) and drain (D). A) Pictorial of N-type channel embedded in P-type substrate and schematic symbol. B) P-channel embedded in N-type substrate and schematic symbol.

3.2.8 Field-Effect Transistors (FET)

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. 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 technique is usually only used to make transistors, although FETs are also available as special-purpose diodes, for use as constant current sources.

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.

Within the FET, current moves in a *channel* as shown in **Fig 3.15**. 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. Current flows from the *source* terminal (where majority carriers are injected) to the *drain* terminal (where majority carriers are removed). A *gate* terminal generates an electric field that controls the current in the channel.

In N-channel devices, the drain potential must be higher than that of the source ($V_{DS} > 0$) for electrons (the majority carriers) to flow in channel. 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. 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, causing the input impedance to be quite low at high frequencies.

The current through an FET only has to pass through a single type of semiconductor material. Depending on the type of material and the construction of the FET, drain-source resistance when the FET is conducting ($r_{DS(ON)}$) may be anywhere from a few hundred ohms to much 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.

In order to achieve a higher gain-bandwidth product, other materials have been used. Gallium-arsenide (GaAs) has *electron mobility* and *drift velocity* (both are measures of how easily electrons are able to move through the crystal lattice) far higher than the standard doped silicon. Amplifiers designed with GaAsFET devices operate at much higher frequencies and with a lower noise factor at VHF and UHF than those made with silicon FETs (although silicon FETs have improved dramatically in recent years).

JFET

One of two basic types of FET, the *junction FET (JFET)* 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 with the gate material in direct contact with the channel. JFETs are used with the junction reverse-biased, since any current in the gate is undesirable. The reverse bias of the junction creates an electric field that "pinches" the channel. Since the magnitude of the electric field is proportional to the reverse-bias voltage, the current in the channel is reduced for higher reverse gate bias voltages. When current in the channel is completely halted by the electric field, this is called *pinch-off* and it is analogous to cutoff in a bipolar transistor. The channel in a JFET is at its maximum conductivity when the gate and source voltages are equal ($V_{GS} = 0$).

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 prohibited condition is for $V_{GS} > 0$. For P-channel JFETs these conditions are reversed (in normal operation $V_{GS} > 0$ and the prohibited condition is for $V_{GS} < 0$).

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 3.16**). The bias voltage on the gate terminal either attracts or repels the majority carriers of the substrate across its PN-junction with the channel. This narrows (*depletes*) or widens (*enhances*) the channel, respectively, as V_{GS} changes polarity. For example, in the N-channel enhancement-mode MOSFET, 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 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 quiescent state, $V_{GS} = 0$.

A *depletion mode* device (also called a *normally-on MOSFET*) has a channel in the quiescent state that gets smaller as a reverse bias is applied; this device conducts current with no bias applied (see Fig 3.16A 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 temporary channel that conducts current (see Fig 3.16C and D).

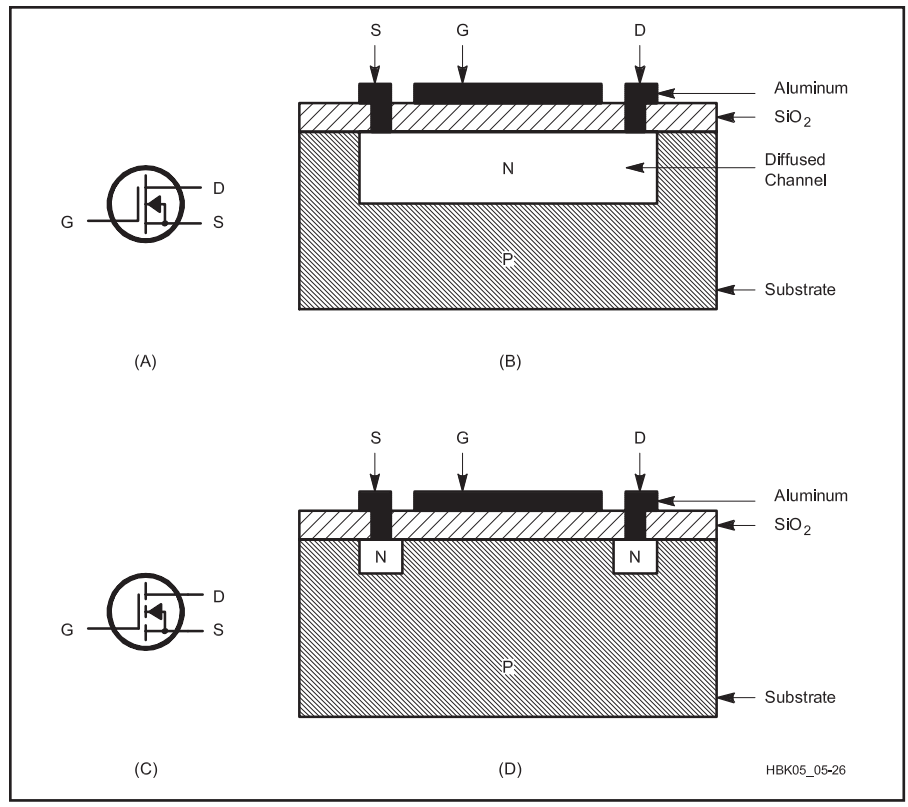

Fig 3.16 — MOSFET devices with terminals labeled: source (S), gate (G) and drain (D). N-channel devices are pictured. P-channel devices have the arrows reversed in the schematic symbols and the opposite type semiconductor material for each of the layers. (A) N-channel depletion mode device schematic symbol and (B) 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. (C) N-channel enhancement mode device schematic and (D) 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 by repelling the P-carriers away from the channel region in the substrate.

Complementary Metal Oxide Semiconductors (CMOS)

Power dissipation in a circuit can be reduced to very small levels (on the order of a few nanowatts) by using MOSFET devices in complementary pairs (CMOS). Each amplifier is constructed of a series circuit of MOSFET devices, as in **Fig 3.17**. 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 $V_{DD} \times I_{D(OFF)}$. Power is only consumed during the switching process, so for ac signals, power consumption is proportional to frequency.

CMOS circuitry could be built with discrete components, but the number of extra parts and the need for the complementary components to be matched has made that an unusual design technique. The low power consumption and ease of fabrication has made CMOS the most

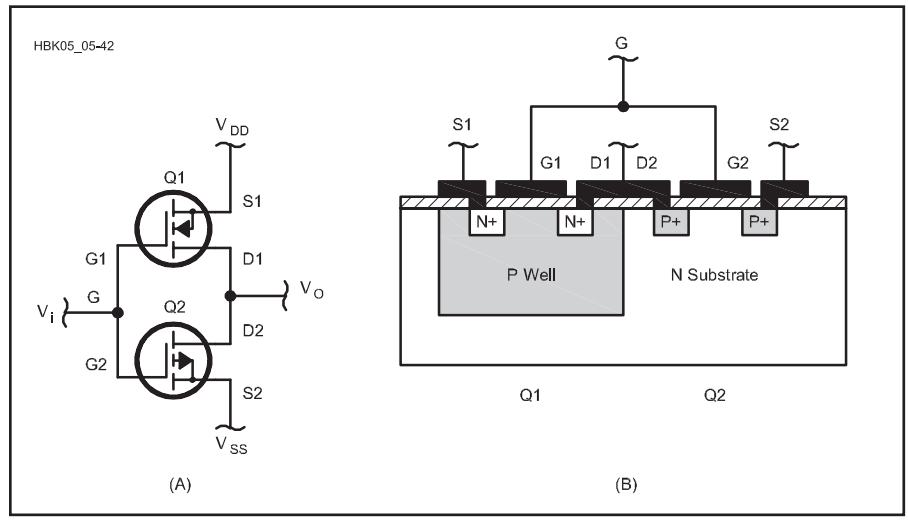

Fig 3.17 — Complementary metal oxide semiconductor (CMOS). (A) CMOS device is made from a pair of enhancement mode MOS transistors. The upper is an N-channel device, and the lower is a P-channel device. When one transistor is biased on, the other is biased off; therefore, there is minimal current from V_{DD} to ground. (B) Implementation of a CMOS pair as an integrated circuit.

common of all IC technologies. Although CMOS is most commonly used in digital integrated circuitry, its low power consumption has also been put to work by manufacturers of analog ICs, as well as digital ICs.

3.2.9 Semiconductor Temperature Effects

The number of excess holes and electrons in semiconductor material 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% per °C and by about 6% per °C for germanium. Semiconductor temperature properties are the opposite of most metals, which *increase* their resistance by about 0.4% per °C. These opposing temperature characteristics permit the design of circuits with opposite temperature coefficients that cancel each other out, making a temperature insensitive circuit.

Semiconductor devices can experience an effect called *thermal runaway* as the current causes an increase in temperature. (This is primarily an issue with bipolar transistors.) 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, so circuit design must include measures that compensate for the effects of temperature.

Semiconductor Failure Caused by Heat

There are several common failure modes for semiconductors that are related to heat. The semiconductor material is connected to the outside world through metallic *bonding leads*. The point at which the lead and the semiconductor are connected is a 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 powered on all the time, but this requires removal of the heat generated during normal operation.

A common failure mode of semiconductors is caused by the heat generated during semiconductor use. If the temperatures of the PN junctions remain at high enough levels for long enough periods of time, the impurities resume their diffusion across the PN junctions. When enough of the impurity atoms cross the depletion region, majority carrier recombination stops functioning properly and the semiconductor device fails permanently.

Excessive temperature can also cause failure anywhere in the semiconductor from heat generation within any current-carrying conductor, such as an FET channel or the bonding leads. Integrated circuits with more than one output may have power dissipation limits that depend on how many of the outputs are active at one time. The high temperature can cause localized melting or cracking of the semiconductor material, causing a permanent failure.

Another heat-driven failure mode, usually not fatal to the semiconductor, is excessive leakage current or a shift in operating point that causes the circuit to operate improperly. This is a particular problem in complex integrated circuits — analog and digital — dissipating significant amounts of heat under normal operating conditions. Computer microprocessors are a good example, often requiring their own cooling systems. Once the device cools, normal operation is usually restored.

To reduce the risk of thermal failures, the designer must comply with the limits stated in the manufacturer's data sheet, devising an adequate heat removal system. (Thermal issues are discussed in the **Electrical Fundamentals** chapter.)

3.2.10 Safe Operating Area (SOA)

Devices intended for use in circuits handling high currents or voltages are specified to have a *safe operating area* (*SOA*). This refers to the area drawn on the device's characteristic curve containing combinations of voltage and current that the device can be expected to control without damage under specific conditions. The SOA combines a number of limits — voltage, current, power, temperature and various breakdown mechanisms — in order to simplify the design of protective circuitry. The SOA is also specified to apply to specific durations of use — steady-state, long pulses, short pulses and so forth. The device may have separate SOAs for resistive and inductive loads.

You may also encounter two specialized types of SOA for turning the device on and off. *Reverse bias safe operating area* (*RBSOA*) applies when the device is turning off. *Forward bias safe operating* area (*FBSOA*) applies when turning the device on. These SOAs are used because the high rate-of-change of current and voltage places additional stresses on the semiconductor.

3.3 Practical Semiconductors

3.3.1 Semiconductor Diodes

Although many types of semiconductor diodes are available, they share many common characteristics. The different types of diodes have been developed to optimize particular characteristics for one type of application. You will find many examples of diode applications throughout this book.

The diode symbol is shown in **Fig 3.18**. Forward current flows in the direction from anode to cathode, in the direction of the arrow. Reverse current flows from cathode to anode. (Current is considered to be conventional current as described in the **Electrical Fundamentals** chapter.) 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 3.18. Most diodes are marked with a band on the cathode end (Fig 3.18).

DIODE RATINGS

Five major characteristics distinguish standard junction diodes from one another: current handling capacity, maximum voltage rating, response speed, reverse leakage current and junction forward voltage. Each of these characteristics can be manipulated during manufacture to produce special purpose diodes.

Current Capacity

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 3.7. Note that the scales of the two graphs 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, R_F.

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). Nevertheless, this resistance causes heat dissipation according to $P = I_F^2 \times R_F$.

In addition, there is a forward voltage, V_F, whenever the forward current is flowing. This

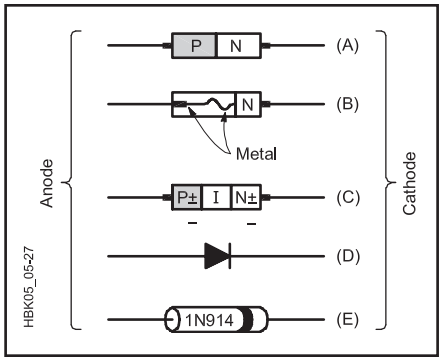

Fig 3.18 — Practical semiconductor diodes. All devices are aligned with anode on the left and cathode on the right. (A) Standard PN junction diode. (B) Point-contact or "cat's whisker" diode. (C) PIN diode formed with heavily doped P-type (P^+), undoped (intrinsic) and heavily doped N-type (N^+) semiconductor material. (D) Diode schematic symbol. (E) Diode package with marking stripe on the cathode end.

also results in heat dissipation as $P = I \times V_F$. In power applications where the average forward current is high, heating from forward resistance and the forward voltage drop can be significant. Since forward current determines the amount of heat dissipation, the diode's power rating is stated as a *maximum average current*. Exceeding the current rating in a diode will cause excessive heating that leads to PN junction failure as described earlier.

Peak Inverse Voltage (PIV)

In Fig 3.7, 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. This sudden change occurs because the diode enters *reverse breakdown* or when the reverse voltage becomes high enough to push current across the junction. The voltage at which this occurs is the *reverse breakdown voltage*. Unless the current is so large that the diode fails from overheating, breakdown is not destructive and the diode will again behave normally when the bias is removed. The maximum reverse voltage that the diode can withstand under normal use is the *peak inverse voltage* (*PIV*) rating. A related effect is *avalanche breakdown* in which the voltage across a device is greater than its ability to control or block current flow.

Response Speed

The speed of a diode's response to a change in voltage polarity limits the frequency of ac current that the diode can rectify. The diode response in Fig 3.7 shows how that diode will act at dc. As the frequency increases, the diode may not be able to turn current on and off as fast as the changing polarity of the signal.

Diode response speed mainly depends on *charge storage* in the depletion region. When forward current is flowing, electrons and holes fill the region near the junction to recombine. When the applied voltage reverses, these excess charges move away from the junction so that no recombination can take place. As reverse bias empties the depletion region of excess charge, it begins to act like a small capacitor formed by the regions containing majority carriers on either side of the junction and the depletion region acting as the dielectric. This *junction capacitance* is inversely proportional to the width of the depletion region and directly proportional to the cross-sectional surface area of the junction.

The effect of junction capacitance is to allow current to flow for a short period after the applied voltage changes from positive to negative. To halt current flow requires that the junction capacitance be charged. Charging this capacitance takes some time; a few µs for regular rectifier diodes and a few hundred nanoseconds for *fast-recovery* diodes. This is the diode's *charge-storage time*. The amount of time required for current flow to cease is the diode's *recovery time*.

Reverse Leakage Current

Because the depletion region is very thin, reverse bias causes a small amount of reverse leakage or reverse saturation current to flow from cathode to anode. This is typically 1 µA or less until reverse breakdown voltage is reached. Silicon diodes have lower reverse leakage currents than diodes made from other materials with higher carrier mobility, such as germanium.

The reverse saturation current I_s is not constant but is affected by temperature, with higher temperatures increasing the mobility of the majority carriers so that more of them cross the depletion region for a given amount of reverse bias. For silicon diodes (and transistors) near room temperature, I_s increases by a factor of 2 every 4.8 °C. This means that for every 4.8 °C rise in temperature, either the current doubles (if the voltage across it is constant), or if the current is held constant by other resistances in the circuit, the diode voltage will *decrease* by $V_T \times \ln 2 = 18$ mV. For germanium, the current doubles every 8 °C and for gallium-arsenide (GaAs), 3.7 °C. This dependence is highly reproducible and may actually be exploited to produce temperature-measuring circuits.

While the change resulting from a rise of several degrees may be tolerable in a circuit design, that from 20 or 30 degrees may not. Therefore it's a good idea with diodes, just as with other components, to specify power ratings conservatively (2 to 4 times margin) to prevent self-heating.

While component derating does reduce self-heating effects, circuits must be designed for the expected operating environment. For example, mobile radios may face temperatures from –20° to +140 °F (–29° to 60 °C).

Forward Voltage

The amount of voltage required to cause majority carriers to enter the depletion region and recombine, creating full current flow, is called a diode's *forward voltage*, V_F. It depends on the type of material used to create the junction and the amount of current. For silicon diodes at normal currents, $V_F = 0.7$ V, and for germanium diodes, $V_F = 0.3$ V. As you saw earlier, V_F also affects power dissipation in the diode.

POINT-CONTACT DIODES

One way to decrease charge storage time in the depletion region is to form a metal-semiconductor junction for which the depletion is very thin. 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 with a steel wire in contact with a crystal of impure lead (galena). Point-contact diodes have high response speed, but poor PIV and current-handling ratings. The 1N34 germanium point-contact diode is the best-known example of point-contact diode still in common use.

SCHOTTKY DIODES

An improvement to point-contact diodes, the *hot-carrier diode* is similar to a point-contact diode, but 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 junctions of Schottky diodes, being smaller, store less charge and as a result, have shorter switching times and junction capacitances than standard PN-junction diodes. Their forward voltage is also lower, typically 0.3 to 0.4 V. In most other respects they behave similarly to PN diodes.

PIN DIODES

The PIN diode, shown in Fig 3.18C is a *slow response* diode that is capable of passing RF and 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 extra amount of doping), creating a *PIN junction*. These devices provide very effective switches for RF signals and are often used in transmit-receive switches

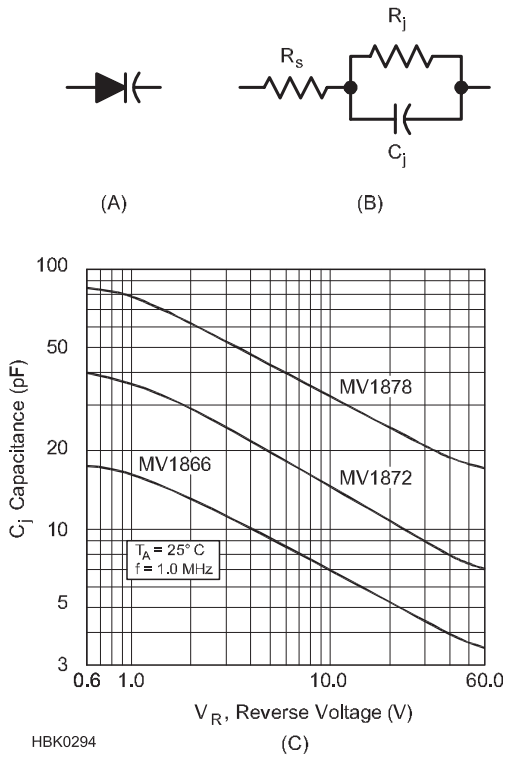

Fig 3.19 — Varactor diode. (A) Schematic symbol. (B) Equivalent circuit of the reverse biased varactor diode. R_S is the junction resistance, R_J is the leakage resistance and C_J is the junction capacitance, which is a function of the magnitude of the reverse bias voltage. (C) Plot of junction capacitance, C_J, as a function of reverse voltage, V_R, for three different varactor devices. Both axes are plotted on a logarithmic scale.

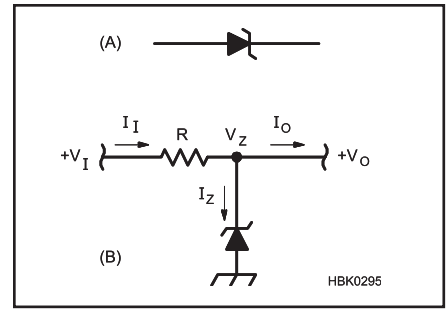

Fig 3.20 — Zener diode. (A) Schematic symbol. (B) Basic voltage regulating circuit. V_Z is the Zener reverse breakdown voltage. Above V_Z, the diode draws current until $V_I - I_I R = V_Z$. The circuit design should select R so that when the maximum current is drawn, $R < (V_I - V_Z) / I_O$. The diode should be capable of passing the same current when there is no output current drawn.

in transceivers and amplifiers. The majority carriers in PIN diodes have longer than normal lifetimes before recombination, resulting in a slow switching process that causes them to act more like resistors than diodes at high radio frequencies. The amount of resistance can be controlled by the amount of forward bias applied to the PIN diode and this allows them to act as current-controlled attenuators. (For additional discussion of PIN diodes and projects in which they are used, see the chapters on **Transmitters and Transceivers**, **RF Power Amplifiers**, and **Test Equipment and Measurements**.)

VARACTOR DIODES

Junction capacitance can be used as a circuit element by controlling the reverse bias voltage across the junction, creating a small variable capacitor. Junction capacitances are small, on the order of pF. As the reverse bias voltage on a diode increases, the width of the depletion region increases, decreasing its capacitance. A *varactor* (also known by the trade name Varicap diode) is a diode with a junction specially formulated to have a relatively large range of capacitance values for a modest range of reverse bias voltages (**Fig 3.19**).

As the reverse bias applied to a diode changes, the width of the depletion layer, and therefore the capacitance, also changes. The diode junction capacitance (C_j) under a reverse bias of V volts is given by

$$C_j = \frac{C_{j0}}{\sqrt{V_{on} - V}} \quad (5)$$

where C_{j0} = measured capacitance with zero applied voltage.

Note that the quantity under the radical is a large *positive* quantity for reverse bias. As seen from the equation, for large reverse biases C_j is inversely proportional to the square root of the voltage.

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. For example, if the varactor is used to tune an oscillator, 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. (For examples of the use of varactors in oscillators and modulators, see the chapters on **Mixers, Modulators, and Demodulators** and **Oscillators and Synthesizers**.)

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 will not destroy the diode if it is limited to less than the device's maximum allowable value. By using heavy levels of doping during manufacture, a diode's PIV can be precisely controlled to be at a specific level, called the *Zener voltage*, creating a type of voltage reference. These diodes are called Zener diodes after their inventor, American physicist Clarence Zener.

When the Zener voltage is reached, the reverse voltage across the Zener diode remains constant even as the current through it changes. With an appropriate series current-limiting resistor, the Zener diode provides an accurate voltage reference (see **Fig 3.20**).

Zener diodes are rated by their reverse-breakdown voltage and their power-handling capacity, where $P = V_Z \times I_Z$. 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 Sources** chapter for more information about using Zener diodes as voltage regulators.) When operating in the breakdown region, Zener diodes can be modeled as a simple voltage source.

The primary 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 5 parts per million per °C. If this is unacceptable, voltage reference integrated circuits based on Zener diodes have been developed

that include additional circuitry to counteract temperature effects.

A variation of Zener diodes, *transient voltage suppressor* (*TVS*) diodes are designed to dissipate the energy in short-duration, high-voltage transients that would otherwise damage equipment or circuits. TVS diodes have large junction cross-sections so that they can handle large currents without damage. These diodes are also known as TransZorbs. Since the polarity of the transient can be positive, negative, or both, transient protection circuits can be designed with two devices connected with opposite polarities.

RECTIFIERS

The most common application of a diode is to perform rectification; that is, permitting current flow in only one direction. Power rectification converts ac current into pulsating dc current. There are three basic forms of power rectification using semiconductor diodes: half wave (1 diode), full-wave center-tapped (2 diodes) and full-wave bridge (4 diodes). These applications are shown in **Fig 3.21A**, B and C and are more fully described in the **Power Sources** chapter.

The most important diode parameters to consider for power rectification are the PIV and current ratings. The peak negative voltages that are blocked 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 average forward current.

Rectification is also used at much lower current levels in modulation and demodulation and other types of analog signal processing circuits. For these applications, the diode's response speed and junction forward voltage are the most important ratings.

3.3.2 Bipolar Junction Transistors (BJT)

The bipolar transistor is a *current-controlled device* with three basic terminals; *emitter*, *collector* and *base*. The current between the emitter and the collector is controlled by the current between the base and emitter. 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 (see the **Electrical Fundamentals** chapter) 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

$$I_c + I_b + I_e = 0 \tag{6}$$

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

$$I_e = -(I_c + I_b) \tag{7}$$

The back-to-back diode model shown in Fig 3.9 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.

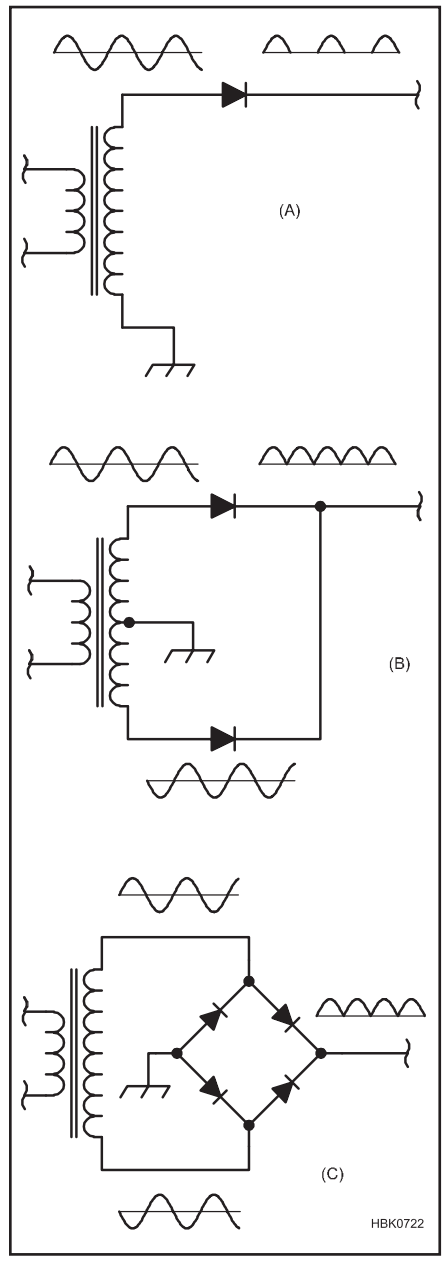

Fig 3.21 — Diode rectifier circuits. (A) Half wave rectifier circuit. Only when the ac voltage is positive does current pass through the diode. Current flows only during half of the cycle. (B) Full-wave center-tapped rectifier circuit. Center-tap on the transformer secondary is grounded and the two ends of the secondary are 180° out of phase. (C) Full-wave bridge rectifier circuit. In each half of the cycle two diodes conduct.

The operation of the bipolar transistor is described graphically by characteristic curves as shown in **Fig 3.22**. These are similar to the I-V characteristic curves for the two-terminal devices described in the preceding sections. The parameters shown by the curves depend on the type of circuit in which they are measured, such as common emitter or common collector. The output characteristic shows a set of curves for either collector or emitter current versus collector-emitter voltage at various values of input current (either base or emitter). The input characteristic shows the voltage between the input and common terminals (such as base-emitter) versus the input current for different values of output voltage.

CURRENT GAIN

Two parameters describe the relationships between the three transistor currents at low frequencies:

$$\alpha \approx -\frac{\Delta I_C}{\Delta I_E} \approx 1 \tag{8}$$

$$\beta = \frac{\Delta I_C}{\Delta I_B} \tag{9}$$

The relationship between α and β is defined as

$$\alpha = -\frac{\beta}{1+\beta} \tag{10}$$

Another designation for β is often used: h_{FE}, the *forward dc current gain*. (The "h" refers to "h parameters," a set of transfer parameters for describing a two-port network and described in more detail in the **RF Techniques** chapter.) The symbol, h_{fe}, in which the subscript is in lower case, is used for the forward current gain of ac signals.

OPERATING REGIONS

Current conduction between collector and emitter is described by *regions* of the transistor's characteristic curves in Fig 3.22. (References such as *common-emitter* or *common-base* refer to the configuration of the circuit in which the parameter is measured.) The transistor is in its *active* or *linear region* when the base-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. In this region, the output circuit of the transistor can be modeled as a constant-current source controlled by

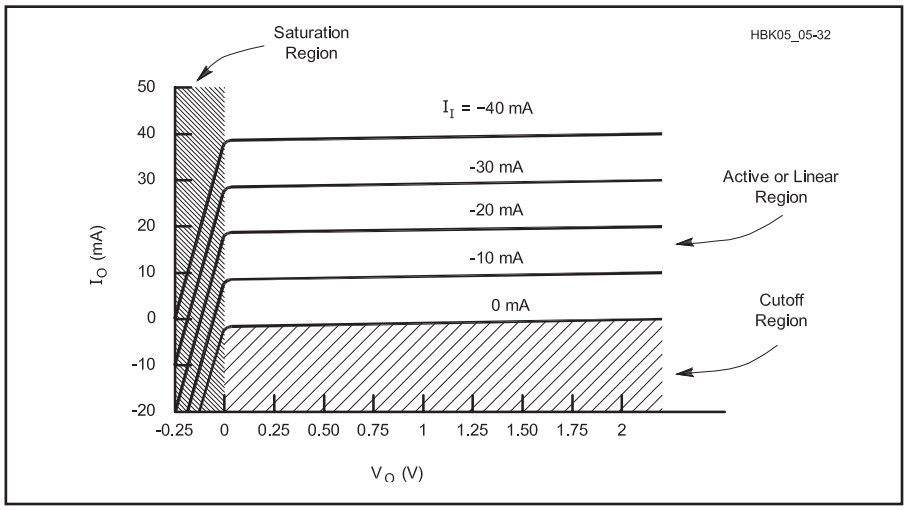

Fig 3.22 — Transistor response curve output characteristics. The x-axis is the output voltage, and the y-axis is the output current. Different curves are plotted for various values of input current. The three regions of the transistor are its cutoff region, where no current flows in any terminal, its 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.

the input current. The slight slope that does exist is due to base-width modulation (known as the *Early effect*).

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. Both junctions in the transistor are reverse-biased in the *cutoff region*. 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. Finally, if V_O is increased to very high values, avalanche breakdown begins as in a PN-junction diode and output current increases rapidly. This is the *breakdown region*, not shown in Fig 3.22.

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.

OPERATING PARAMETERS

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. Exceeding any of these parameters is likely to cause the transistor to be destroyed. (The "O" in the suffixes of the junction voltages indicates that the remaining terminal is not connected, or open.)

In the OPERATING PARAMETERS section, three guaranteed minimum junction breakdown voltages are listed $V_{(BR)CEO}$, $V_{(BR)CBO}$ and $V_{(BR)EBO}$. Exceeding these voltages is likely to cause the transistor to enter avalanche breakdown, but if current is limited, permanent damage may not result.

Under ON CHARACTERISTICS are the guaranteed minimum dc current gain (β or h_{FE}), guaranteed maximum collector-emitter saturation voltage, $V_{CE(SAT)}$, and the guaranteed maximum base-emitter on voltage, $V_{BE(ON)}$. Two guaranteed maximum collector cutoff currents, I_{CEO} and I_{CBO}, are listed under OFF CHARACTERISTICS.

The next section is SMALL-SIGNAL CHARACTERISTICS, where the guaranteed minimum current gain-bandwidth product, BW or 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.

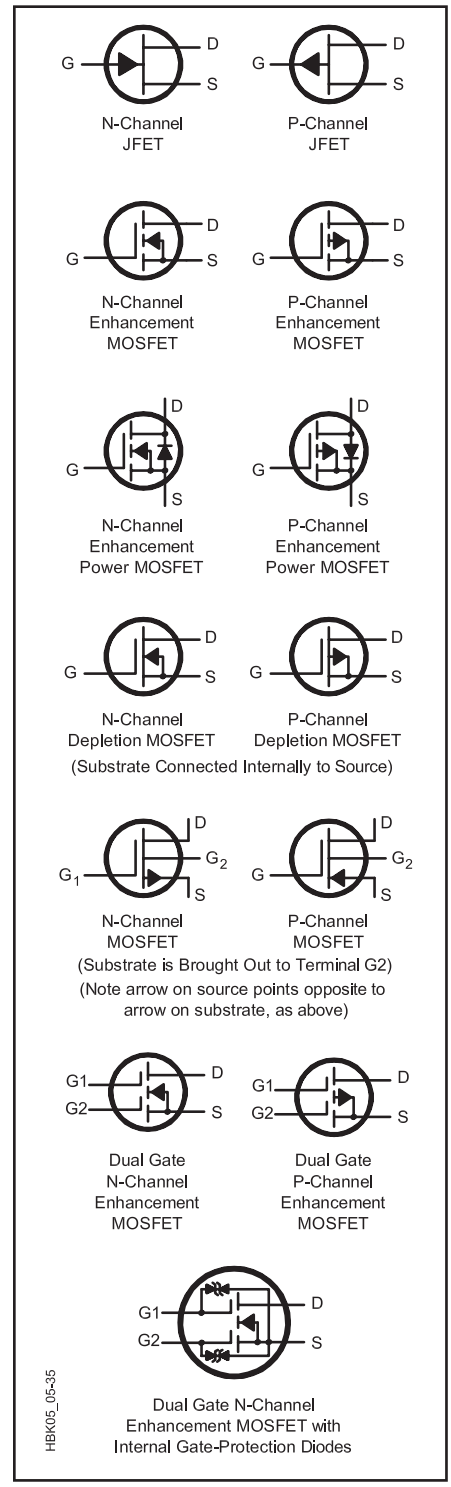

Fig 3.23—FET schematic symbols.

3.3.3 Field-Effect Transistors (FET)

FET devices 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*. They are analogous to bipolar transistor terminals: the gate to the base, the source to the emitter, and the drain to the collector. Symbols for the various forms of FET devices are pictured in **Fig 3.23**.

The FET gate has a very high impedance, so the input can be modeled as an open circuit. The voltage between gate and source, V_{GS}, controls the resistance of the drain-source

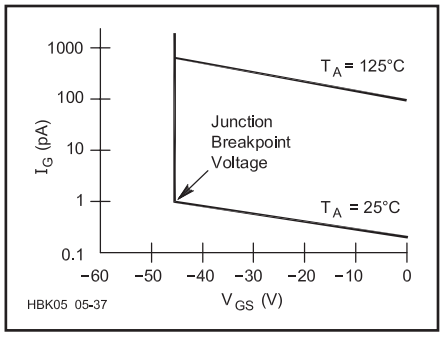

Fig 3.24 — JFET input leakage curves for common source amplifier configuration. Input voltage (V_{GS}) on the x-axis versus input current (I_G) on the y-axis, with two curves plotted for different operating temperatures, 25 °C and 125 °C. Input current increases greatly when the gate voltage exceeds the junction breakpoint voltage.

channel, r_{DS}, and so the output of the FET is modeled as a current source, whose output current is controlled by the input voltage.

The action of the FET channel is so nearly ideal that, as long as the JFET gate does not become forward biased and inject current from the base into the channel, 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* (see **Fig 3.24**). The point at which there is a significant increase in I_G is called the *junction breakpoint voltage*. Because the gate of MOSFETs is insulated from the channel, gate leakage current is insignificant in these devices.

The dc channel resistance, r_{DS}, is specified in data sheets to be less than a maximum value when the device is biased on ($r_{DS(on)}$). When the 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. Channel resistance is approximately the same for ac and dc signals until at high frequencies the capacitive reactances inherent in the FET structure become significant.

FETs also have strong similarities to vacuum tubes in that input voltage between the grid and cathode controls an output current between the plate and cathode. (See the chapter on **RF Power Amplifiers** for more information on vacuum tubes.)

FORWARD TRANSCONDUCTANCE

The change in FET drain current caused by a change in gate-to-source voltage is called *forward transconductance, g_m.*

$$g_m = \frac{\Delta I_{DS}}{\Delta V_{GS}}$$

or

$$\Delta I_{DS} = g_m \Delta V_{GS} \qquad (11)$$

The input voltage, $V_{GS,}$ is measured between the FET gate and source and drain current, I_{DS}, flows from drain to source. Analogous to a bipolar transistor's current gain, the units of transconductance are Siemens (S) because it is the ratio of current to voltage. (Both g_m and g_{fs} are used interchangeably to indicate transconductance. Some sources specify g_{fs} as the *common-source forward transconductance*. This chapter uses g_m, the most common convention in the reference literature.)

OPERATING REGIONS

The most useful relationships for FETs are the output and transconductance response characteristic curves in **Fig 3.25**. (References such as *common-source* or *common-gate* refer to the configuration of the circuit in which the parameter is measured.) Transconductance curves relate the drain current, I_D, to gate-to-source voltage, V_{GS}, at various drain-source voltages, V_{DS}. The FET's forward transconductance, g_m, is the slope of the lines in the forward transconductance curve. The same parameters are interrelated in a different way in the output characteristic, in which I_D is shown versus V_{DS} for different values of V_{GS}.

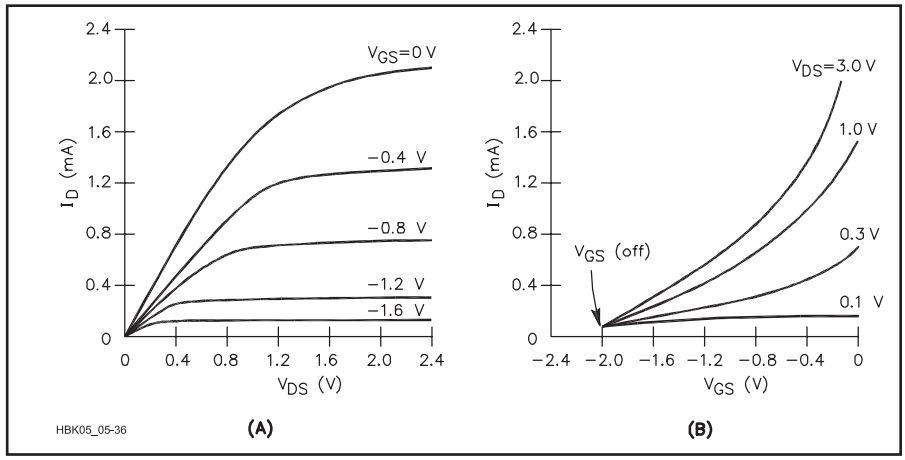

Fig 3.25 — JFET output and transconductance response curves for common source amplifier configuration. (A) Output voltage (V_{DS}) on the x-axis versus output current (I_D) on the y-axis, with different curves plotted for various values of input voltage (V_{GS}). (B) Transconductance curve with the same three variables rearranged: V_{GS} on the x-axis, I_D on the y-axis and curves plotted for different values of V_{DS}.

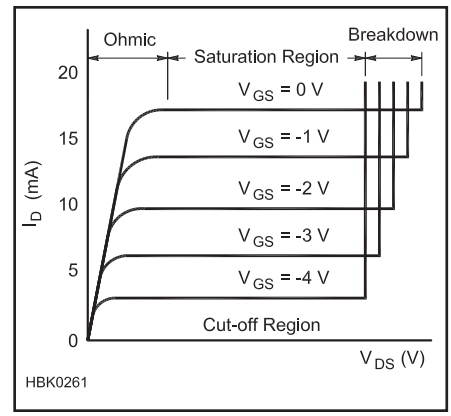

Fig 3.26 — JFET operating regions. At the left, I_D is increasing rapidly with V_{GS} and the JFET can be treated as resistance (R_{DS}) controlled by V_{GS}. In the saturation region, drain current, I_D, is relatively independent of V_{GS}. As V_{DS} increases further, avalanche breakdown begins and I_D increases rapidly.

Like the bipolar transistor, FET operation can be characterized by regions. The *ohmic region* is shown at the left of the FET output characteristic curve in **Fig 3.26** where I_D is increasing nearly linearly with V_{DS} and the FET is acting like a resistance controlled by V_{GS}. As V_{DS} continues to increase, I_D saturates and becomes nearly constant. This is the FET's *saturation region* in which the channel of the FET can be modeled as a constant-current source. V_{DS} can become so large that V_{GS} no longer controls the conduction of the device and avalanche breakdown occurs as in bipolar transistors and PN-junction diodes. This is the *breakdown region*, shown in Fig 3.26 where the curves for I_D break sharply upward. If V_{GS} is less than V_P, so that transconductance is zero, the FET is in the *cutoff region*.

OPERATING PARAMETERS

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. Exceeding these limits usually results in destruction of the FET.

Under OPERATING PARAMTERS the OFF CHARACTERISTICS list the gate-source breakdown voltage, $V_{GS(BR)}$, the reverse gate current, I_{GSS} and the gate-source cutoff voltage,

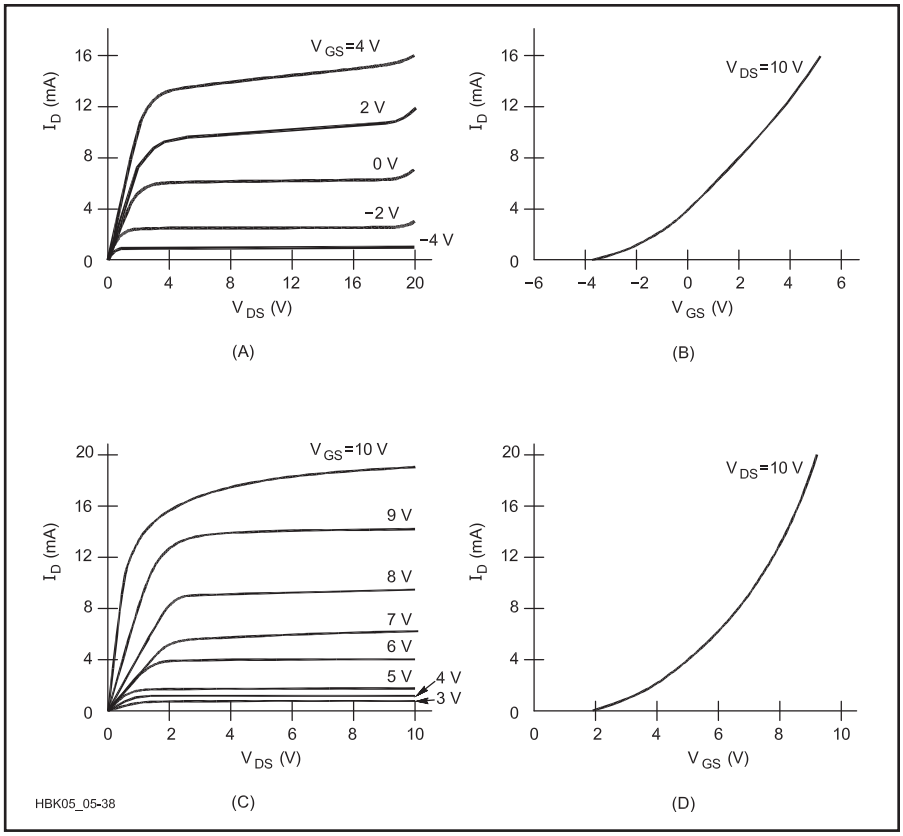

Fig 3.27 — MOSFET output [(A) and (C)] and transconductance [(B) and (D)] response curves. Plots (A) and (B) are for an N-channel depletion mode device. Note that V_{GS} varies from negative to positive values. Plots (C) and (D) are for an N-channel enhancement mode device. V_{GS} has only positive values.

$V_{GS(OFF)}$. Exceeding $V_{GS(BR)}$ will not permanently damage the device if current is limited. The primary ON CHARACTERISTIC parameters are the channel resistance, r_{DS}, and the zero-gate-voltage drain current (I_{DSS}). An FET's 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 in the FET becomes significant.

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}.

MOSFETS

As described earlier, the MOSFET's gate is insulated from the channel by a thin layer of nonconductive oxide, doing away with any appreciable gate leakage current. Because of this isolation of the gate, MOSFETs do not need input and reverse transconductance curves. Their output curves (**Fig 3.27**) are similar to those of the JFET. The gate acts as a small capacitance between the gate and both the source and drain.

The output and transconductance curves in Fig 3.27A and 3.27B show that the depletion-mode N-channel MOSFET's transconductance is positive at $V_{GS} = 0$, like that of the N-channel JFET. Unlike the JFET, however, increasing V_{GS} does not forward-bias the gate-source junction and so the device can be operated with $V_{GS} > 0$.

In the enhancement-mode MOSFET, transconductance is zero at $V_{GS} = 0$. As V_{GS} is increased, the MOSFET enters the ohmic region. If V_{GS} increases further, the saturation region is reached and the MOSFET is said to be *fully-on*, with r_{DS} at its minimum value. The behavior of the enhancement-mode MOSFET is similar to that of the bipolar transistor in this regard.

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. 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 as a mixer in which the two input signals are applied to the pair of gates.

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 transconductance of the device but this makes it susceptible to damage from high voltage levels, such as *electrostatic discharge* (ESD) from static electricity. 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 electrical discharges from handling.

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 diode junction conducts, bleeding 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, keeping the Zener reverse biased. In depletion mode devices however, V_{GS} can be both positive and negative; when negative, a gate-protection Zener diode would be forward biased and the MOSFET gate would not be driven properly. In some depletion mode MOSFETs, 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 an electrostatic discharge is capable of delivering little energy, it can generate thousands of volts and high peak currents. 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 workbench or mat. 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 little excess charge flows when the device is inserted into the circuit board.

Power MOSFETs

Power MOSFETs are designed for use as switches, with extremely low values of $r_{DS(on)}$; values of 50 milliohms (mΩ) are common. The largest devices of this type can switch tens of amps of current with V_{DS} voltage ratings of hundreds of volts. The **Component Data and References** chapter includes a table of Power FET ratings. The schematic symbol for power MOSFETs (see Fig 3.23) includes a *body diode* that allows the FET to conduct in the reverse direction, regardless of V_{GS}. This is useful in many high-power switching applications. Power MOSFETs used for RF amplifiers are discussed in more detail in the **RF Power Amplifiers** chapter.

While the maximum ratings for current and voltage are high, the devices cannot withstand both high drain current and high drain-to-source voltage at the same time because of the power dissipated; $P = V_{DS} \times I_D$. It is important to drive the gate of a power MOSFET such that the device is fully on or fully off so that either V_{DS} or I_D is at or close to zero. When switching, the device should spend as little time as possible in the linear region where both current and voltage are nonzero because their product (P) can be substantial. This is not a big problem if switching only takes place occasionally, but if the switching is repetitive (such as in a switching power supply) care should be taken to drive the gate properly and remove excess heat from the device.

Because the gate of a power MOSFET is capacitive (up to several hundred pF for large devices), charging and discharging the gate quickly results in short current peaks of more than 100 mA. Whatever circuit is used to drive the gate of a power MOSFET must be able to handle that current level, such as an integrated circuit designed for driving the capacitive load an FET gate presents.

The gate of a power MOSFET should not be left open or connected to a high-impedance circuit. Use a pull-down or pull-up resistor connected between the gate and the appropriate power supply to ensure that the gate is placed at the right voltage when not being driven by the gate drive circuit.

GaAsFETs

FETs made from gallium-arsenide (GaAs) material are used at UHF and microwave frequencies because they have gain at these frequencies and add little noise to the signal. The reason GaAsFETs have gain at these frequencies is the high mobility of the electrons in GaAs material. Because the electrons are more mobile than in silicon, they respond to the gate-source input signal more quickly and strongly than silicon FETs, providing gain at higher frequencies (f_T is directly proportional to electron mobility). The higher electron mobility also reduces thermally-generated noise generated in the FET, making the GaAsFET especially suitable for weak-signal preamps.

Because electron mobility is always higher than hole mobility, N-type material is used in GaAsFETs to maximize high-frequency gain. Since P-type material is not used to make a gate-channel junction, a metal Schottky junction is formed by depositing metal directly on the surface of the channel. This type of device is also called a *MESFET* (metal-semiconductor field-effect transistor).

3.3.4 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 photons of light to strike the surface of the semiconductor material, the energy absorbed by electrons disrupts the bonds between atoms, creating free electrons and holes. This increases the conductivity of the material (*photoconductivity*). The photon can also transfer enough energy to an electron to allow it to cross a PN junction's depletion region as current flow through the semiconductor (*photoelectricity*).

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.

PHOTODIODES

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, shown in **Fig 3.28**, 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, I_{SC}, (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

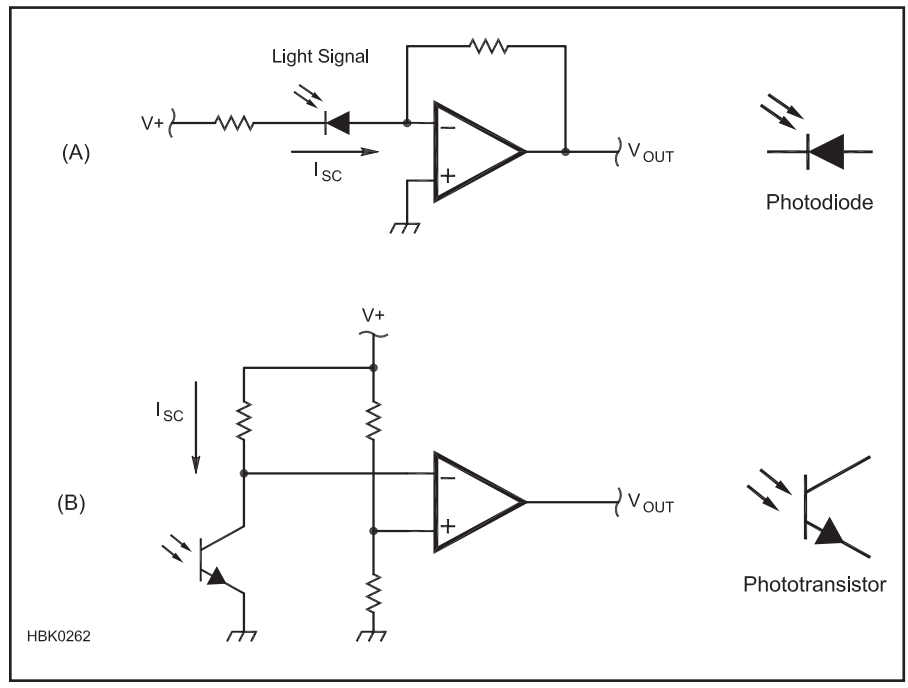

Fig 3.28 — The photodiode (A) is used to detect light. An amplifier circuit changes the variations in photodiode current to a change in output voltage. At (B), a phototransistor conducts current when its base is illuminated. This causes the voltage at the collector to change causing the amplifier's output to switch between ON and OFF.

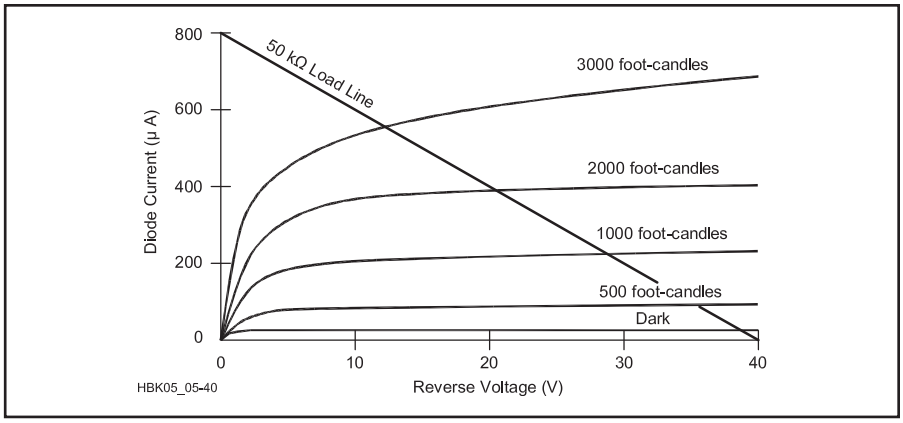

Fig 3.29 — Photodiode I-V curve. Reverse voltage is plotted on the x-axis and current through diode is plotted on the y-axis. Various response lines are plotted for different illumination. Except for the zero illumination line, the response does not pass through the origin since there is current generated at the PN junction by the light energy. A load line is shown for a 50-kΩ resistor in series with the photodiode.

(reverse current increases as reverse voltage increases). See **Fig 3.29** 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 3.29 intersect the left side of the graph.

Photoconductors and photodiodes are generally used to produce light-related analog signals that require further processing. For example, a photodiode is used to detect infrared light signals from remote control devices as in Fig 3.28A. The light falling on the reverse-biased photodiode causes a change in I_{SC} that is detected as a change in output voltage.

Light falling on the phototransistor acts as base current to control a larger current between the collector and emitter. Thus the phototransistor acts as an amplifier whose input signal is light and whose output is current. Phototransistors are more sensitive to light than the other devices. Phototransistors have lots of light-to-current gain, but photodiodes normally have less noise, so they make more sensitive detectors. The phototransistor in Fig 3.28B is being used as a detector. Light falling on the phototransistor causes collector current to flow, dropping the collector voltage below the voltage at the amplifier's + input and causing a change in V_{OUT}.

PHOTOVOLTAIC CELLS

When illuminated, the reverse-biased photodiode has a reverse current caused by 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 as the *terminal voltage,* V_T, or *open-circuit voltage,* V_{OC}.

Devices that use light from the sun to produce electricity in this way are called *photovoltaic (PV)* or *solar cells* or *solar batteries*. The symbol for a photovoltaic cell is shown in **Fig 3.30A**. The electrical equivalent circuit of the cell is shown in Fig 3.30B. The cell is basically a large, flat diode junction exposed to light. Metal electrodes on each side of the junction collect the current generated.

When illuminated, the cell acts like a current source, with some of the current flowing through a diode (made of the same material as the cell), a shunt resistance for leakage current and a series resistor that represents the resistance of the cell. Two quantities define the electrical characteristics of common silicon photovoltaic cells. These are an open-circuit voltage, V_{OC} of 0.5 to 0.6 V and the output *short-circuit current*, I_{SC} as above, that depends on the area of the cell exposed to light and the degree of illumination. A measure of the cell's effectiveness at converting light into current is the *conversion efficiency*. Typical silicon solar cells have a conversion efficiency of 10 to 15% although special cells with stacked junctions or using special light-absorbing materials have shown efficiencies as high as 40%.

Solar cells are primarily made from single-crystal slices of silicon, similar to diodes and transistors, but with a much greater area. *Polycrystalline silicon* and *thin-film* cells are less expensive, but have lower conversion efficiency. Technology is advancing rapidly in the field of photovoltaic energy and there are a number of different types of materials and fabrication techniques that have promise in surpassing the effectiveness of the single-junction silicon cells.

Solar cells are assembled into arrays called *solar panels*, shown in Fig 3.30C. Cells are connected in series so that the combined output voltage is a more useful voltage, such as 12 V. Several strings of cells are then connected in parallel to increase the available output current. Solar panels are available with output powers from a few watts to hundreds of watts. Note that unlike batteries, strings of solar cells can be connected directly in parallel because they act as sources of constant current instead

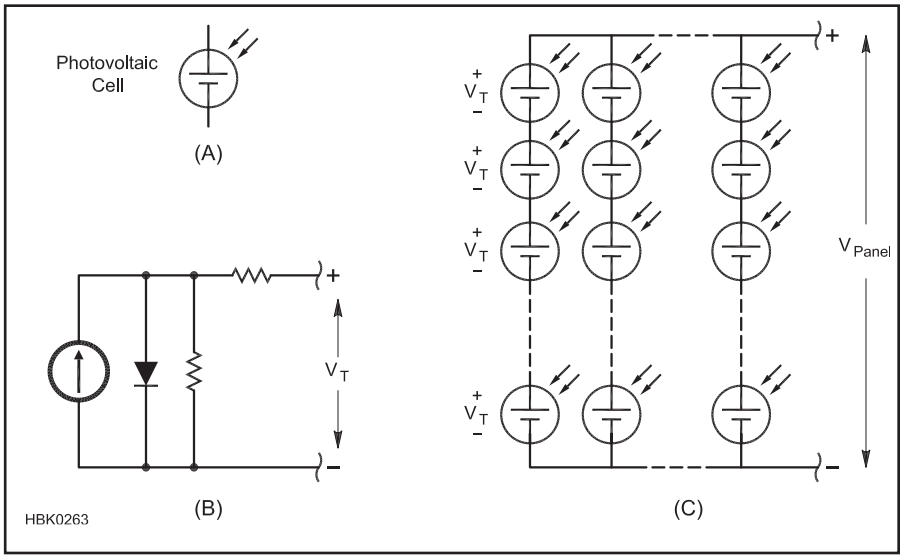

Fig 3.30 — A photovoltaic cell's symbol (A) is similar to a battery. Electrically, the cell can be modeled as the equivalent circuit at (B). Solar panels (C) consist of arrays of cells connected to supply power at a convenient voltage.

of voltage. (More information on the use of solar panels for powering radio equipment can be found in the chapter on **Power Sources**.)

LIGHT EMITTING DIODES AND LASER DIODES

In the photodiode, energy from light falling on the semiconductor material is absorbed to create 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 (non-thermal) 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-carbide (SiC) creates a blue LED.

White LEDs are made by coating the inside of the LED lens with a white-light emitting phosphor and illuminating the phosphor with light from a single-color LED. White LEDs are currently approaching the cost of cold-florescent (CFL) bulbs and will eventually displace CFL technology for lighting, just as CFL is replacing the incandescent bulb.

The LED, shown in **Fig 3.31**, is very simple to use. It is connected across a voltage source (V) with a series resistor (R) that limits the current to the desired level (I_F) for the amount of light to be generated.

$$R = \frac{V - V_F}{I_F} \qquad (12)$$

where V_F is the forward voltage of the LED.

The cathode lead is connected to the lower potential, and is specially marked as shown in the manufacturer's data sheet. LEDs may be connected in series for additional light, with the same current flowing in all of the diodes. Diodes connected in parallel without current-limiting resistors for each diode are likely to share the current unequally, thus the series connection is preferred.

The laser diode operates by similar principles to the LED except that all of the light produced is *monochromatic* (of the same color and wavelength) and it is *coherent*, meaning that all of the light waves emitted by the device are in phase. Laser diodes generally require higher current than an LED and will not emit light until the *lasing current* level is reached. Because the light is monochromatic and coherent, laser diodes can be used for applications requiring precise illumination and modulation, such as high-speed data links, and in data storage media such as CD-ROM and DVD. LEDs are not used for high-speed or high-frequency analog modulation because of recovery time limitations, just as in regular rectifiers.

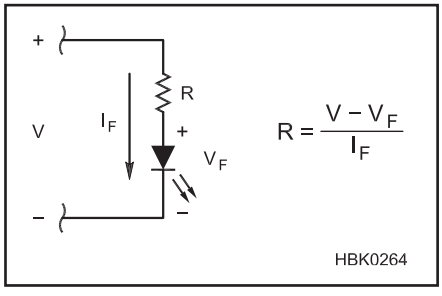

Fig 3.31 — A light-emitting diode (LED) emits light when conducting forward current. A series current-limiting resistor is used to set the current through the LED according to the equation.

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 (see **Fig 3.32**). The optoisolator, as its name suggests, isolates different circuits from each other. Typically, isolation resistance is on the order of 10^{11} Ω and isolation capacitance is less than 1 pF. Maximum voltage isolation varies from 1000 to 10,000 V ac. The most common optoisolators are available in 6-pin DIP packages.

Optoisolators are primarily used for voltage level shifting and signal isolation. Voltage level shifting allows signals (usually digital signals) to pass between circuits operating at greatly different voltages. The isolation has two purposes: to protect circuitry (and operators) from excessive voltages and to isolate noisy circuitry from noise-sensitive circuitry.

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. Optoisolators with bandwidths of several MHz are available, but are somewhat expensive.

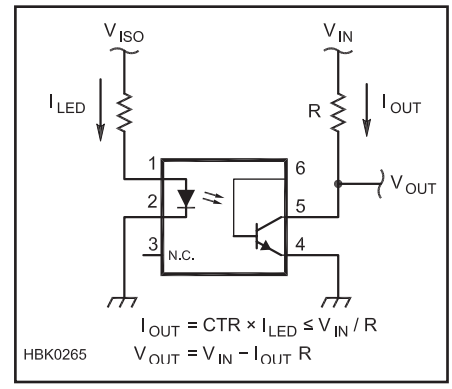

Fig 3.32 — The optoisolator consists of an LED (input) that illuminates the base of a phototransistor (output). The phototransistor then conducts current in the output circuit. CTR is the optoisolator's current transfer ratio.

As an example of voltage level shifting, an optoisolator can be used to allow a low-voltage, solid-state electronic Morse code keyer to activate a vacuum-tube grid-block keying circuit that operates at a high negative voltage (typically about –100 V) but low current. No common ground is required between the two pieces of equipment.

Optoisolators can act as input protection for circuits that are exposed to high voltages or transients. For example, a short 1000-V transient that can destroy a semiconductor circuit will only saturate the LED in the optoisolator, preventing damage to the circuit. The worst that will happen is the LED in the optoisolator will be destroyed, but that is usually quite a bit less expensive than the circuit it is protecting.

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 optoisolators is when a computer is used to control radio equipment. The computer signal, and even its ground reference, typically contains considerable wide-band noise caused by 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 greatly different from the design of circuits with LEDs and with transistors. On the input side, the LED is forward-biased and driven with a series current-limiting resistor whose value limits current to less than the maximum value for the device (for example, 60 mA is the maximum LED current for a 4N25). This is identical to designing with standalone LEDs.

On the output side, instead of current gain for a transistor, the optoisolator's *current transfer ratio* (*CTR*) is used. CTR is a ratio

given in percent between the amount of current through the LED to the output transistor's maximum available collector current. For example, if an optoisolator's CTR = 25%, then an LED current of 20 mA results in the output transistor being able to conduct up to 20 × 0.25 = 5 mA of current in its collector circuit.

If the optoisolator is to be used for an analog signal, the input signal must be appropriately dc shifted so that the LED is always forward biased. A phototransistor with all three leads available for connection (as in Fig 3.32) is required. The base lead is used for biasing, allowing the optical signal to create variations above and below the transistor's operating point. The collector and emitter leads are used as they would be in any transistor amplifier circuit. (There are also linear optoisolators that include built-in linearizing circuitry.) The use of linear optoisolators is not common.

FIBER OPTICS

An interesting variation on the optoisolator is the *fiber-optic* connection. Like the optoisolator, the input signal is used to drive an LED or laser diode that produces modulated light (usually light pulses). 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 signal is recovered by a photo detector (photoresistor, photodiode or phototransistor). Because the fiber optic cable is nonconductive, the transmitting and receiving systems are electrically isolated.

Fiber optic cables generally have far less loss than coaxial cable transmission lines. They do not leak RF energy, nor do they pick up electrical noise. Fiber optic cables are virtually immune to electromagnetic interference! 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 gigahertz bandwidth.

3.3.5 Linear Integrated Circuits

If you look inside 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. The individual components are attached together on a small circuit board or with bonding wires.

Once widespread in electronics, hybrid ICs have largely been displaced by fully integrated devices with all the components on the same piece of semiconductor material. Manufacturers often use hybrids when small size is needed, but there is insufficient volume to justify the expense of a custom IC.

A current application for hybrid circuitry is UHF and microwave amplifiers — they are in wide use by the mobile phone industry. For example, the Motorola MW4IC915N Wideband Integrated Power Amplifier is a complete 15-W transmitting module. Its TO-272 package is only about 1 inch long by ⅜-inch wide. This particular device is designed for use between 750 and 1000 MHz and can be adapted for use on the amateur 902 MHz band. Other devices available as hybrid circuits include oscillators, signal processors, preamplifiers and so forth. Surplus hybrids can be hard to adapt to amateur use unless they are clearly identified with manufacturing identification such that a data sheet can be obtained.

MONOLITHIC INTEGRATED CIRCUITS

In order to build entire circuits on a single piece of semiconductor, it must be possible to fabricate resistors and capacitors, as well as transistors and diodes. Only then can the entire circuit be created on one piece of silicon called a *monolithic integrated circuit*.

An integrated circuit (IC) or "chip" is fabricated in layers. An example of a semiconductor circuit schematic and its implementation in an IC is pictured in **Fig 3.33**. 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, except for the demanding military and aerospace applications.

On top of the P-type substrate is a thin layer of N-type material in which the active and passive 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 aluminum (copper is used in some high-speed digital ICs) 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 3.33B illustrates the various layers. Connections 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.

Fabricating Resistors and Capacitors

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

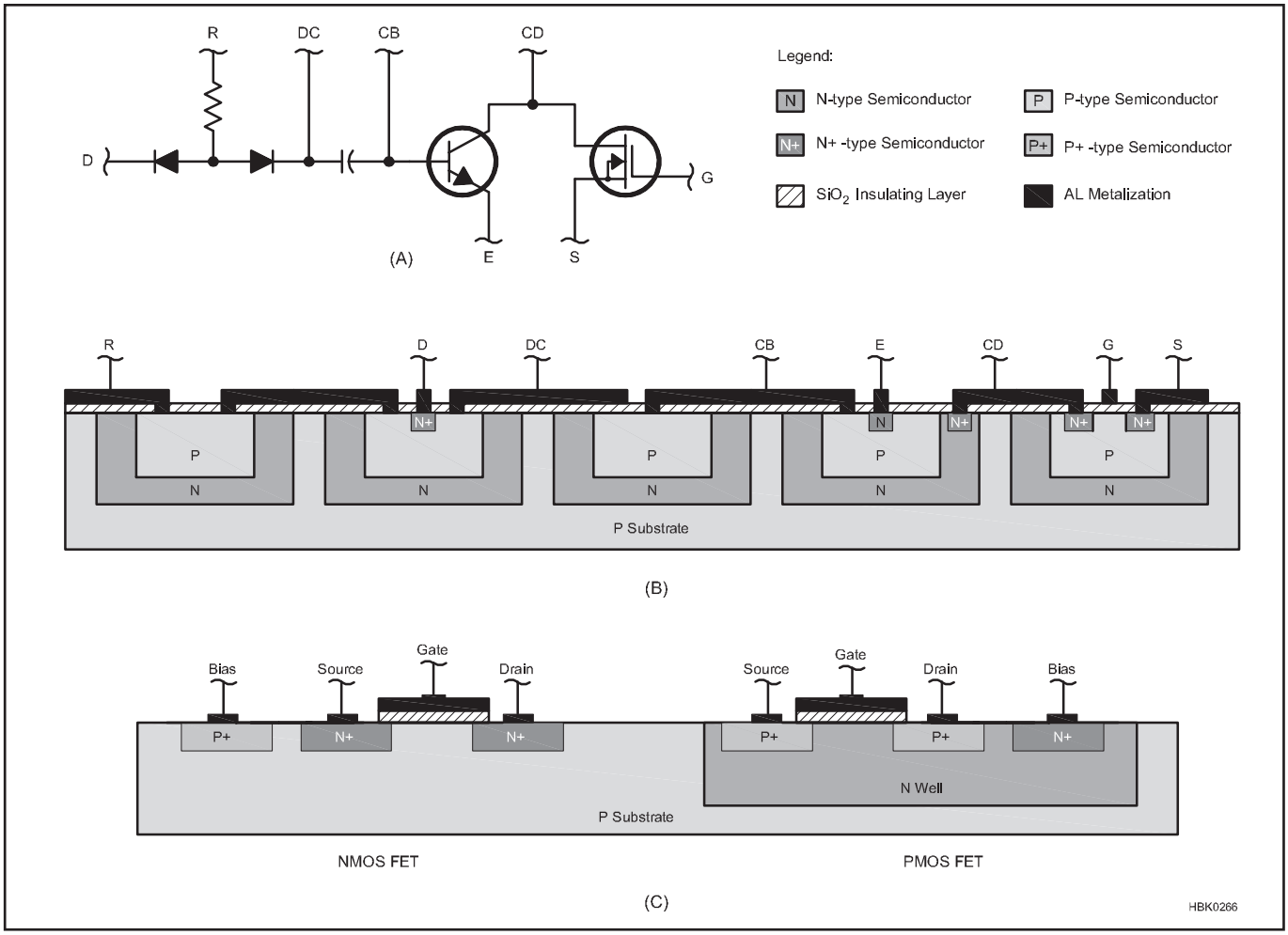

Fig 3.33 — Integrated circuit layout. (A) Circuit containing two diodes, a resistor, a capacitor, an NPN transistor and an N-channel MOSFET. Labeled leads are D for diode, R for resistor, DC for diode-capacitor, E for emitter, S for source, CD for collector-drain and G for gate. (B) Integrated circuit 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).

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 reduce or eliminate the need for external inductors.

Fabricating Diodes and 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.

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.

FETs can also be fabricated in IC form as shown in Fig 3.33C. Due to its many func-

tional 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. CMOS ICs can contain 20 times more circuitry than bipolar ICs with the same chip size, making the devices more powerful and less expensive than those based on bipolar technology. CMOS is the most popular form of integrated circuit.

The final configuration of the switching circuit is CMOS as described in a previous section of this chapter. CMOS gates require two FETs, one of each form (NMOS and PMOS as shown in the figure). NMOS requires fewer processing steps, and the individual FETs have lower on-resistance than PMOS. The fabrication of NMOS FETs is the same as for individual semiconductors; P+ wells form the source and drain in a P-type substrate. A metal gate electrode is formed on top of an insulating SiO_2 layer so that the channel forms in the P-type substrate between the source and drain. For the PMOS FET, the process is similar, but begins with an N-type well in the P-type substrate.

MOSFETs fabricated in this manner also have bias (B) terminals connected to the positive power supply to prevent destructive *latch-up*. This can occur in CMOS gates because the two MOSFETs form a *parasitic SCR*. If the SCR mode is triggered and both transistors conduct at the same time, large currents can flow through the FET and destroy the IC unless power is removed. 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.

While CMOS is the most widely used technology, 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.

INTEGRATED CIRCUIT ADVANTAGES

The primary advantages of using integrated circuits as opposed to discrete components are the greatly decreased physical size of the circuit and improved reliability. In fact, studies show that failure rate of electronic circuitry is most closely related to the number of interconnections between components. Thus, using integrated circuits not only reduces volume, but makes the equipment more reliable.

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 the size of the individual features that can be created on the semiconductor wafer. Since the invention of the monolithic IC in the mid-1960s, feature size limits have dropped below 100 nanometers (1/10th of a millionth of a meter) as of 2009. Currently, it is not unusual to find chips with more than one million transistors on them.

In addition to size and reliability of the ICs themselves, the relative properties of the devices on a single chip are very predictable. 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 are nearly 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. With closely matched components on the single substrate, this type of design works very well without requiring external components to adjust or "trim" IC performance.

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. 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, particularly the ubiquitous mobile phone and other wireless devices. 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 vacuum 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 of more than 50 W output 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. As described previously, monolithic microwave integrated circuits — MMICs — are available for operation to 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 a result, analog ICs are now available with gain-bandwidth products over 1 GHz.

A major impediment to the introduction

of new integrated circuits, particularly with special applications, is the very high cost of development of new designs for full custom ICs. The masking cost alone for a designed and tested integrated circuit can exceed $100,000 so these devices must sell in high volume to recoup the development costs. Over the past decade, however, IC design tools and fabrication services have become available that greatly reduce the cost of IC development by using predefined circuit layouts and functional blocks. These *application-specific integrated circuits* (*ASICs*) and *programmable gate arrays (PGA)* are now routinely created and used for even limited production runs. While still well beyond the reach of the individual amateur's resources, the ASIC or PGA is widely used in nearly all consumer electronics and in many pieces of radio equipment.

The drawback of the ASIC and PGA is that servicing and repairing the equipment at the integrated circuit level is almost impossible for the individual without access to the manufacturer's inventory of parts and proprietary information. Nevertheless, just as earlier amateurs moved beyond replacing individual components to diagnosing ICs, today's amateurs can troubleshoot to the module or circuit-board level, treating them as "components" in their own right.

3.3.6 Comparison of Semiconductor Devices for Analog Applications

Analog signal processing deals with changing a signal to a desired form. The three primary types of devices — bipolar transistors, field-effect transistors and integrated circuits — perform similar functions, each with specific advantages and disadvantages. The vacuum tube, once the dominant signal processing component, is relegated to high-power amplifier and display applications and is found only in the **RF Power Amplifiers** chapter of this *Handbook*. *Cathode-ray tubes* (CRTs) are covered in a supplement on the *Handbook* CD-ROM.

Bipolar transistors, when treated properly, can have virtually unlimited life spans. They are relatively small and, if they do not handle high currents, do not generate much heat. 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 final amplifiers generally include circuitry to protect the transistors from the high voltages generated by reflections under high SWR conditions. Bipolar transistors and ICs, like all semiconductors, are susceptible to damage from power and lightning transients.

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. FETs are less expensive to fabricate in ICs and so are gradually replacing bipolar transistors in many IC applications.

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, including combinations of analog and digital circuitry that previously required multiple devices. *Charge-coupled devices* (CCD) imaging technology has replaced vidicon tubes in video cameras and film cameras of all types. *Liquid crystal displays* (LCDs) in laptop computers and standalone computer displays have largely displaced the bulky and power-hungry cathode-ray tube (CRT), although it is still found in analog oscilloscopes and some types of video display equipment.

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. ASIC technology, as mentioned earlier, brings the power of custom electronics to the radio, but also makes it nearly impossible to repair at the level of the IC, even if the problem is known. If field repair and service at the component level are to be performed, it is important to use standard ICs wherever possible. Even so, when demand for a particular component drops, a manufacturer will discontinue its production. This happens on an ever-decreasing timeline.

A further consideration is the trend toward digital signal processing and software-defined radio systems. (See the chapter on **DSP and Software Radio Design**.) More and more analog functions are being performed by microprocessors and the analog signals converted to digital at higher and higher frequencies. There will always be a need for analog circuits, but the balance point between analog and digital is shifting towards the latter. In future years, radio and test equipment will consist of a powerful, general-purpose digital signal processor, surrounded by the necessary analog circuitry to convert the signals to digital form and supply the processor with power.

3.4 Analog Systems

Many kinds of electronic equipment are developed by combining basic analog signal processing circuits, often treating them as independent functional blocks. This section describes several topics associated with building analog systems from multiple blocks. Although not all basic electronic functions are discussed here, the concepts associated with combining them can be applied generally.

An analog circuit can contain any number of discrete components (or may be implemented as an IC). Since our main concern is the effect that circuitry has on a signal, we often describe the circuit by its actions rather than by its specific components. A *black box* is a circuit that can be described entirely by the behavior of its interfaces with other blocks and circuitry. When circuits are combined in such as way as to perform sequential operations on a signal, the individual circuits are called *stages*.

The most general way of referring to a circuit is as a *network*. 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. Interfaces between the network and the rest of the network are called *ports*.

Many analog circuits are analyzed as *two-port networks* with an input and an output port. The signal is fed into the input port, is modified inside the network and then exits from the output port. (See the chapter on **RF Techniques** for more information on two-port networks.)

3.4.1 Transfer Functions

The specific way in which the analog circuit modifies the signal can be described mathematically as a *transfer function*. The mathematical operation that combines a signal with a transfer function is pictured symbolically in **Fig 3.34**. The transfer function, h(t) or h(f), describes the circuit's modification of the input signal in the time domain where all values are functions of time, such as a(t) or b(t), or in the frequency domain where all values are functions of frequency, such as a(f) or b(f). The mathematical operation by which h(t) operates on a(t) is called *convolution* and is represented as a dot, as in $a(t) \bullet h(t) = b(t)$. In the frequency domain, the transfer function multiplies the input, as in $a(f) \times h(f) = b(f)$.

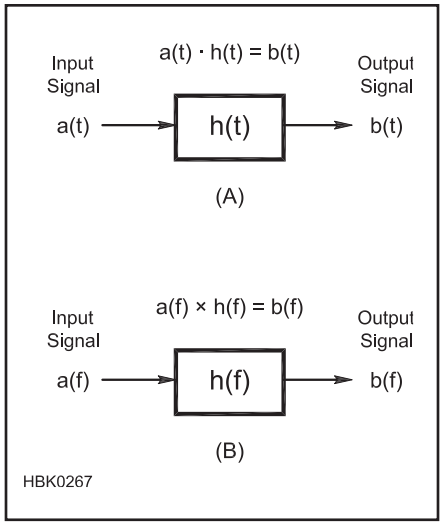

Fig 3.34 — Linear function blocks and transfer functions. The transfer function can be expressed in the time domain (A) or in the frequency domain (B). The transfer function describes how the input signal a(t) or a(f) is transformed into the output signal b(t) or b(f).

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.

3.4.2 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 of which they are made. 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 the functions of the stages are completely independent. In other words, when the action of a stage is unchanged, regardless of the characteristics of any stages connected to its input or output.

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 in the cascaded circuits.

While each stage in a series may use feedback within itself, feedback around more than one stage may create a function — and resultant performance — different from any of the included stages. Examples include oscillation or negative feedback.

3.4.3 Amplifier Frequency Response

At higher frequencies a typical amplifier acts as a low-pass filter, decreasing amplification with increasing frequency. Signals within a range of frequencies are amplified consistently but outside that range the amplification changes. At high gains many amplifiers work properly only over a small range of frequencies. The combination of gain and frequency response is often expressed as a *gain-bandwidth product*. For many amplifiers, gain times bandwidth is approximately constant. As gain increases, bandwidth decreases, and vice versa.

Performance at lower frequencies depends on whether the amplifier is *dc- or ac-coupled*. Coupling refers to the transfer of signals between circuits. A dc-coupled amplifier amplifies signals at all frequencies down to dc. An ac-coupled amplifier acts as a high-pass filter, decreasing amplification as the frequency decreases toward dc. Ac-coupled circuits usually use capacitors to allow ac signals to flow between stages while blocking the dc bias voltages of the circuit.

3.4.4 Interstage Loading and Impedance Matching

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, 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.

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. 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 from the output of the stage to a load connected to the output. In this case, the output impedance of the stage is *matched* or transformed to that of the load (or vice versa). This allows the stage to operate at its optimum voltage and current levels. 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 required. Instead, current between stages can be minimized by using mismatched impedances. Ideally, if the output impedance of a network is very low 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.

3.4.5 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 circuit's bandwidth or by reducing the temperature in the system. Thermal noise voltage and current vary with the circuit impedance and follow Ohm's Law. Low-noise-amplifier-design techniques are based on these relationships.

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 amplifier 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:

$$NF = 10 \log \frac{P_{NO}}{A\, P_{N\,TH}} \quad (13)$$

where

P_{NO} = total noise output power,
A = amplification gain
$P_{N\,TH}$ = input thermal noise power.

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, such as a communications receiver, each stage contributes to the total noise of the system. The noise factor of the first stage dominates the noise factor of the entire system because noise added at the first stage is then multiplied by each following stage. Noise added by later stages is not multiplied to the same degree and so is a smaller contribution to the overall noise at the output.

Designers try to optimize system noise factor by using a first stage with a minimum possible noise factor and maximum possible gain. (Caution: A circuit that overloads is often as useless as one that generates too much noise.) See the **RF Techniques** chapter for a more complete discussion on noise. Circuit overload is discussed in the **Receivers** chapter.

3.4.6 Buffering

It is often necessary to isolate the stages of an analog circuit. This isolation reduces the loading, coupling and feedback between stages. It is often necessary to connect circuits that operate at different impedance levels between stages. An intervening stage, a type of amplifier called a *buffer*, is often used for this purpose.

Buffers can have high values of amplification but this is unusual. A buffer used for impedance transformation generally has a low or unity gain. 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. Combinations of buffer stages can also be effective at isolating the stages from each other and making impedance transformations, as well.

3.5 Amplifiers

By far, the most common type of analog circuit is the amplifier. The basic component of most electronics — the transistor — is an amplifier in which a small input signal controls a larger signal. Transistor circuits are designed to use the amplifying characteristics of transistors in order to create useful signal processing functions, regardless of whether the input signal is amplified at the output.

3.5.1 Amplifier Configurations

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 that is used by both the input and output ports of the circuit. The most common reference is ground, but positive and negative power sources are also valid references.

The type of circuit reference used depends on the type of device (transistor [NPN or PNP] or FET [P-channel or N-channel]), which lead is chosen as 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. (Vacuum tube amplifiers are discussed in the chapter on **RF Power Amplifiers**.)

DC power sources are usually constructed so that ac signals at the output terminals are bypassed to ground through a very low impedance. This allows the power source to be treated as an *ac ground*, even though it may be supplying dc voltages to the circuit. When a circuit is being analyzed for its ac behavior, ac grounds are usually treated as ground, since dc bias is ignored in the ac analysis. Thus, a transistor's collector can be considered the "common" part of the circuit, even though in actual operation, a dc voltage is applied to it.

Fig 3.35 shows the three basic types of bipolar transistor amplifiers: the common-base, common-emitter, and common-collector. The common terminal is shown connected to ground, although as mentioned earlier, a dc bias voltage may be present. Each type of amplifier is described in the following sections. Following the description of the amplifier, additional discussion of biasing transistors and their operation at high frequencies and for large signals is presented.

3.5.2 Transistor Amplifiers

Creating a useful transistor amplifier depends on using an appropriate model for the transistor itself, choosing the right configuration of the amplifier, using the design equations for that configuration and insuring that the amplifier operates properly at different temperatures. This section follows that sequence, first introducing simple transistor models and then extending that knowledge to the point of design guidelines for common circuits that use bipolar and FETs.

DEVICE MODELS AND CLASSES

Semiconductor circuit design is based on equivalent circuits that describe the physics of the devices. These circuits, made up of voltage and current sources and passive components such as resistors, capacitors and inductors, are called models. A complete model that describes a transistor exactly over a wide frequency range is a fairly complex circuit. As a result, simpler models are used in specific circumstances. For example, the *small-signal model* works well when the device is operated close to some nominal set of characteristics such that current and voltage interact fairly linearly. The *large-signal model* is used when the device is operated so that it enters its saturation or cut-off regions, for example.

Different frequency ranges also require different models. The *low-frequency models* used in this chapter can be used to develop circuits for dc, audio and very low RF applications. At higher frequencies, small capacitances and inductances that can be ignored at low frequencies begin to have significant effects on device behavior, such as gain or impedance. In addition, the physical structure of the device also becomes significant as gain begins to drop or phase shifts between input and output signals start to grow. In this region, *high-frequency models* are used.

Amplifiers are also grouped by their *operating class* that describes the way in which the input signal is amplified. There are several classes of analog amplifiers; A, B, AB, AB1, AB2 and C.

The analog class designators specify over how much of the input cycle the active device is conducting current. A class-A amplifier's active device conducts current for 100 percent of the input signal cycle, such as shown in Fig 3.6. A class-B amplifier conducts during one-half of the input cycle, class-AB, AB1, and AB2 some fraction between 50 and 100 percent of the input cycle, and class-C for less than 50 percent of the input signal cycle.

Digital amplifiers, in which the active device is operated as a switch that is either fully-on or fully-off, similarly to switchmode power supplies, are also grouped by classes beginning with the letter D and beyond. Each different class uses a different method of con-

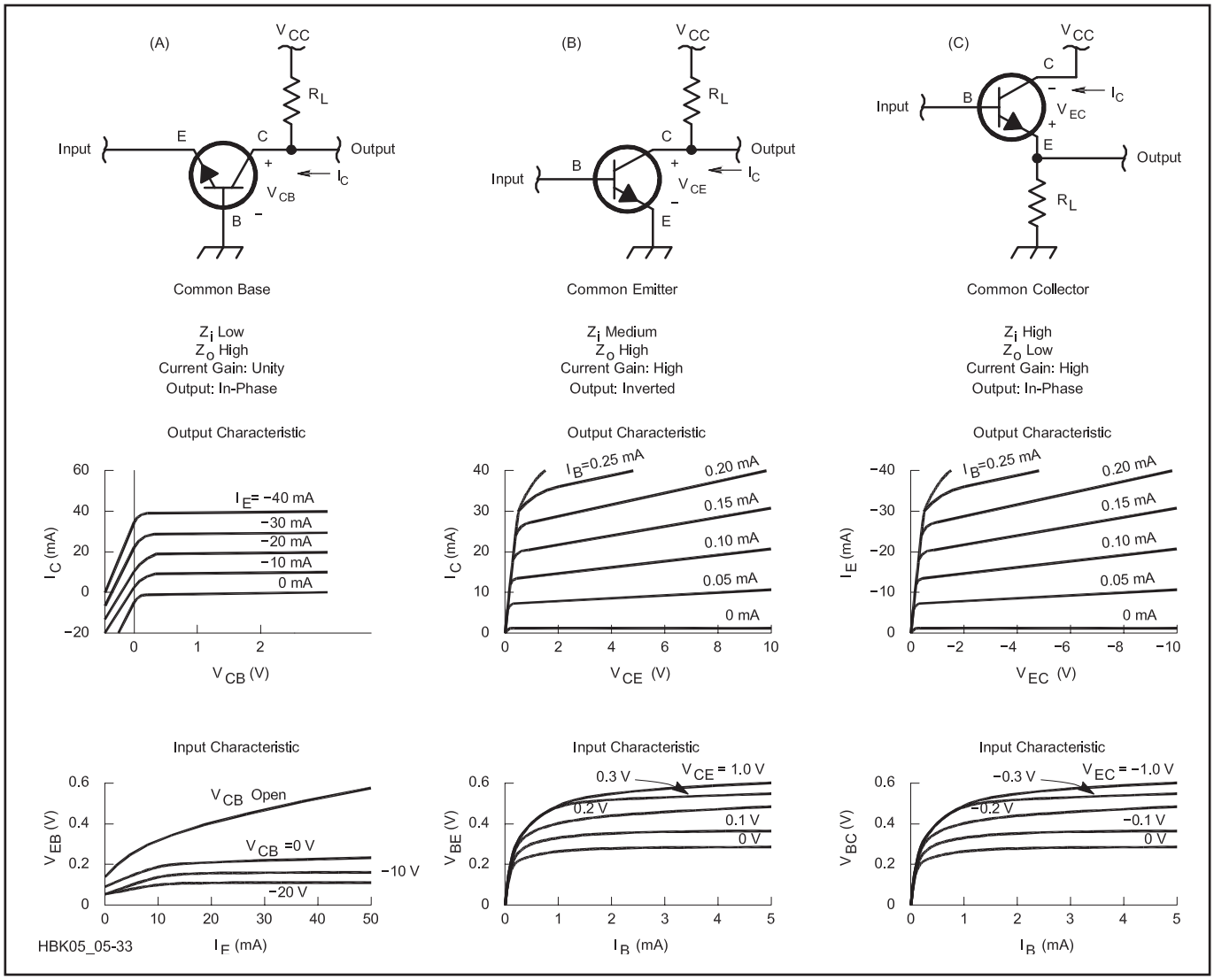

Fig 3.35 — The three configurations of bipolar transistor amplifiers. Each has a table of its relative impedance and current gain. The output characteristic curve is plotted for each, with the output voltage along the x-axis, the output current along the y-axis and various curves plotted for different 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.

verting the switch's output waveform to the desired RF waveform.

Amplifier classes, models and their use at high-frequencies are discussed in more detail in the chapter on **RF Techniques**. In addition, the use of models for circuit simulation is discussed at length in the **Computer-Aided Circuit Design** chapter.

3.5.3 Bipolar Transistor Amplifiers

In this discussion, we will focus on simple models for bipolar transistors (BJTs). This discussion is centered on NPN BJTs but applies equally well to PNP BJTs if the bias voltage and current polarities are reversed. This section assumes the small-signal, low-frequency models for the transistors.

SMALL-SIGNAL BJT MODEL

The transistor is usually considered as a *current-controlled* device in which the base current controls the collector current:

$$I_c = \beta I_b \qquad (14)$$

where

I_c = collector current
I_b = base current
β = *common-emitter current gain*, beta.

(The term "common-emitter" refers to the type of transistor circuit described below in which the transistor operates with base current as its input and collector circuit as its output.) Current is positive if it flows *into* a device terminal.

The transistor can also be treated as a voltage-controlled device in which the transistor's emitter current, I_e, is controlled by the base-emitter voltage, V_{be}:

$$I_c = I_{es}\left[e^{(qV_{be}/kT)} - 1\right] \approx I_{es}e^{(qV_{be}/kT)} \qquad (15)$$

where

q = electronic charge
k = Boltzmann's constant
T = temperature in degrees Kelvin (K)
I_{es} = emitter saturation current, typically 1×10^{-13} A.

The subscripts for voltages indicate the direction of positive voltage, so that V_{be} indicates positive is from the base to the emitter. It is simpler to design circuits using the current-controlled device, but accounting for the transistor's behavior with temperature requires an understanding of the voltage-controlled model.

Transistors are usually driven by both biasing and signal voltages. Equations 14 and 15 apply to both transistor dc biasing and signal design. Both of these equations are approximations of the more complex behavior exhibited by actual transistors. Equation 15 applies to a simplification of the first *Ebers-Moll model* in Reference 1. More sophisticated models for BJTs are described by Getreu in Reference 2. Small-signal models treat only the signal components. We will consider bias later.

The next step is to use these basic equations to design circuits. We will begin with small-signal amplifier design and the limits of where the techniques can be applied. Later, we'll discuss large-signal amplifier design and the distortion that arises from operating the transistor in regions where the relationship between the input and output signals is nonlinear.

Common-Emitter Model

Fig 3.36 shows a BJT amplifier connected in the common-emitter configuration. (The emitter, shown connected to ground, is common to both the input circuit with the voltage source and the output circuit with the transistor's collector.) The performance of this circuit is adequately described by equation 14. **Fig 3.37** shows the most common of all transistor small-signal models, a controlled current source with emitter resistance.

There are two variations of the model shown in the figure. Fig 3.37B shows the base as a direct connection to the junction of a current-controlled current source ($I_c = \beta I_b$) and a resistance, r_e, the *dynamic emitter resistance* representing the change in V_{be} with I_e. This resistance also changes with emitter current:

$$r_e = \frac{kT}{qI_c} \approx \frac{26}{I_e} \qquad (16)$$

where I_e is the dc bias current in milliamperes.

The simplified approximation only applies at a typical ambient temperature of 300 K because r_e increases with temperature. In Fig 3.37A, the emitter resistance has been moved to the base connection, where it has the value $(\beta+1)r_e$. These models are electrically equivalent.

The transistor's output resistance (the Thevenin or Norton equivalent resistance between the collector and the grounded emitter) is infinite because of the current source.

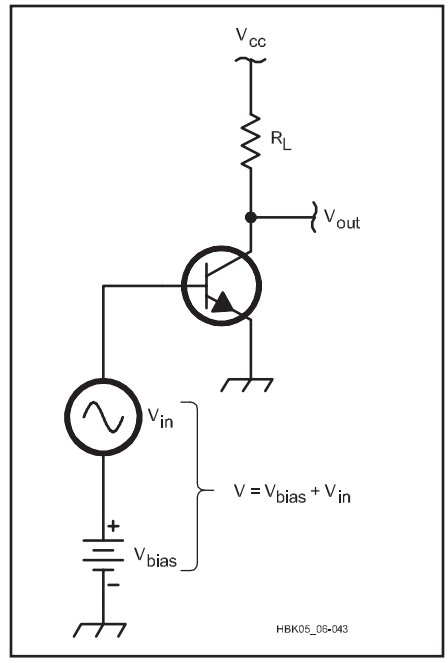

Fig 3.36 — Bipolar transistor with voltage bias and input signal.

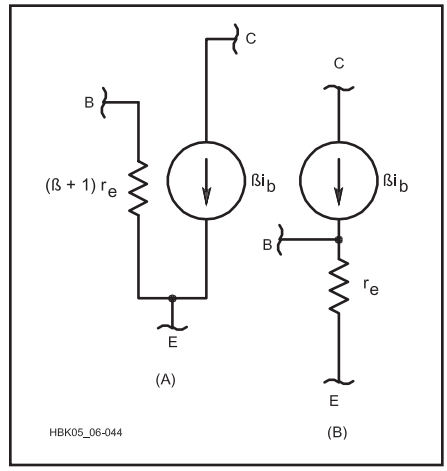

Fig 3.37 — Simplified low-frequency model for the bipolar transistor, a "beta generator with emitter resistance." r_e = 26 / I_e (mA dc).

This is a good approximation for most silicon transistors at low frequencies (well below the transistor's gain-bandwidth product, F_T) and will be used for the design examples that follow.

As frequency increases, the capacitance inherent in BJT construction becomes significant and the *hybrid-pi model* shown in **Fig 3.38** is used, adding C_π in parallel with the input resistance. In this model the transfer parameter h_{ie} often represents the input impedance, shown here as a resistance at low frequencies.

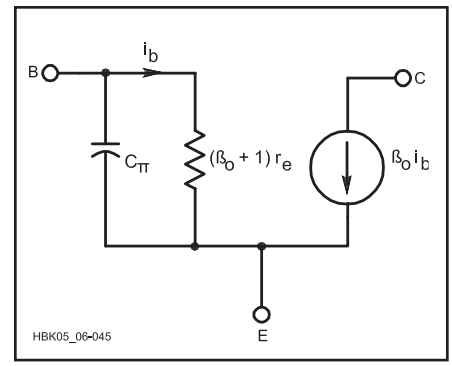

Fig 3.38 — The hybrid-pi model for the bipolar transistor.

THREE BASIC BJT AMPLIFIERS

Fig 3.39 shows a small-signal model applied to the three basic bipolar junction transistor (BJT) amplifier circuits: *common-emitter (*CE), *common-base* (CB) and *common-collector* (CC), more commonly known as the *emitter-follower* (EF). As defined earlier, the word "common" indicates that the referenced terminal is part of both the input and output circuits.

In these simple models, transistors in both the CE and CB configurations have infinite output resistance because the collector current source is in series with the output current. (The amplifier circuit's output impedance must include the effects of R_L.) The transistor connected in the EF configuration, on the other hand, has a finite output resistance because the current source is connected in parallel with the base circuit's equivalent resistance. Calculating the EF amplifier's output resistance requires including the input voltage source, V_s, and its impedance.

The three transistor amplifier configurations are shown as simple circuits in Fig 3.35. Each circuit includes the basic characteristics of the amplifier and characteristic curves for a typical transistor in each configuration. Two sets of characteristic curves are presented: one describing the input behavior and the other describing the output behavior in each amplifier configuration. The different transistor amplifier configurations have different gains, input and output impedances and phase relationships between the input and output signals.

Examining the performance needs of the amplifier (engineers refer to these as the circuit's *performance requirements*) determines which of the three circuits is appropriate. Then, once the amplifier configuration is chosen, the equations that describe the circuit's behavior are used to turn the performance requirements into actual circuit component values.

This text presents design information for the CE amplifier in some detail, then summarizes designs for the CC and CB ampli-

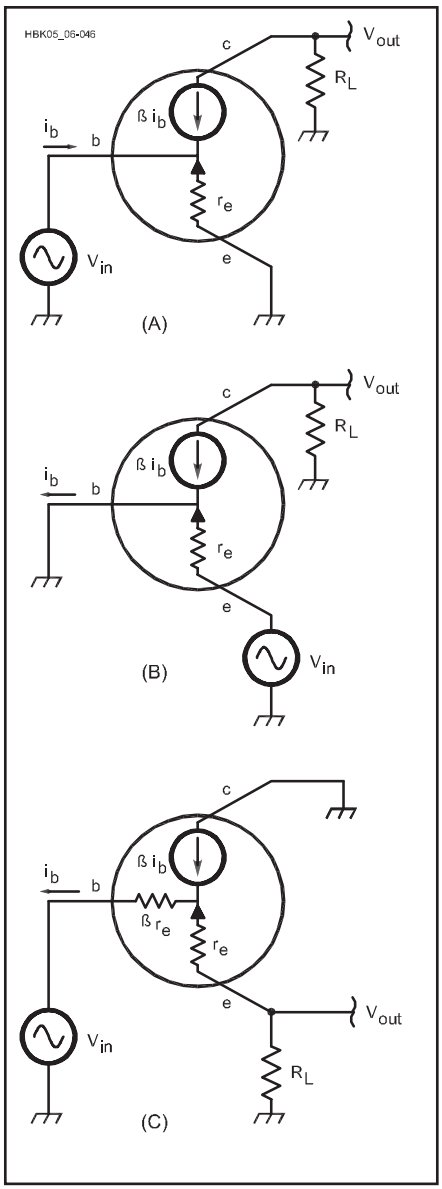

Fig 3.39 — Application of small-signal models for analysis of (A) the CE amplifier, (B) the CB and (C) the EF (CC) bipolar junction transistor amplifiers.

fiers. Detailed design analysis for all three amplifiers is described in the texts listed in the reference section for this chapter. All of the analysis in the following sections assume the small-signal, low-frequency model and ignore the effects of the coupling capacitors. High-frequency considerations are discussed in the **RF Techniques** chapter and some advanced discussion of biasing and large signal behavior of BJT amplifiers is available on the companion CD-ROM.

LOAD LINES AND Q-POINT

The characteristic curves in Fig 3.35 show that the transistor can operate with an infinite number of combinations of current (collector, emitter and base) and voltage (collector-emitter, collector-base or emitter-collector). The particular combination at which the amplifier is operating is its *operating point*. The operating point is controlled by the selection of component values that make up the amplifier circuit so that it has the proper combination of gain, linearity and so forth. The result is that the operating point is restricted to a set of points that fall along a *load line*. The operating point with no input signal applied is the circuit's *quiescent point* or *Q-point*. As the input signal varies, the operating point moves along the load line, but returns to the Q-point when the input signal is removed.

Fig 3.40 shows the load line and Q-point for an amplifier drawn on a transistor's set of characteristic curves for the CE amplifier circuit. The two end-points of the load line correspond to transistor saturation (I_{Csat} on the I_C current axis) and cutoff (V_{CC} on the V_{CE} voltage axis).

When a transistor is in saturation, further increases in base current do not cause a further increase in collector current. In the CE amplifier, this means that V_{CE} is very close to zero and I_C is at a maximum. In the circuit of Fig 3.35B, imagine a short circuit across the collector-to-emitter so that all of V_{CC} appears across R_L. Increasing base current will not result in any additional collector current. At cutoff, base current is so small that V_{CE} is at a maximum because no collector current is flowing and further reductions in base current cause no additional increase in V_{CE}.

In this simple circuit, $V_{CE} = V_{CC} - I_C R_L$ and the relationship between I_C and V_{CE} is a straight line between saturation and cutoff. This is the circuit's *load line* and it has a slope of $R_L = (V_{CC} - V_{CE}) / I_C$. No matter what value of base current is flowing in the transistor, the resulting combination of I_C and V_{CE} will be somewhere on the load line.

With no input signal to this simple circuit, the transistor is at cutoff where $I_C = 0$ and $V_{CE} = V_{CC}$. As the input signal increases so that base current gets larger, the operating point begins to move along the load line to the left, so that I_C increases and the voltage drop across the load, $I_C R_L$, increases, reducing V_{CE}. Eventually, the input signal will cause enough base current to flow that saturation is reached, where $V_{CE} \approx 0$ (typically 0.1 to 0.3 V for silicon transistors) and $I_C \approx V_{CC} / R_L$. If R_L is made smaller, the load line will become steeper and if R_L increases, the load line's slope is reduced.

This simple circuit cannot reproduce negative input signals because the transistor is already in cutoff with no input signal. In addition, the shape and spacing of the characteristic curves show that the transistor responds nonlinearly when close to saturation and cutoff (the nonlinear regions) than it does in the middle of the curves (the linear or active region). Biasing is required so that the

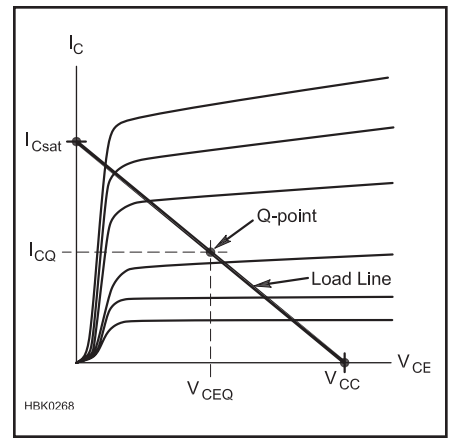

Fig 3.40 — A load line. A circuit's load line shows all of the possible operating points with the specific component values chosen. If there is no input signal, the operating point is the quiescent or Q-point.

circuit does not operate in nonlinear regions, distorting the signal as shown in Fig 3.6.

If the circuit behaves differently for ac signals than for dc signals, a separate *ac load line* can be drawn as discussed below in the section "AC Performance" for the common-emitter amplifier. For example, in the preceding circuit, if R_L is replaced by a circuit that includes inductive or capacitive reactance, ac collector current will result in a different voltage drop across the circuit than will dc collector current. This causes the slope of the ac load line to be different than that of the dc load line.

The ac load line's slope will also vary with frequency, although it is generally treated as constant over the range of frequencies for which the circuit is designed to operate. The ac and dc load lines intersect at the circuit's Q-point because the circuit's ac and dc operation is the same if the ac input signal is zero.

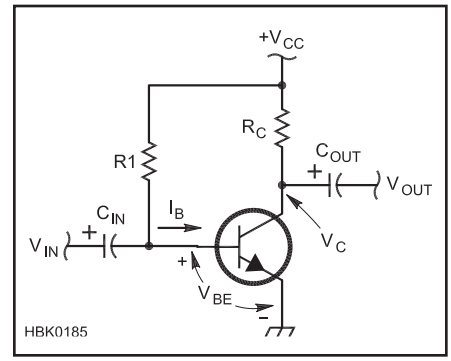

Fig 3.41 — Fixed-bias is the simplest common-emitter (CE) amplifier circuit.

COMMON-EMITTER AMPLIFIER

The *common-emitter amplifier* (*CE*) is the most common amplifier configuration of all — found in analog and digital circuits, from dc through microwaves, made of discrete components and fabricated in ICs. If you understand the CE amplifier, you've made a good start in electronics.

The CE amplifier is used when modest voltage gain is required along with an *input impedance* (the load presented to the circuit supplying the signal to be amplified) of a few hundred to a few kΩ. The current gain of the CE amplifier is the transistor's current gain, β.

The simplest practical CE amplifier circuit is shown in **Fig 3.41**. This circuit includes both coupling and biasing components. The capacitors at the input (C_{IN}) and output (C_{OUT}) block the flow of dc current to the load or to the circuit driving the amplifier. This is an ac-coupled design. These capacitors also cause the gain at very low frequencies to be reduced — gain at dc is zero, for example, because dc input current is blocked by C_{IN}. Resistor R_1 provides a path for bias current to flow into the base, offsetting the collector current from zero and establishing the Q-point for the circuit.

As the input signal swings positive, more current flows into the transistor's base through C_{IN}, causing more current to flow from the collector to emitter as shown by equation 14. This causes more voltage drop across R_L and so the voltage at the collector also drops. The reverse is true when the input signal swings negative. Thus, the output from the CE amplifier is inverted from its input.

Kirchoff's Voltage Law (KVL, see the **Electrical Fundamentals** chapter) is used to analyze the circuit. We'll start with the collector circuit and treat the power supply as a voltage source.

$$V_{cc} = I_c R_c + V_{ce}$$

We can determine the circuit's voltage gain, A_V, from the variation in output voltage caused by variations in input voltage. The output voltage from the circuit at the transistor collector is

$$V_c = V_{CC} - I_c R_c = V_{CC} - \beta I_B R_C \qquad (17)$$

It is also necessary to determine how base current varies with input voltage. Using the transistor's equivalent circuit of Fig 3.37A,

$$I_B = \frac{V_B}{(\beta + 1)\, r_e}$$

so that

$$V_c = V_{CC} - V_B \frac{\beta}{\beta + 1} \times \frac{R_C}{r_e} \qquad (18)$$

We can now determine the circuit's *voltage gain*, the variation in output voltage, ΔV_C, due to variations in input voltage, ΔV_B. Since V_{CC} is constant and β is much greater than 1 in our model:

$$A_V \approx -\frac{R_C}{r_e} \qquad (19)$$

Because r_e is quite small (typically a few ohms, see equation 16), A_V for this circuit can be quite high.

The circuit load line's end-points are $V_{CE} = V_{CC}$ and $I_C = V_{CC} / R_C$. The circuit's Q-point is determined by the collector resistor, R_C, and resistor R_1 that causes bias current to flow into the base. To determine the Q-point, again use KVL starting at the power source and assuming that $V_{BE} = 0.7$ V for a silicon transistor's PN junction when forward-biased.

$$V_{CC} - I_B R_1 = V_B = V_{BE} = 0.7\ \text{V}$$

so

$$I_B = \frac{V_{CC} - 0.7\text{V}}{R_1} \qquad (20)$$

And the Q-point is therefore

$$V_{CEQ} = V_{CC} - \beta I_B R_C \qquad (21a)$$

and

$$I_{CQ} = \beta I_B \qquad (21b)$$

The actual V_{BE} of silicon transistors will vary from 0.6-0.75 V, depending on the level of base current, but 0.7 V is a good compromise value and widely used in small-signal, low-frequency design. Use 0.6 V for very low-power amplifiers and 0.75 V (or more) for high-current switch circuits.

This simple *fixed-bias* circuit is a good introduction to basic amplifiers, but is not entirely practical because the bias current will change due to the change of V_{BE} with temperature, leading to thermal instability. In addition, the high voltage gain can lead to instability due to positive feedback at high frequencies.

To stabilize the dc bias, **Fig 3.42** adds R_E, a technique called *emitter degeneration* because the extra emitter resistance creates negative feedback: as base current rises, so does V_E, the voltage drop across R_E. This reduces the base-emitter voltage and lowers base current. The benefit of emitter degeneration comes from stabilizing the circuit's dc behavior with temperature, but there is a reduction in gain because of the increased resistance in the emitter circuit. Ignoring the effect of R_L for the moment,

$$A_V \approx -\frac{R_C}{R_E} \qquad (22)$$

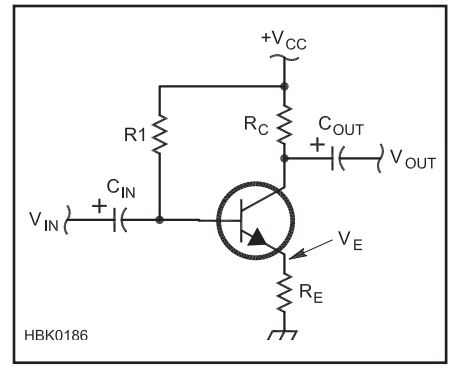

Fig 3.42 — Emitter degeneration. Adding R_E produces negative feedback to stabilize the bias point against changes due to temperature. As the bias current increases, the voltage drop across R_E also increases and causes a decrease in V_{BE}. This reduces bias current and stabilizes the operating point.

In effect, the load resistor is now split between R_C and R_E, with part of the output voltage appearing across each because the changing current flows through both resistors. While somewhat lower than with the emitter connected directly to ground, voltage gain becomes easy to control because it is the ratio of two resistances.

Biasing the CE Amplifier

Fig 3.43 adds R_1 and R_2 from a voltage divider that controls bias current by fixing the base voltage at:

$$V_B = V_{CC} \frac{R_2}{R_1 + R_2}$$

Since

$$V_B = V_{BE} + (I_B + I_C)\, R_E = 0.7\ \text{V} + (\beta + 1)\, I_B R_E$$

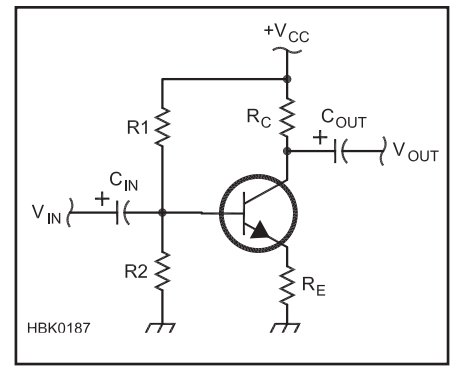

Fig 3.43 — Self-bias. R1 and R2 form a voltage divider to stabilize V_B and bias current. A good rule of thumb is for current flow through R1 and R2 to be 10 times the desired bias current. This stabilizes bias against changes in transistor parameters and component values.

base current is

$$I_B = \frac{V_B - 0.7\ V}{(\beta + 1)\, R_E} \qquad (23a)$$

and Q-point collector current becomes for high values of β

$$I_{CQ} = \beta I_B \approx \frac{V_{CC} \dfrac{R_1}{R_2} - 0.7}{R_E} \qquad (23b)$$

This is referred to as *self-bias* in which the Q-point is much less sensitive to variations in temperature that affect β and V_{BE}.

A good rule-of-thumb for determining the sum of R_1 and R_2 is that the current flowing through the voltage divider, $V_{CC}/(R_1+R_2)$, should be at least 10 times the bias current, I_B. This keeps V_B relatively constant even with small changes in transistor parameters and temperature.

Q-point V_{CEQ} must now also account for the voltage drop across both R_C and R_E,

$$V_{CEQ} \approx V_{CC} - \beta I_B\,(R_C + R_E) \qquad (24)$$

More sophisticated techniques for designing the bias networks of bipolar transistor circuits are described in reference texts listed at the end of this chapter.

Input and Output Impedance

With R_E in the circuit, the small changes in input current, I_B, when multiplied by the transistor's current gain, β, cause a large voltage change across R_E equal to $\beta I_B R_E$. This is the same voltage drop as if I_B was flowing through a resistance equal to βR_E. Thus, the effect of β on impedance at the base is to multiply the emitter resistance, R_E by β, as well. At the transistor's base,

$$Z_B \approx (\beta + 1)\, R_E$$

The input source doesn't just drive the base, of course, it also has to drive the combination of R1 and R2, the biasing resistors. From an ac point of view, both R1 and R2 can be considered as connected to ac ground and they can be treated as if they were connected in parallel. When R1//R2 are considered along with the transistor base impedance, Z_B, the impedance presented to the input signal source is:

$$Z_{IN} = R1\,//\,R2\,//\,(\beta + 1)\, R_E \qquad (25)$$

where // designates "in parallel with."

For both versions of the CE amplifier, the collector output impedance is high enough that

$$Z_{OUT} \approx R_C \qquad (26)$$

CE Amplifier Design Example

The general process depends on the circuit's primary performance requirements, including voltage gain, impedances, power consumption and so on. The most common situation in which a specific voltage gain is required and the circuit's Q-point has been selected based on the transistor to be used, and using the circuit of Fig 3.43, is as follows:

1) Start by determining the circuit's design constraints and assumptions: power supply V_{CC} = 12 V, transistor β = 150 and V_{BE} = 0.7 V. State the circuit's design requirements: $|A_V|$ = 5, Q-point of I_{CQ} = 4 mA and V_{CEQ} = 5 V. (A $V_{CEQ} \approx ½\ V_{CC}$ allows a wide swing in output voltage with the least distortion.)

2) Determine the values of R_C and R_E using equation 24: $R_C + R_E = (V_{CC} - V_{CEQ})/I_{CQ}$ = 1.75 kΩ

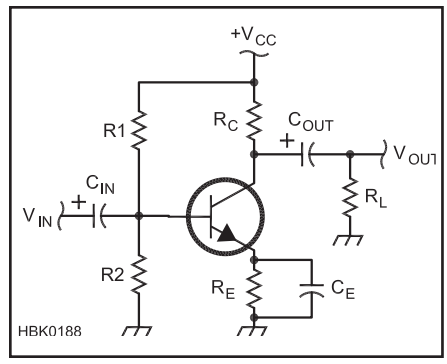

Fig 3.44 — Emitter bypass. Adding C_E allows ac currents to flow "around" R_E, returning ac gain to the value for the fixed-bias circuit while allowing R_E to stabilize the dc operating point.

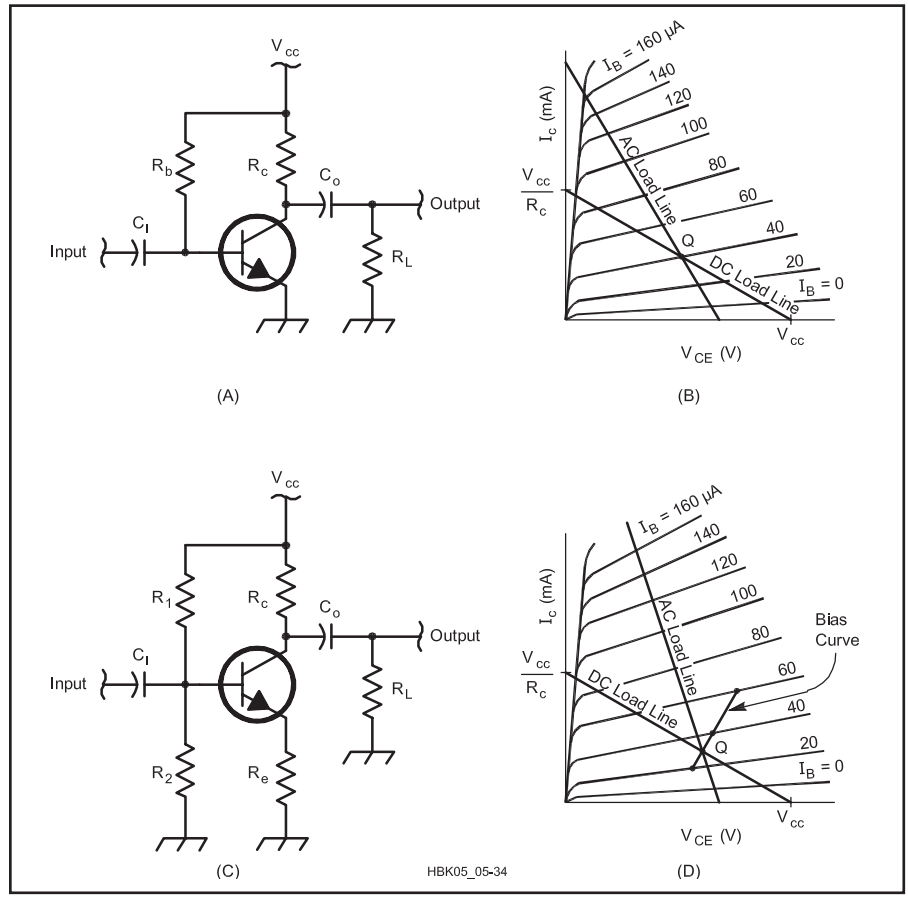

Fig 3.45 — Amplifier biasing and ac and dc load lines. (A) Fixed bias. Input signal is ac coupled through C_i. The output has a voltage that is equal to $V_{CC} - I_C \times R_C$. This signal is ac coupled to the load, R_L, through C_O. For dc signals, the entire output voltage is based on the value 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 // R_L). The quiescent-point, Q, is based on the base bias current with no input signal applied and the point 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. If C_E was added as in Fig 3.44, the slope of the ac load line would increase further.

3) $A_V = -5$, so from equation 22: $R_C = 5\,R_E$, thus $6R_E = 1.75\ k\Omega$ and $R_E = 270\ \Omega$

4) Use equation 14 to determine the base bias current, $I_B = I_{CQ}/\beta = 27\ \mu A$. By the rule of thumb, current through R_1 and $R_2 = 10\ I_B = 270\ \mu A$

5) Use equation 23 to find the voltage across $R_2 = V_B = V_{BE} + I_C\ R_E = 0.7 + 4\ mA\ (0.27\ k\Omega) = 1.8\ V$. Thus, $R_2 = 1.8\ V / 270\ \mu A = 6.7\ k\Omega$

6) The voltage across $R_1 = V_{CC} - V_{R2} = 12 - 1.8 = 10.2\ V$ and $R_1 = 10.2\ V / 270\ \mu A = 37.8\ k\Omega$

Use the nearest standard values ($R_E = 270\ \Omega$, $R_1 = 39\ k\Omega$, $R_2 = 6.8\ k\Omega$) and circuit behavior will be close to that predicted.

AC Performance

To achieve high gains for ac signals while maintaining dc bias stability, the *emitter-bypass capacitor*, C_E, is added in **Fig 3.44** to provide a low impedance path for ac signals around R_E. In addition, a more accurate formula for ac gain includes the effect of adding R_L through the dc blocking capacitor at the collector. In this circuit, the ac voltage gain is

$$A_V \approx -\frac{R_C\,//\,R_L}{r_e} \qquad (27)$$

Because of the different signal paths for ac and dc signals, the ac performance of the circuit is different than its dc performance. This is illustrated in **Fig 3.45** by the intersecting load lines labeled "AC Load Line" and "DC Load Line." The load lines intersect at the Q-point because at that point dc performance is the same as ac performance if no ac signal is present.

The equation for ac voltage gain assumes that the reactances of C_{IN}, C_{OUT}, and C_E are small enough to be neglected (less than one-tenth that of the components to which they are connected at the frequency of interest). At low frequencies, where the capacitor reactances become increasingly large, voltage gain is reduced. Neglecting C_{IN} and C_{OUT}, the low-frequency 3 dB point of the amplifier, f_L, occurs where $X_{CE} = 0.414\ r_e$,

$$f_L = \frac{2.42}{2\pi r_e C_E} \qquad (28)$$

This increases the emitter circuit impedance such that A_V is lowered to 0.707 of its mid-band value, lowering gain by 3 dB. (This ignores the effects of C_{IN} and C_{OUT}, which will also affect the low-frequency performance of the circuit.)

The ac input impedance of this version of the CE amplifier is lower because the effect of R_E on ac signals is removed by the bypass capacitor. This leaves only the internal emitter resistance, r_e, to be multiplied by the current gain,

$$Z_{IN} \approx R1\,//\,R2\,//\,\beta r_e \qquad (29)$$

and

$$Z_{OUT} \approx R_C \qquad (30)$$

again neglecting the reactance of the three capacitors.

The power gain, A_P, for the emitter-bypassed CE amplifier is the ratio of output power, V_O^2/Z_{OUT}, to input power, V_I^2/Z_{IN}. Since $V_O = V_I A_V$,

$$A_P = A_V{}^2\,\frac{R1\,//\,R2\,//\,\beta r_e}{R_C} \qquad (31)$$

COMMON-COLLECTOR (EMITTER-FOLLOWER) AMPLIFIER

The common-collector (CC) amplifier in **Fig 3.46** is also known as the *emitter-follower* (EF) because the emitter voltage "follows" the input voltage. In fact, the amplifier has no voltage gain (voltage gain ≈ 1), but is used as a buffer amplifier to isolate sensitive circuits such as oscillators or to drive low-impedance loads, such as coaxial cables. As in the CE amplifier, the current gain of the emitter-follower is the transistor's current gain, β. It has relatively high input impedance with low output impedance and good power gain.

The collector of the transistor is connected directly to the power supply without a resistor and the output signal is created by the voltage drop across the emitter resistor. There is no 180° phase shift as seen in the CE amplifier; the output voltage follows the input signal with 0° phase shift because increases in the input signal cause increases in emitter current and the voltage drop across the emitter resistor.

The EF amplifier has high input impedance: following the same reasoning as for the CE amplifier with an unbypassed emitter resistor as described by equation 25,

$$Z_{IN} = R1\,//\,R2\,//\,(\beta + 1)\,R_E \qquad (32)$$

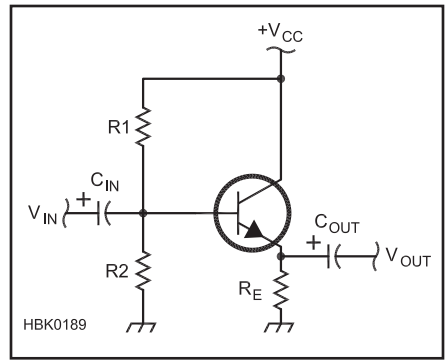

Fig 3.46 — Emitter follower (EF) amplifier. The voltage gain of the EF amplifier is unity. The amplifier has high input impedance and low output impedance, making it a good choice for use as a buffer amplifier.

The impedance at the EF amplifier's output consists of the emitter resistance, R_E, in parallel with the series combination of the internal emitter resistance, r_e, the parallel combination of biasing resistors R1 and R2, and the internal impedance of the source providing the input signal. In this case, current gain acts to *reduce* the effect of the input circuit's impedance on output impedance:

$$Z_{OUT} = \left[\frac{R_S\,//\,R1\,//\,R2}{(\beta + 1)}\right]\,//\,R_E \qquad (33)$$

In practice, with transistor β of 100 or more, $Z_{OUT} \approx R_S/\beta$. However, if a very high impedance source is used, such as an crystal microphone element or photodetector, the effects of the biasing and emitter resistors must be considered.

Because the voltage gain of the EF amplifier is unity, the power gain is simply the ratio of input impedance to output impedance,

$$A_P \approx \frac{R1\,//\,R2\,//\,(\beta + 1)\,R_E}{R_E} \qquad (34)$$

EF Amplifier Design Example

The following procedure is similar to the design procedure in the preceding section for the CE amplifier, except $A_V = 1$.

1) Start by determining the circuit's design constraints and assumptions: $V_{cc} = 12\ V$ (the power supply voltage), a transistor's β of 150 and $V_{BE} = 0.7\ V$. State the circuit's design requirements: Q-point of $I_{CQ} = 5\ mA$ and $V_{CEQ} = 6\ V$.

2) $R_E = (V_{CC} - V_{CEQ})/I_{CQ} = 1.2\ k\Omega$

3) Base current, $I_B = I_{CQ}/\beta = 33\ \mu A$

4) Current through R1 and R2 = $10\ I_B = 330\ \mu A$ ($10\ I_B$ rule of thumb as with the CE amplifier)

5) Voltage across $R2 = V_{BE} + I_C\,R_E = 0.7 + 5\ mA\ (1.2\ k\Omega) = 6.7\ V$ and $R2 = 6.7\ V / 330\ \mu A = 20.3\ k\Omega$ (use the standard value 22 kΩ)

6) Voltage across $R1 = V_{CC} - 6.7\ V = 5.3\ V$

7) $R1 = 5.3\ V / 330\ \mu A = 16.1\ k\Omega$ (use 16 kΩ)

8) $Z_{IN} = R1\ //\ R2\ //\ R_E(\beta + 1) \approx 8.5\ k\Omega$

COMMON-BASE AMPLIFIER

The common-base (CB) amplifier of **Fig 3.47** is used where low input impedance is needed, such as for a receiver preamp with a coaxial feed line as the input signal source. Complementary to the EF amplifier, the CB amplifier has unity current gain and high output impedance.

Fig 3.47A shows the CB circuit as it is usually drawn, without the bias circuit resistors connected and with the transistor symbol

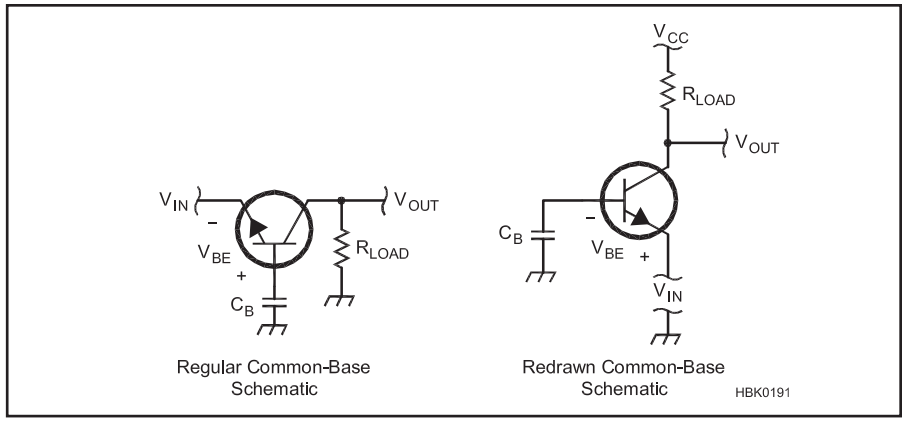

Fig 3.47 — The common-base (CB) amplifier is often drawn in an unfamiliar style (A), but is more easily understood when drawn similarly to the CE and EF amplifiers (B). The input signal to the CB amplifier is applied to the emitter instead of the base.

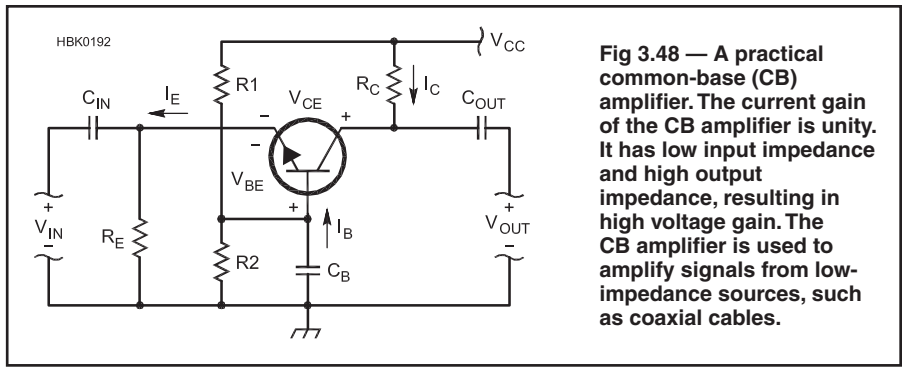

Fig 3.48 — A practical common-base (CB) amplifier. The current gain of the CB amplifier is unity. It has low input impedance and high output impedance, resulting in high voltage gain. The CB amplifier is used to amplify signals from low-impedance sources, such as coaxial cables.

turned on its side from the usual orientation so that the emitter faces the input. In order to better understand the amplifier's function, Fig 3.47B reorients the circuit in a more familiar style. We can now clearly see that the input has just moved from the base circuit to the emitter circuit.

Placing the input in the emitter circuit allows it to cause changes in the base-emitter current as for the CE and EF amplifiers, except that for the CB amplifier a positive change in input amplitude reduces base current by lowering V_{BE} and raising V_C. As a result, the CB amplifier is noninverting, just like the EF, with output and input signals in-phase.

A practical circuit for the CB amplifier is shown in **Fig 3.48**. From a dc point of view (replace the capacitors with open circuits), all of the same resistors are there as in the CE amplifier. The input capacitor, C_{IN}, allows the dc emitter current to bypass the ac input signal source and C_B places the base at ac ground while allowing a dc voltage for biasing. (All voltages and currents are labeled to aid in understanding the different orientation of the circuit.)

The CB amplifier's current gain,

$$A_I = \frac{i_C}{i_E} = \frac{\beta}{\beta + 1} \tag{35}$$

is relatively independent of input and output impedance, providing excellent isolation between the input and output circuits. Output impedance does not affect input impedance, allowing the CB amplifier to maintain stable input impedance, even with a changing load.

Following reasoning similar to that for the CE and EF amplifiers for the effect of current gain on R_E, we find that input impedance for the CB amp is

$$Z_{IN} = R_E \,//\, (\beta + 1)\, r_e \tag{36}$$

The output impedance for the CB amplifier is approximately

$$Z_{OUT} = R_C \,//\, \frac{1}{h_{oe}} \approx R_C \tag{37}$$

where h_{oe} is the transistor's collector output admittance. The reciprocal of h_{oe} is in the range of 100 kΩ at low frequencies.

Voltage gain for the CB amplifier is

$$A_V \approx \frac{R_C \,//\, R_L}{r_e} \tag{38}$$

As a result, the usual function of the CB amplifier is to convert input current from a low-impedance source into output voltage at a higher impedance.

Power gain for the CB amplifier is approximately the ratio of output to input impedance,

$$A_P \approx \frac{R_C}{R_E \,//\, (\beta + 1)\, r_e} \tag{39}$$

CB Amplifier Design Example

Because of its usual function as a current-to-voltage converter, the design process for the CB amplifier begins with selecting R_E and A_V, assuming that R_L is known.

1) Start by determining the circuit's design constraints and assumptions: V_{cc} = 12 V (the power supply voltage), a transistor's β of 150 and V_{BE} = 0.7 V. State the circuit's design requirements: R_E = 50 Ω, R_L = 1 kΩ, I_{CQ} = 5 mA, V_{CEQ} = 6 V.

2) Base current, $I_B = I_{CQ}/\beta$ = 33 μA

3) Current through R1 and R2 = 10 I_B = 330 μA (10 I_B rule of thumb as with the CE amplifier)

4) Voltage across R2 = V_{BE} + I_C R_E = 0.7 + 5 mA (1.2 kΩ) = 6.7 V and R2 = 6.7 V / 330 μA = 20.3 kΩ (use the standard value 22 kΩ)

5) Voltage across R1 = V_{CC} – 6.7 V = 5.3 V

6) R1 = 5.3 V / 330 μA = 16.1 kΩ (use 16 kΩ)

7) $R_C = (V_{CC} - I_{CQ}\, R_E - V_{CEQ}) / I_{CQ}$ = (12 – 0.25 – 5) / 5 mA = 1.35 kΩ (use 1.5 kΩ)

8) A_V = (1.5 kΩ // 1 kΩ) / (26 / I_E) = 115

3.5.4 FET Amplifiers

The field-effect transistor (FET) is widely used in radio and RF applications. There are many types of FETs, with JFETs (junction FET) and MOSFETs (metal-oxide-semiconductor FET) being the most common types. In this section we will discuss JFETs, with the understanding that the use of MOSFETs is similar. (This discussion is based on N-channel JFETs, but the same discussion applies to P-channel devices if the bias voltages and currents are reversed.)

SMALL-SIGNAL FET MODEL

While bipolar transistors are most commonly viewed as current-controlled devices, the JFET, however, is purely a voltage-controlled device — at least at low frequencies. The input gate is treated as a reverse-biased diode junction with virtually no current flow. As with the bipolar transistor amplifier circuits, the circuits in this section are very basic and more thorough treatments of FET amplifier design can be found in the references at the end of the chapter.

The operation of an N-channel JFET for both biasing and signal amplification can be

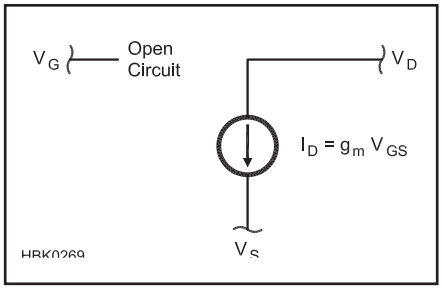

Fig 3.49 — Small-signal FET model. The FET can be modeled as a voltage-controlled current source in its saturation region. The gate is treated as an open-circuit due to the reverse-biased gate-channel junction.

characterized by the following equation:

$$I_D = I_{DSS}\left(\frac{V_P - V_{GS}}{V_P}\right)^2 \quad (40)$$

where

I_{DSS} = drain saturation current
V_{GS} = the gate-source voltage
V_P = the pinch-off voltage.

I_{DSS} is the maximum current that will flow between the drain and source for a given value of drain-to-source voltage, V_{DS}. Note that the FET is a *square-law* device in which output current is proportional to the square of an input voltage. (The bipolar transistor's output current is an exponential function of input current.)

Also note that V_{GS} in this equation has the opposite sense of the bipolar transistor's V_{BE}. For this device, as V_{GS} increases (making the source more positive than the gate), drain current decreases until at V_P the channel is completely "pinched-off" and no drain current flows at all. This equation applies only if V_{GS} is between 0 and V_P. JFETs are seldom used with the gate-to-channel diode forward-biased ($V_{GS} < 0$).

None of the terms in Equation 40 depend explicitly on temperature. Thus, the FET is relatively free of the thermal instability exhibited by the bipolar transistor. As temperature increases, the overall effect on the JFET is to reduce drain current and to stabilize the operation.

The small-signal model used for the JFET is shown in **Fig 3.49**. The drain-source channel is treated as a current source whose output is controlled by the gate-to-source voltage so that $I_D = g_m V_{GS}$. The high input impedance allows the input to be modeled as an open circuit (at low frequencies). This simplifies circuit modeling considerably as biasing of the FET gate can be done by a simple voltage divider without having to consider the effects of bias current flowing in the JFET itself.

The FET has characteristic curves as shown in Fig 3.25 that are similar to those of a bipolar transistor. The output characteristic curves are similar to those of the bipolar transistor, with the horizontal axis showing V_{DS} instead of V_{CE} and the vertical axis showing I_D instead of I_C. Load lines, both ac and dc, can be developed and drawn on the output characteristic curves in the same way as for bipolar transistors.

The set of characteristic curves in Fig 3.25 are called *transconductance response curves* and they show the relationship between input voltage (V_{GS}), output current (I_D) and output voltage (V_{DS}). The output characteristic curves show I_D and V_{DS} for different values of V_{GS} and are similar to the BJT output characteristic curve. The input characteristic curves show I_D versus V_{GS} for different values of V_{DS}.

MOSFETs act in much the same way as JFETs when used in an amplifier. They have a higher input impedance, due to the insulation between the gate and channel. The insulated gate also means that they can be operated with the polarity of V_{GS} such that a JFET's gate-channel junction would be forward biased, beyond V_P. Refer to the discussion of depletion- and enhancement-mode MOSFETs in the previous section on Practical Semiconductors.

THREE BASIC FET AMPLIFIERS

Just as for bipolar transistor amplifiers, there are three basic configurations of amplifiers using FETs; the *common-source* (CS)(corresponding to the common-emitter), *common-drain* (CD) or *source-follower* (corresponding to the emitter-follower) and the *common-gate* (CG) (corresponding to the common base). Simple circuits and design methods are presented here for each, assuming low-frequency operation and a simple, voltage-controlled current-source model for the FET. Discussion of the FET amplifier at high frequencies is available in the **RF Techniques** chapter and an advanced discussion of biasing FET amplifiers and their large-signal behavior is contained on the companion CD-ROM.

COMMON-SOURCE AMPLIFIER

The basic circuit for a common-source FET amplifier is shown in **Fig 3.50**. In the ohmic region (see the previous discussion on FET characteristics), the FET can be treated as a variable resistance as shown in **Fig 3.50A** where V_{GS} effectively varies the resistance between drain and source. However, most FET amplifiers are designed to operate in the saturation region and the model of Fig 3.49 is used in the circuit of **Fig 3.50B** in which,

$$I_D = g_m V_{GS} \quad (41)$$

where g_m is the FET's forward transconductance.

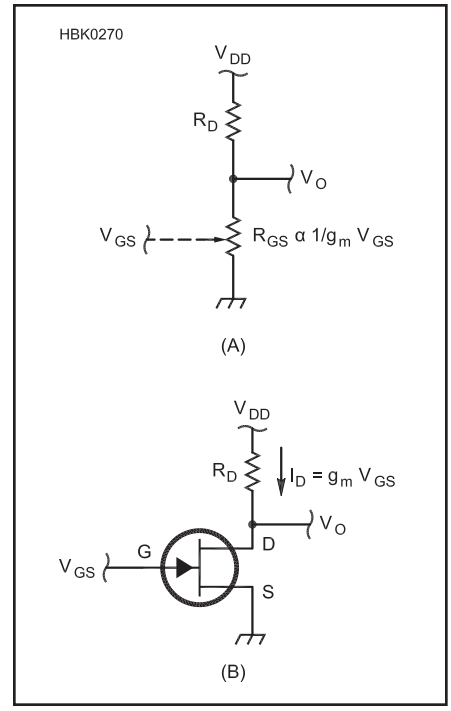

Fig 3.50 — In the ohmic region (A), the FET acts like a variable resistance, R_{DS}, with a value controlled by V_{GS}. The alpha symbol (α) means "is proportional to". In the saturation region (B), the drain-source channel of the FET can be treated like a current source with $I_D = g_m V_{GS}$.

If V_O is measured at the drain terminal (just as the common-emitter output voltage is measured at the collector), then

$$\Delta V_O = -g_m \Delta V_{GS} R_D$$

The minus sign results from the output voltage decreasing as drain current and the voltage drop across R_D increases, just as in the CE amplifier. Like the CE amplifier, the input and output voltages are thus 180° out of phase. Voltage gain of the CS amplifier in terms of transconductance and the drain resistance is:

$$A_V = -g_m R_D \quad (42)$$

As long as $V_{GS} < 0$, this simple CS amplifier's input impedance at low frequencies is that of a reverse-biased diode — nearly infinite with a very small leakage current. Output impedance of the CS amplifier is approximately R_D because the FET drain-to-source channel acts like a current source with very high impedance.

$$Z_{IN} = \infty \text{ and } Z_{OUT} \approx R_D \quad (43)$$

As with the BJT, biasing is required to create a Q-point for the amplifier that allows reproduction of ac signals. The practical circuit

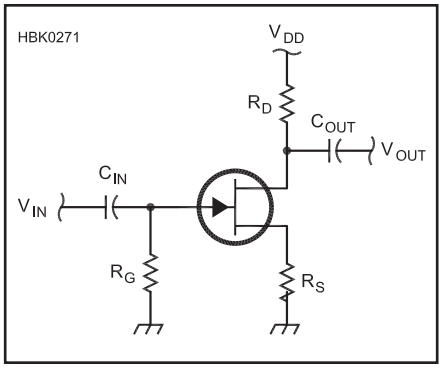

Fig 3.51 — Common-source (CS) amplifier with self-bias.

of Fig 3.50B is used to allow control of V_{GS} bias. A load line is drawn on the JFET output characteristic curves, just as for a bipolar transistor circuit. One end point of the load line is at $V_{DS} = V_{DD}$ and the other at $I_{DS} = V_{DD} / R_D$. The Q-point for the CS amplifier at I_{DQ} and V_{DSQ} is thus determined by the dc value of V_{GS}.

The practical JFET CS amplifier shown in **Fig 3.51** uses self-biasing in which the voltage developed across the source resistor, R_S, raises V_S above ground by $I_D R_S$ volts and $V_{GS} = -I_D R_S$ since there is no dc drop across R_G. This is also called *source degeneration*. The presence of R_S changes the equation of voltage gain to

$$A_V = -\frac{g_m R_D}{1 + g_m R_S} \approx -\frac{R_D}{R_S} \qquad (44)$$

The value of R_S is obtained by substituting $V_{GS} = I_D R_S$ into Equation 40 and solving for R_S as follows:

$$R_S = \frac{V_P}{I_{DQ}} \left(1 - \sqrt{\frac{I_{DQ}}{I_{DSS}}}\right) \qquad (45)$$

Once R_S is known, the equation for voltage gain can be used to find R_D.

The input impedance for the circuit of Fig 3.51 is essentially R_G. Since the gate of the JFET is often ac coupled to the input source through a dc blocking capacitor, C_{IN}, a value of 100 kΩ to 1 MΩ is often used for R_G to provide a path to ground for gate leakage current. If R_G is omitted in an ac-coupled JFET amplifier, a dc voltage can build up on the gate from leakage current or static electricity, affecting the channel conductivity.

$$Z_{IN} = R_G \qquad (46)$$

Because of the high impedance of the drain-source channel in the saturation region, the output impedance of the circuit is:

$$Z_{OUT} \approx R_D \qquad (47)$$

Designing the Common-Source Amplifier

The design of the CS amplifier begins with selection of a Q-point $I_{DQ} < I_{DSS}$. Because of variations in V_P and I_{DSS} from JFET to JFET, it may be necessary to select devices individually to obtain the desired performance.

1) Start by determining the circuit's design constraints and assumptions: $V_{DD} = 12$ V (the power supply voltage) and the JFET has an I_{DSS} of 35 mA and a V_P of –3.0V, typical of small-signal JFETs. State the circuit's design requirements: $|A_V| = 10$ and $I_{DQ} = 10$ mA.

2) Use equation 45 to determine R_S = 139 Ω

3) Since $|A_V| = 10$, $R_D = 10\ R_S = 1390$ Ω.

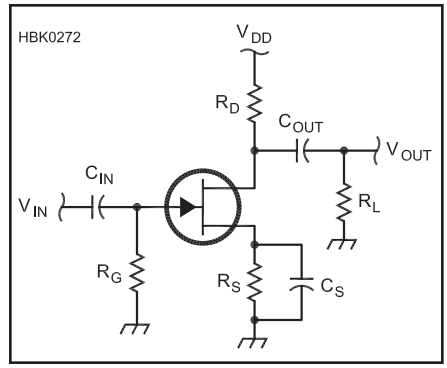

Fig 3.52 — Common-source amplifier with source bypass capacitor, C_S, to increase voltage gain without affecting the circuit's dc performance.

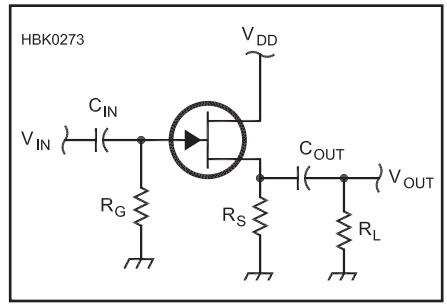

Fig 3.53 — Similar to the EF amplifier, the common-drain (CD) amplifier has a voltage gain of unity, but makes a good buffer with high input and low output impedances.

Use standard values for R_S = 150 Ω and R_D = 1.5 kΩ.

AC Performance

As with the CE bipolar transistor amplifier, a bypass capacitor can be used to increase ac gain while leaving dc bias conditions unchanged as shown in **Fig 3.52**. In the case of the CS amplifier, a *source bypass* capacitor is placed across R_S and the load, R_L, connected through a dc blocking capacitor. In this circuit voltage gain becomes:

$$A_V = -g_m (R_D \,//\, R_L) \qquad (48)$$

Assuming C_{IN} and C_{OUT} are large enough to ignore their effects, the low-frequency cutoff frequency of the amplifier, f_L, is approximately where $X_{CS} = 0.707\ (R_D \,//\, R_L)$,

$$f_L = \frac{1.414}{2\pi (R_D \,//\, R_L) C_S} \qquad (49)$$

as this reduces A_V to 0.707 of its mid-band value, resulting in a 3 dB drop in output amplitude.

The low-frequency ac input and output impedances of the CS amplifier remain

$$Z_{IN} = R_G \text{ and } Z_{OUT} \approx R_D \qquad (50)$$

COMMON-DRAIN (SOURCE-FOLLOWER) AMPLIFIER

The *common-drain* amplifier in **Fig 3.53** is also known as a *source-follower* (SF) because the voltage gain of the amplifier is unity, similar to the emitter follower (EF) bipolar transistor amplifier. The SF amplifier is used primarily as a buffer stage and to drive low-impedance loads.

At low frequencies, the input impedance of the SF amplifier remains nearly infinite. The SF amplifier's output impedance is the source resistance, R_S, in parallel with the impedance of the controlled current source, $1/g_m$.

$$Z_{OUT} = R_S \,//\, \frac{1}{g_m} = \frac{R_S}{g_m R_S + 1} \approx \frac{1}{g_m} \text{ for } g_m R_S >> 1 \qquad (51)$$

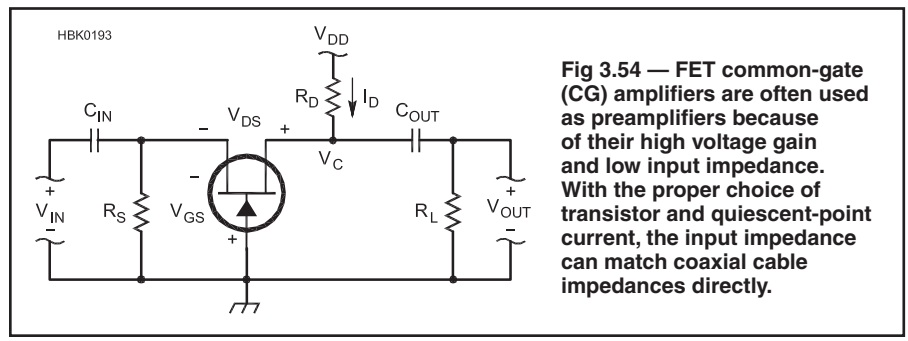

Fig 3.54 — FET common-gate (CG) amplifiers are often used as preamplifiers because of their high voltage gain and low input impedance. With the proper choice of transistor and quiescent-point current, the input impedance can match coaxial cable impedances directly.

Design of the SF amplifier follows essentially the same process as the CS amplifier, with $R_D = 0$.

THE COMMON-GATE AMPLIFIER

The *common-gate* amplifier in **Fig 3.54** has similar properties to the bipolar transistor common-base (CB) amplifier; unity current gain, high voltage gain, low input impedance and high output impedance. (Refer to the discussion of the CB amplifier regarding placement of the input and how the circuit schematic is drawn.) It is used as a voltage amplifier, particularly for low-impedance sources, such as coaxial cable inputs.

The CG amplifier's voltage gain is

$$A_V = g_m(R_D \;//\; R_L) \tag{52}$$

The output impedance of the CG amplifier is very high, we must take into account the output resistance of the controlled current source, r_o. This is analogous to the appearance of h_{oe} in the equation for output impedance of the bipolar transistor CG amplifier.

$$Z_O \approx r_o\,(g_m R_S + 1) \;//\; R_D \tag{53}$$

The CG amplifier input impedance is approximately

$$Z_I = R_S \;//\; \frac{1}{g_m} \tag{54}$$

Because the input impedance is quite low, the cascode circuit described later in the section on buffers is often used to present a higher-impedance input to the signal source.

Occasionally, the value of R_S must be fixed in order to provide a specific value of input impedance. Solving equation 40 for I_{DQ} results in the following equation:

$$I_{DQ} = \frac{V_P}{2R_S{}^2 I_{DSS}}\left(V_P + \sqrt{V_P{}^2 - 4R_S I_{DSS} V_P}\right) - \frac{V_P}{R_S} \tag{55}$$

Designing the Common-Gate Amplifier

Follow the procedure for designing a CS amplifier, except determine the value of R_D as shown in equation 52 for voltage gain above.

1) Start by determining the circuit's design constraints and assumptions: $V_{DD} = 12$ V (the power supply voltage) and the JFET has a g_m of 15 mA/V, an I_{DSS} of 60 mA and $V_P = -6$ V. State the circuit's design requirements: $A_V = 10$, $R_L = 1$ kΩ and $R_S = 50$ Ω.

2) Solve equation 52 for R_D: $10 = 0.015 \times R_D//R_L$, so $R_D//R_L = 667$ Ω. $R_D = 667\,R_L / (R_L - 667) = 2$ kΩ.

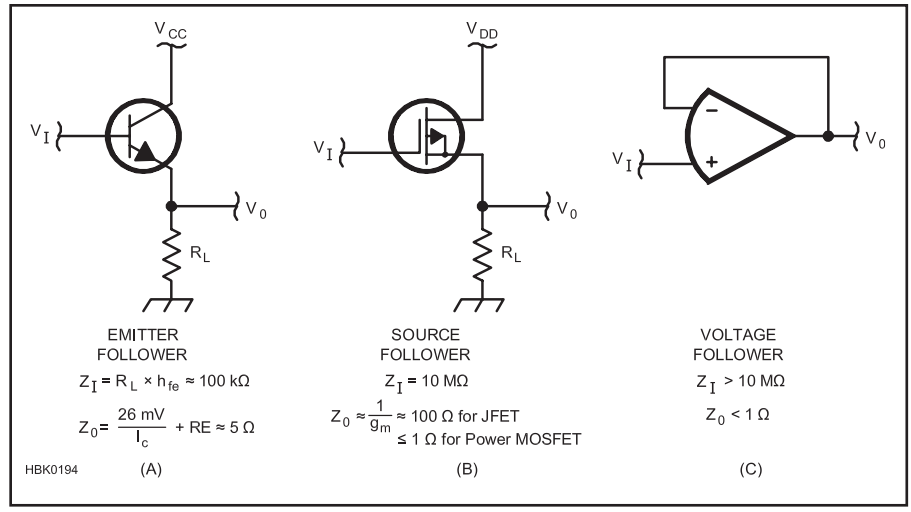

Fig 3.55 — Common buffer stages and some typical input (Z_I) and output (Z_O) impedances. (A) Emitter follower, made with an NPN bipolar transistor; (B) Source follower, made with an FET; and (C) Voltage follower, made with an operational amplifier. All of these buffers are terminated with a load resistance, R_L, and have an output voltage that is approximately equal to the input voltage (gain ≈ 1).

3) I_{DQ} is determined from equation 55: $I_{DQ} = 10$ mA. If I_{DQ} places the Q-point in the ohmic region, reduce A_V and repeat the calculations.

3.5.5 Buffer Amplifiers

Fig 3.55 shows common forms of buffers with low-impedance outputs: the *emitter follower* using a bipolar transistor, the *source follower* using a field-effect transistor and the *voltage follower*, using an operational amplifier. (The operational amplifier is discussed later in this chapter.) These circuits are called "followers" because the output "follows" the input very closely with approximately the same voltage and little phase shift between the input and output signals.

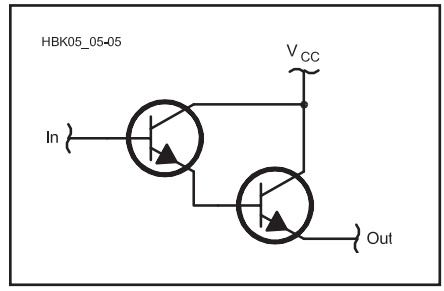

Fig 3.56 — Darlington pair made with two emitter followers. Input impedance, Z_I, is far higher than for a single transistor and output impedance, Z_O, is nearly the same as for a single transistor. DC biasing has been omitted for simplicity.

3.5.6 Cascaded Buffers

THE DARLINGTON PAIR

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 transistors connected as emitter followers as shown in **Fig 3.56.** The current gain of the Darlington pair is the product of the current gains for the two transistors, $\beta_1 \times \beta_2$.

What makes the Darlington pair so useful is that its input impedance is equal to the load impedance times the current gain, effectively multiplying the load impedance;

$$Z_I = Z_{LOAD} \times \beta_1 \times \beta_2 \tag{56}$$

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

$$Z_I = 15\ \text{k}\Omega \times 100 \times 100 = 150\ \text{M}\Omega$$

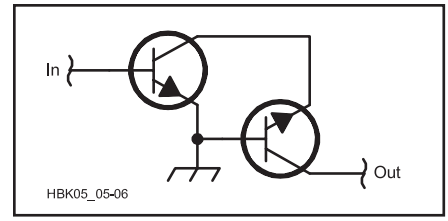

Fig 3.57 — Cascode buffer made with two NPN bipolar transistors has a medium input impedance and high output impedance. DC biasing has been omitted for simplicity.

This impedance places almost no load on the circuit connected to the Darlington pair's input. The shunt capacitance at the input of real transistors can lower the actual impedance as the frequency increases.

Drawbacks of the Darlington pair include lower bandwidth and switching speed. The extremely high dc gain makes biasing very sensitive to temperature and component tolerances. For these reasons, the circuit is usually used as a switch and not as a linear amplifier.

CASCODE AMPLIFIERS

A common-emitter amplifier followed by a common-base amplifier is called a *cascode buffer*, shown in its simplest form in **Fig 3.57**. (Biasing and dc blocking components are omitted for simplicity — replace the transistors with the practical circuits described earlier.) Cascode stages using FETs follow a common-source amplifier with a common-gate configuration. The input impedance and current gain of the cascode amplifier are approximately the same as those of the first stage. The output impedance of the common-base or –gate stage is much higher than that of the common-emitter or common-source amplifier. The power gain of the cascode amplifier is the product of the input stage current gain and the output stage voltage gain.

As an example, a typical cascode buffer made with BJTs has moderate input impedance (Z_{IN} = 1 kΩ), high current gain (h_{fe} = 50), and high output impedance (Z_{OUT} = 1 MΩ). Cascode amplifiers have excellent input/output isolation (very low unwanted internal feedback), resulting in high gain with good stability. Because of its excellent isolation, the cascode amplifier has little effect on external tuning components. Cascode circuits are often used in tuned amplifier designs for these reasons.

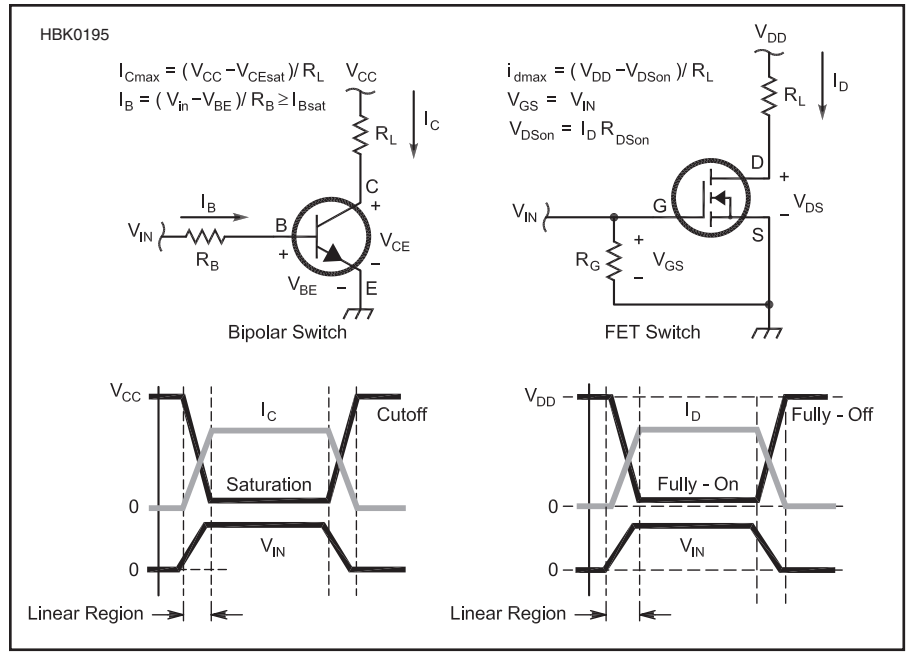

Fig 3.58 — A pair of transistor driver circuits using a bipolar transistor and a MOSFET. The input and output signals show the linear, cutoff and saturation regions.

3.5.7 Using the Transistor as a Switch

When designing amplifiers, the goal was to make the transistor's output a replica of its input, requiring that the transistor stay within its linear region, conducting some current at all times. A switch circuit has completely different properties — its output current is either zero or some maximum value. **Fig 3.58** shows both a bipolar and metal-oxide semiconductor field-effect transistor (or MOSFET) switch circuit. Unlike the linear amplifier circuits, there are no bias resistors in either circuit. When using the bipolar transistor as a switch, it should operate in saturation or in cutoff. Similarly, an FET switch should be either fully-on or fully-off. The figure shows the waveforms associated with both types of switch circuits.

This discussion is written with power control in mind, such as to drive a relay or motor or lamp. The concepts, however, are equally applicable to the much lower-power circuits that control logic-level signals. The switch should behave just the same — switch between on and off quickly and completely — whether large or small.

DESIGNING SWITCHING CIRCUITS

First, select a transistor that can handle the load current and dissipate whatever power is dissipated as heat. Second, be sure that the input signal source can supply an adequate input signal to drive the transistor to the required states, both on and off. Both of these conditions must be met to insure reliable driver operation.

To choose the proper transistor, the load current and supply voltage must both be known. Supply voltage may be steady, but sometimes varies widely. For example, a car's 12 V power bus may vary from 9 to 18 V, depending on battery condition. The transistor must withstand the maximum supply voltage, V_{MAX}, when off. The load resistance, R_L, must also be known. The maximum steady-state current the switch must handle is:

$$I_{MAX} = \frac{V_{MAX}}{R_L} \tag{57}$$

If you are using a bipolar transistor, calculate how much base current is required to drive the transistor at this level of collector current. You'll need to inspect the transistor's data sheet because β decreases as collector current increases, so use a value for β specified at a collector current at or above I_{MAX}.

$$I_B = \frac{I_{MAX}}{\beta} \tag{58}$$

Now inspect the transistor's data sheet values for V_{CEsat} and make sure that this value of I_B is sufficient to drive the transistor fully into saturation at a collector current of I_{MAX}. Increase I_B if necessary — this is I_{Bsat}. The transistor must be fully saturated to minimize heating when conducting load current.

Using the *minimum* value for the input voltage, calculate the value of R_B:

$$R_B = \frac{V_{IN(min)} - V_{BE}}{I_{Bsat}} \tag{59}$$

The minimum value of input voltage must be used to accommodate the *worst-case* combination of circuit voltages and currents.

Designing with a MOSFET is a little easier because the manufacturer usually specifies the value V_{GS} must have for the transistor to be fully on, $V_{GS(on)}$. The MOSFET's gate, being insulated from the conducting channel, acts like a small capacitor of a few hundred pF and draws very little dc current. R_G in Fig 3.58 is required if the input voltage source does not actively drive its output to zero volts when off, such as a switch connected to a positive voltage. The MOSFET won't turn off reliably if its gate is allowed to "float." R_G pulls the gate voltage to zero when the input is open-circuited.

Power dissipation is the next design hurdle. Even if the transistors are turned completely on, they will still dissipate some heat. Just as

for a resistor, for a bipolar transistor switch the power dissipation is:

$$P_D = V_{CE} I_C = V_{CE(sat)}\ I_{MAX}$$

where $V_{CE(sat)}$ is the collector-to-emitter voltage when the transistor is saturated.

Power dissipation in a MOSFET switch is:

$$P_D = V_{DS} I_D = R_{DS(on)}\ I_{MAX}^2 \qquad (60)$$

$R_{DS(on)}$ is the resistance of the channel from drain to source when the MOSFET is on. MOSFETs are available with very low on-resistance, but still dissipate a fair amount of power when driving a heavy load. The transistor's data sheet will contain $R_{DS(on)}$ specification and the V_{GS} required for it to be reached.

Power dissipation is why a switching transistor needs to be kept out of its linear region. When turned completely off or on, either current through the transistor or voltage across it are low, also keeping the product of voltage and current (the power to be dissipated) low. As the waveform diagrams in Fig 3.58 show, while in the linear region, both voltage and current have significant values and so the transistor is generating heat when changing from off to on and vice versa. It's important to make the transition through the linear region quickly to keep the transistor cool.

The worst-case amount of power dissipated during each on-off transition is approximately

$$P_{transition} = \frac{1}{4} V_{MAX}\ I_{MAX}$$

assuming that the voltage and current increase and decrease linearly. If the circuit turns on and off at a rate of f, the total average power dissipation due to switching states is:

$$P_D = \frac{f}{2} V_{MAX}\ I_{MAX} \qquad (61)$$

since there are two on-off transitions per switching cycle. This power must be added to the power dissipated when the switch is conducting current.

Once you have calculated the power the switch must dissipate, you must check to see whether the transistor can withstand it. The manufacturer of the transistor will specify a *free-air dissipation* that assumes no heat-sink and room temperature air circulating freely around the transistor. This rating should be at least 50% higher than your calculated power dissipation. If not, you must either use a larger transistor or provide some means of getting rid of the heat, such as heat sink. Methods of dissipating heat are discussed in the **Electrical Fundamentals** chapter.

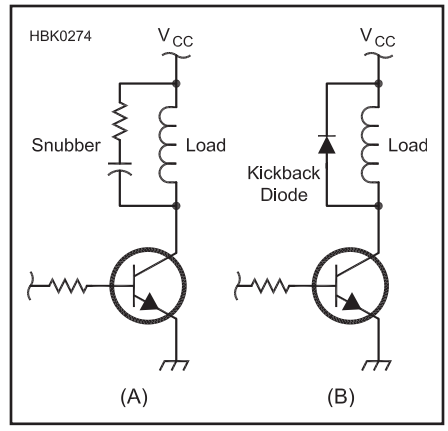

Fig 3.59 — The snubber RC circuit at (A) absorbs energy from transients with fast rise- and fall-times. At (B) a kickback diode protects the switching device when current is interrupted in the inductive load, causing a voltage transient, by conducting the energy back to the power source.

INDUCTIVE AND CAPACITIVE LOADS

Voltage transients for inductive loads, such as solenoids or relays can easily reach dozens of times the power supply voltage when load current is suddenly interrupted. To protect the transistor, the voltage transient must be clamped or its energy dissipated. Where switching is frequent, a series-RC *snubber* circuit (see **Fig 3.59A**) is connected across the load to dissipate the transient's energy as heat. The most common method is to employ a "kickback" diode that is reverse-biased when the load is energized as shown in Fig 3.59B. When the load current is interrupted, the diode routes the energy back to the power supply, clamping the voltage at the power supply voltage plus the diode's forward voltage drop.

Capacitive loads such as heavily filtered power inputs may temporarily act like short circuits when the load is energized or de-energized. The surge current is only limited by the internal resistance of the load capacitance. The transistor will have to handle the temporary overloads without being damaged or overheating. The usual solution is to select a transistor with an I_{MAX} rating greater to the surge current. Sometimes a small current-limiting resistor can be placed in series with the load to reduce the peak surge current at the expense of dissipating power continuously when the load is drawing current.

HIGH-SIDE AND LOW-SIDE SWITCHING

The switching circuits shown in Fig 3.58 are *low-side switches*. This means the switch is connected between the load and ground. A *high-side switch* is connected between the power source and the load. The same concerns for power dissipation apply, but the methods of driving the switch change because of the voltage of the emitter or source of the switching device will be at or near the power supply voltage when the switch is on.

To drive an NPN bipolar transistor or an N-channel MOSFET in a high-side circuit requires the switch input signal to be at least $V_{BE(sat)}$ or $V_{GS(on)}$ *above* the voltage supplied to the load. If the load expects to see the full power supply voltage, the switch input signal will have to be *greater* than the power supply voltage. A small step-up or boost dc-to-dc converter is often used to supply the extra voltage needed for the driver circuit.

One alternate method of high-side switching is to use a PNP bipolar transistor as the switching transistor. A small input transistor turns the main PNP transistor on by controlling the larger transistor's base current. Similarly, a P-channel MOSFET could also be employed with a bipolar transistor or FET acting as its driver. P-type material generally does not have the same high conductivity as N-type material and so these devices dissipate somewhat more power than N-type devices under the same load conditions.

3.5.8 Choosing a Transistor

With all the choices for transistors — Web sites and catalogs can list hundreds — selecting a suitable transistor can be intimidating. Start by determining the maximum voltage (V_{CEO} or $V_{DS(MAX)}$), current (I_{MAX}) and power dissipation ($P_{D(MAX)}$) the transistor must handle. Determine what dc current gain, β, or transconductance, g_m, is required. Then determine the highest frequency at which full gain is required and multiply it by either the voltage or current gain to obtain f_T or h_{fe}. This will reduce the number of choices dramatically.

The chapter on **Component Data and References** has tables of parameters for popular transistors that tend to be the lowest-cost and most available parts, as well. You will find that a handful of part types satisfy the majority of your building needs. Only in very special applications will you need to choose a corresponding special part.

3.6 Operational Amplifiers

An *operational amplifier*, or *op amp*, is one of the most useful linear devices. While it is possible to build an op amp with discrete components, and early versions were, the symmetry of the circuit demanded for high performance requires a close match of many components. It is more effective, and much easier, to implement as an integrated circuit. (The term "operational" comes from the op amp's origin in analog computers where it was used to implement mathematical operations.)

The op amp's performance approaches that of an ideal analog circuit building block: 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. These attributes allow the circuit designer to implement many different functions with an op amp and only a few external components.

3.6.1 Characteristics of Practical Op-Amps

An op amp has three signal terminals (see **Fig 3.60**). There are two input terminals, the *noninverting input* marked with a + sign and the *inverting input* marked with a – sign. Voltages applied to the noninverting input cause the op amp output voltage to change with the same polarity.

The output of the amplifier is a single terminal with the output voltage referenced to the external circuit's reference voltage. Usually, that reference is ground, but the op amp's internal circuitry allows all voltages to *float*, that is, to be referenced to any arbitrary voltage between the op amp's power supply voltages. The reference can be negative, ground or positive. For example, an op amp powered from a single power supply voltage amplifies just as well if the circuit reference voltage is halfway between ground and the supply voltage.

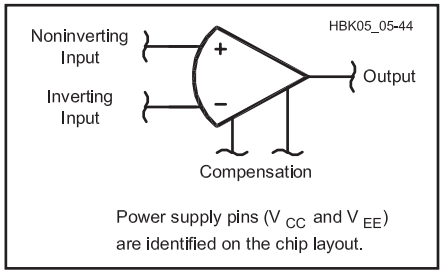

Fig 3.60 — Operational amplifier schematic symbol. The terminal marked with a + sign is the noninverting input. The terminal marked with a – sign is the inverting input. The output is to the right. On some op amps, external compensation is needed and leads are provided, pictured here below the device. Usually, the power supply leads are not shown on the op amp itself but are specified in the data sheet.

GAIN-BANDWIDTH PRODUCT AND COMPENSATION

An ideal op amp would have infinite frequency response, but just as transistors have an f_T that marks their upper frequency limit, the op amp has a *gain-bandwidth product* (GBW or BW). GBW represents the maximum product of gain and frequency available to any signal or circuit: voltage gain × frequency = GBW. If an op-amp with a GBW of 10 MHz is connected as a ×50 voltage amplifier, the maximum frequency at which that gain could be guaranteed is GBW / gain = 10 MHz / 50 = 200 kHz. GBW is an important consideration in high-performance filters and signal processing circuits whose design equations require high-gain at the frequencies over which they operate.

Older operational amplifiers, such as the LM301, have an additional two connections for *compensation*. To keep the amplifier from oscillating at very high gains it is often necessary to place a capacitor across the compensation terminals. This also decreases the frequency response of the op amp but increases its stability by making sure that the output signal can not have the right phase to create positive feedback at its inputs. Most modern op amps are internally compensated and do not have separate pins to add compensation capacitance. Additional compensation can be created by connecting a capacitor between the op amp output and the inverting input.

CMRR AND PSRR

One of the major advantages of using an op amp is its very high *common mode rejection ratio* (CMRR). *Common mode* signals are those that appear equally at all terminals. For example, if both conductors of an audio cable pick up a few tenths of a volt of 60 Hz signal from a nearby power transformer, that 60 Hz signal is a common-mode signal to whatever device the cable is connected. Since the op amp only responds to *differences* in voltage at its inputs, it can ignore or reject common mode signals. CMRR is a measure of how well the op amp rejects the common mode signal. High CMRR results from the symmetry between the circuit halves. CMRR is important when designing circuits that process low-level signals, such as microphone audio or the mV-level dc signals from sensors or thermocouples.

The rejection of power-supply imbalance is also an important op amp parameter. Shifts in power supply voltage and noise or ripple on the power supply voltages are coupled directly to the op amp's internal circuitry. The op amp's ability to ignore those disturbances is expressed by the *power supply rejection ratio* (PSRR). A high PSRR means that the op amp circuit will continue to perform well even if the power supply is imbalanced or noisy.

INPUT AND OUTPUT VOLTAGE LIMITS

The op amp is capable of accepting and amplifying signals at levels limited by the power supply voltages, also called *rails*. The difference in voltages between the two rails limits the range of signal voltages that can be processed. The voltages can be symmetrical positive and negative voltages (±12 V), a positive voltage and ground, ground and a negative voltage or any two different, stable voltages.

In most op amps the signal levels that can be handled are one or two diode forward voltage drops (0.7 V to 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 was illustrated in Fig 3.1.

"Rail-to-rail" op amps have been developed to handle signal levels within a few tens of mV of rails (for example, the MAX406, from Maxim Integrated Products processes signals to within 10 mV of the power supply voltages). Rail-to-rail op-amps are often used in battery-powered products to allow the circuits to operate from low battery voltages for as long as possible.

INPUT BIAS AND OFFSET

The inputs of an op amp, while very high impedance, still allow some input current to flow. This is the *input bias current* and it is in the range of nA in modern op amps. Slight asymmetries in the op amp's internal circuitry result in a slight offset in the op amp's out-put voltage, even with the input terminals shorted together. The amount of voltage difference between the op amp's inputs required to cause the output voltage to be exactly zero is the *input offset voltage*, generally a few mV or less. Some op amps, such as the LM741, have special terminals to which a potentiometer can be connected to *null* the offset by correcting the internal imbalance. 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.

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 voltage must not be too close to the clamping limits or distortion will occur.

A TYPICAL OP AMP

As an example of typical values for these parameters, one of today's garden-variety op amps, the TL084, which contains both JFET and bipolar transistors, has a guaranteed minimum CMRR of 80 dB, an input bias current guaranteed to be below 200 pA (1 pA = 1 millionth of a μA) and a gain-bandwidth product of 3 MHz. Its input offset voltage is 3 mV. CMRR and PSRR are 86 dB, meaning that an unwanted signal or power supply imbalance of 1 V will only result in a 2.5 nV change at the op amp's output! All this for 33 cents even purchased in single quantities and there are four op-amps per package — that's a lot of performance.

3.6.2 Basic Op Amp Circuits

If a signal is connected to the input terminals of an op amp without any other circuitry attached, it will be amplified at the device's *open-loop gain* (typically 200,000 for the TL084 at dc and low frequencies, or 106 dB). This will quickly saturate the output at the power supply rails. Such large gains are rarely used. In most applications, negative feedback is used to limit the circuit gain by providing a feedback path from the output terminal to the inverting input terminal. The resulting *closed-loop gain* of the circuit depends solely on the values of the passive components used to form the loop (usually resistors and, for frequency-selective circuits, capacitors). The higher the op-amp's open-loop gain, the closer the circuit's actual gain will approach that predicted from the component values. Note that the gain of the op amp itself has not changed — it is the configuration of the external components that determines the overall gain of the circuit. Some examples of different circuit configurations that manipulate the closed-loop gain follow.

INVERTING AND NONINVERTING AMPLIFIERS

The op amp is often used in either an *inverting* or a *noninverting* amplifier circuit as shown in **Fig 3.61**. (Inversion means that the output signal is inverted from the input signal about the circuit's voltage reference as described below.) The amount of amplification is determined by the two resistors: the feedback resistor, R_f, and the input resistor, R_i.

In the noninverting configuration shown in Fig 3.61A, the input signal is connected to the op-amp's noninverting input. The feedback resistor is connected between the output and the inverting input terminal. The inverting input terminal is connected to R_i, which is connected to ground (or the circuit reference voltage).

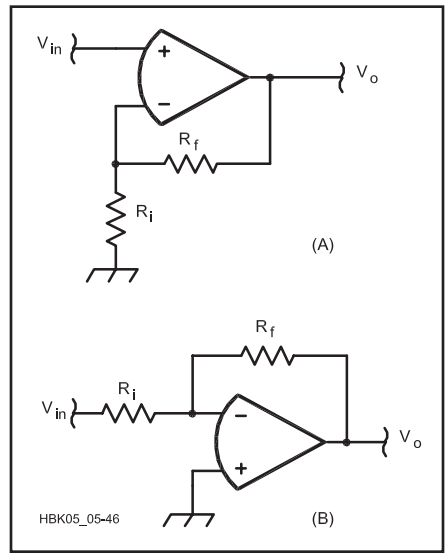

Fig 3.61 — Operational amplifier circuits. (A) Noninverting configuration. (B) Inverting configuration.

This circuit illustrates how op amp circuits use negative feedback, the high open-loop gain of the op amp itself, and the high input impedance of the op amp inputs to create a stable circuit with a fixed gain. The signal applied to the noninverting input causes the output voltage of the op-amp to change with the same polarity. That is, a positive input signal causes a positive change in the op amp's output voltage. This voltage causes current to flow in the voltage divider formed by R_f and R_i. Because the current into the inverting input is so low, the current through R_f is the same as R_i.

The voltage at the *summing junction*, the connection point for the two resistors and the inverting terminal, V_{INV}, is:

$$V_{INV} = V_O\left(\frac{R_i}{R_i + R_f}\right) \qquad (62)$$

The op amp's output voltage will continue to rise until the *loop error signal*, the difference in voltage between the inverting and noninverting inputs, is close to zero. At this point, the voltage at the inverting terminal is approximately equal to the voltage at the noninverting terminal, V_{in}, so that $V_{INV} = V_{in}$. Substituting in equation 62, the gain of this circuit is:

$$\frac{V_O}{V_{in}} = \left(1 + \frac{R_f}{R_i}\right) \qquad (63)$$

where

V_o = the output voltage

V_{in} = the input voltage.

The higher the op amp's open-loop gain, the closer will be the voltages at the inverting and noninverting terminals when the circuit is balanced and the more closely the circuit's closed-loop gain will equal that of Equation 63. So the negative feedback creates an electronic balancing act with the op amp increasing its output voltage so that the input error signal is as small as possible.

In the inverting configuration of Fig 3.61B, the input signal (V_{in}) 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 is connected to ground (or the circuit reference voltage). In this configuration the feedback action results in the output voltage changing to whatever value is needed such that the current through R_i is balanced by an equal and opposite current through R_f. The gain of this circuit is:

$$\frac{V_O}{V_{in}} = -\frac{R_f}{R_{in}} \qquad (64)$$

where V_{in} represents the voltage input to R_{in}.

For the remainder of this section, "ground" or "zero voltage" should be understood to be the circuit reference voltage. That voltage may not be "earth ground potential." For example, if a single positive supply of 12 V is used, 6 V may be used as the circuit reference voltage. The circuit reference voltage is a fixed dc voltage that can be considered to be an ac ground because of the reference source's extremely low ac impedance.

The negative sign in equation 64 indicates that the signal is inverted. For ac signals, inversion represents a 180° phase shift. The gain of the noninverting configuration can vary from a minimum of 1 to the maximum of which the op amp is capable, as indicated by A_v for dc signals, or the gain-bandwidth product for ac signals. The gain of the inverting configuration can vary from a minimum of 0 (gains from 0 to 1 attenuate the signal while gains of 1 and higher amplify the signal) to the maximum of which the device is capable.

The inverting amplifier configuration results in a special condition at the op amp's inverting input called *virtual ground*. Because the op amp's high open-loop gain drives the two inputs to be very close together, if the noninverting input is at ground potential, the inverting input will be very close to ground as well and the op amp's output will change with the input signal to maintain the inverting input at ground. Measured with a voltmeter, the input appears to be grounded, but it is merely maintained at ground potential by the action of the op amp and the feedback loop. This point in the circuit may not be connected

to any other ground connection or circuit point because the resulting additional current flow will upset the balance of the circuit.

The *voltage follower* or *unity-gain buffer* circuit of **Fig 3.62** 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. This configuration has unity gain because the circuit is balanced when the output and input voltages are the same (error voltage equals zero). It also provides the maximum possible input impedance and the minimum possible output impedance of which the device is capable.

Differential and Difference Amplifier

A *differential amplifier* is a special application of an operational amplifier (see **Fig 3.63**). It amplifies the difference between two analog signals and is very useful to cancel noise under certain 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

$$V_O = \frac{R_f}{R_i}\left[\frac{1}{\frac{R_n}{R_g}+1}\left(\frac{R_i}{R_f}+1\right) V_n - V_i\right] \quad (65)$$

which, if the ratios R_i/R_f and R_n/R_g are equal, simplifies to:

$$V_O = \frac{R_f}{R_i}(V_n - V_I) \quad (66)$$

Note that the differential amplifier response is identical to the inverting amplifier response (equation 64) if the voltage applied to the noninverting terminal is equal to zero. If the voltage applied to the inverting terminal (V_i) is zero, the analysis is a little more complicated but it is possible to derive the noninverting amplifier response (equation 62)

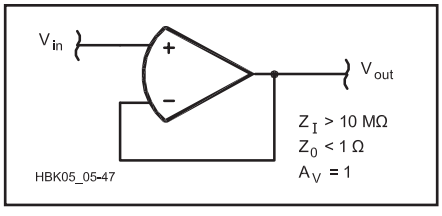

Fig 3.62 — Voltage follower. This operational amplifier circuit makes a nearly ideal buffer with a voltage gain of about one, and with extremely high input impedance and extremely low output impedance.

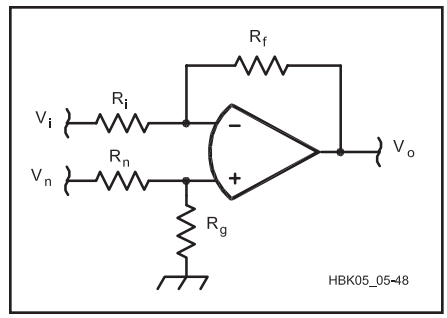

Fig 3.63 — Difference amplifier. This operational amplifier circuit amplifies the difference between the two input signals.

from the differential amplifier response by taking into account the influence of R_n and R_g. If all four resistors have the *same* value the *difference amplifier* is created and V_O is just the difference of the two voltages.

$$V_O = V_n - V_I \quad (67)$$

Instrumentation Amplifier

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 3.64** is called an *instrumentation amplifier*. It has three parts; each of the two inputs is connected to a noninverting buffer amplifier with a gain of 1 + R2/R1. The outputs of these buffer amplifiers are then connected to a differential amplifier with a gain of R4/R3. V2 is the circuit's inverting input and V1 the noninverting input.

The three amplifier modules are usually all part of the same integrated circuit. This means that they have essentially the same temperature and the internal transistors and resistors are very well matched. This causes the subtle gain and tracking errors caused by temperature differences and mismatched components between individual op amps to be cancelled out or dramatically reduced. In addition, the external resistors using the same designators (R2, R3, R4) are carefully matched as well, sometimes being part of a single integrated *resistor pack*. The result is a circuit with better performance than any single-amplifier circuit over a wider temperature range.

Summing Amplifier

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 3.65**. For the inverting summing amplifier, the gain of each input signal can be calculated individually using equation 64 and, because of the superposition property of linear circuits, the output is 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:

$$V_O = V_{n1}\frac{R_{p1}}{R_1 + R_{p1}} + V_{n2}\frac{R_{p2}}{R_2 + R_{p2}} + \ldots$$

$$+V_{nm}\frac{R_{pm}}{R_m + R_{pm}} \quad (68)$$

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.

Comparators

A *voltage comparator* is another special form of op amp circuit, shown in **Fig 3.66**. It has two analog signals as its inputs and its output is either TRUE or FALSE depending on whether the noninverting or inverting signal voltage is higher, respectively. Thus, it "compares" the input voltages. TRUE generally corresponds to a positive output voltage and

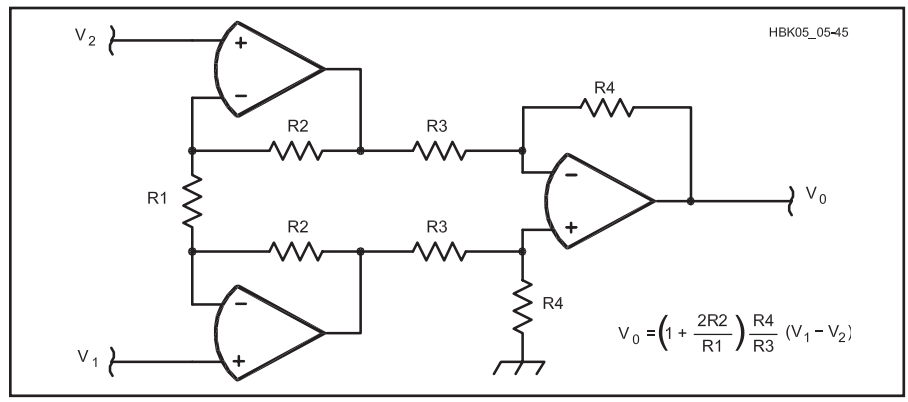

Fig 3.64 — Operational amplifiers arranged as an instrumentation amplifier. The balanced and cascaded series of op amps work together to perform differential amplification with good common-mode rejection and very high input impedance (no load resistor required) on both the inverting (V_1) and noninverting (V_2) inputs.

FALSE to a negative or zero voltage. The circuit in Fig 3.66 uses external resistors to generate a reference voltage, called the *setpoint*, to which the input signal is compared. A comparator can also compare two variable voltages.

A standard operational amplifier can be made to act as a comparator by connecting the two input 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 driven to the positive clamping limit. If the inverting input is at a higher potential than the noninverting input, the output voltage will be driven to the negative clamping limit. If the comparator is comparing an unknown voltage to a known voltage, the known voltage is called the *setpoint* and the comparator output indicates whether the unknown voltage is above or below the setpoint.

An op amp that has been intended for use as a comparator, such as the LM311, is optimized to respond quickly to the input signals. In addition, comparators often have *open-collector outputs* that use an external *pull-up* resistor, R_{OUT}, connected to a positive power supply voltage. When the comparator output is TRUE, the output transistor is turned off and the pull-up resistor "pulls up" the output voltage to the positive power supply voltage. When the comparator output is FALSE, the transistor is driven into the saturation and the output voltage is the transistor's $V_{CE(sat)}$.

Hysteresis

Comparator circuits also use *hysteresis* to prevent "chatter" — the output of the comparator switching rapidly back and forth when the input voltage is at or close to the setpoint voltage. There may be noise on the input signal, as shown in **Fig 3.67A**, that causes the input voltage to cross the setpoint threshold repeatedly. The rapid switching of the output can be confusing to the circuits monitoring the comparator output.

Hysteresis is a form of positive feedback that "moves" the setpoint by a few mV in the direction *opposite* to that in which the input signal crossed the setpoint threshold. As shown in Fig 3.67B, the slight shift in the setpoint tends to hold the comparator output in the new state and prevents switching back to the old state. **Fig 3.68** shows how the output of the comparator is fed back to the positive input through resistor R3, adding or subtracting a small amount of current from the divider and shifting the setpoint.

Some applications of a voltage comparator

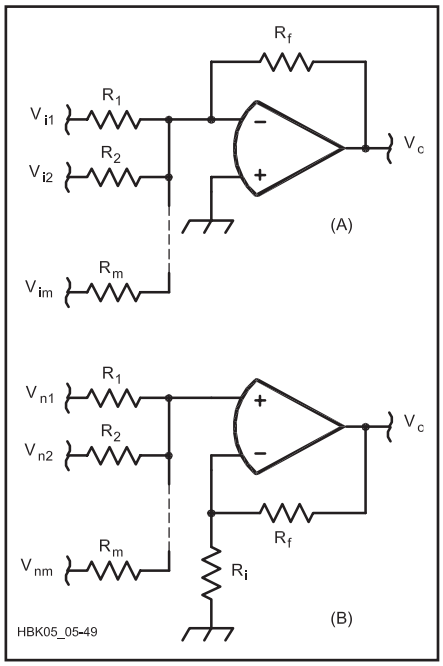

Fig 3.65 — Summing operational amplifier circuits. (A) Inverting configuration. (B) Noninverting configuration.

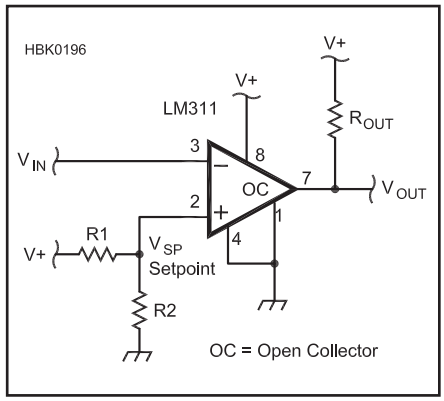

Fig 3.66 — A comparator circuit in which the output voltage is low when voltage at the inverting input is higher than the setpoint voltage at the noninverting input.

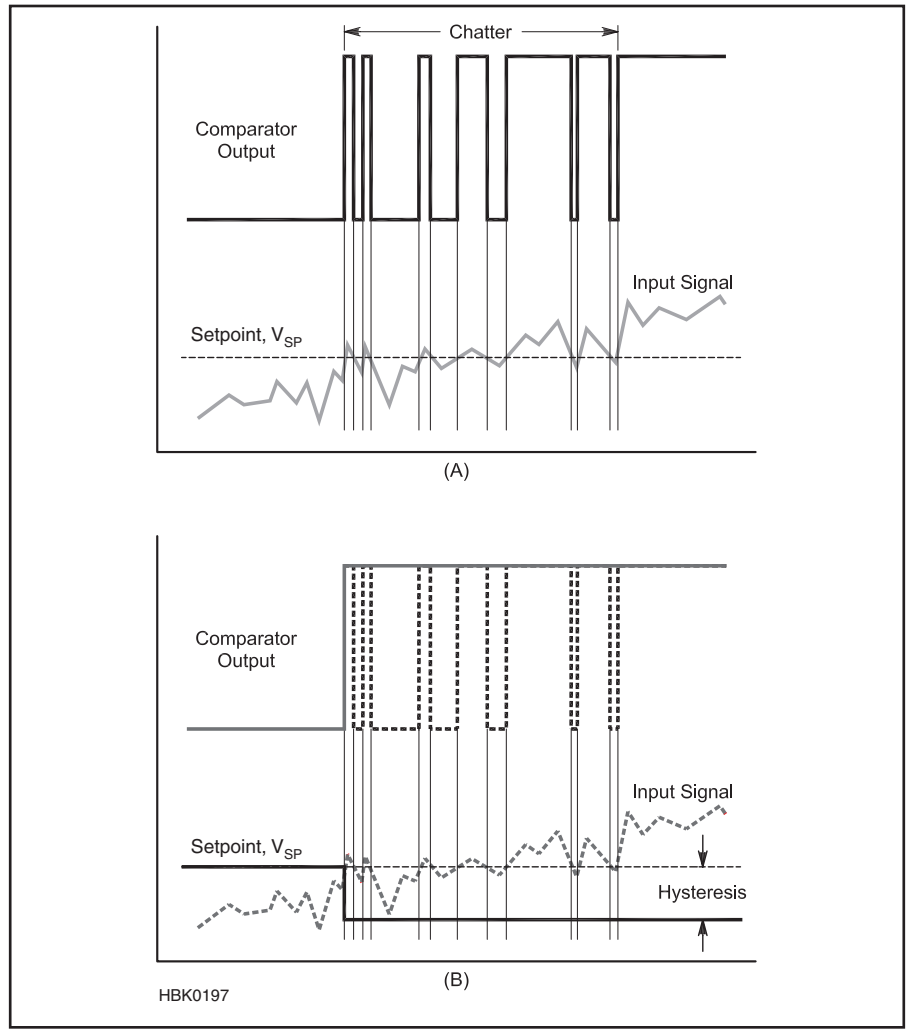

Fig 3.67 — Chatter (A) is caused by noise when the input signal is close to the setpoint. Chatter can also be caused by voltage shifts that occur when a heavy load is turned on and off. Hysteresis (B) shifts the setpoint a small amount by using positive feedback in which the output pulls the setpoint farther away from the input signal after switching.

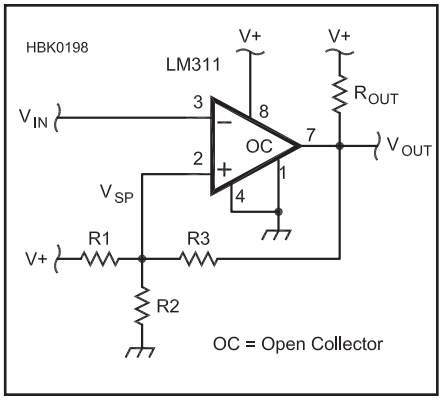

Fig 3.68 — Comparator circuit with hysteresis. R3 causes a shift in the comparator setpoint by allowing more current to flow through R1 when the comparator output is low.

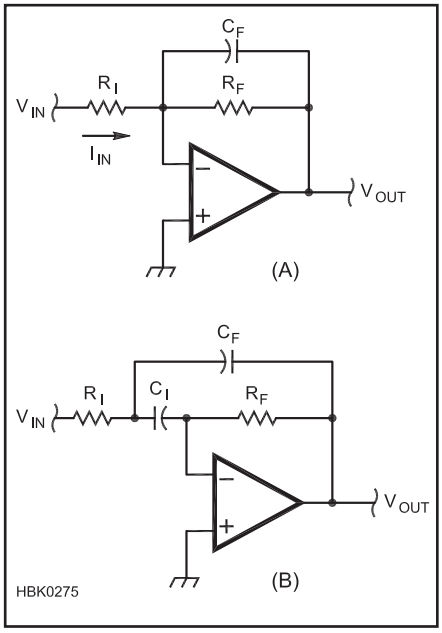

Fig 3.69 — Op amp active filters. The circuit at (A) has a low-pass response identical to an RC filter. The –3 dB frequency occurs when the reactance of C_F equals R_F. The band-pass filter at (B) is a multiple-feedback filter.

are a zero crossing detector, a signal squarer (which turns other cyclical wave forms into square waves) and a peak detector. An amateur station application: Circuits that monitor the CI-V band data output voltage from ICOM HF radios use a series of comparators to sense the level of the voltage and indicate on which band the radio is operating.

FILTERS

One of the most important type of op amp circuits is the *active filter*. Two examples of op amp filter circuits are shown in **Fig 3.69.** The simple noninverting low-pass filter in Fig 3.69A has the same response as a passive single-pole RC low-pass filter, but unlike the passive filter, the op amp filter circuit has a very high input impedance and a very low output impedance so that the filter's frequency and voltage response are relatively unaffected by the circuits connected to the filter input and output. This circuit is a low-pass filter because the reactance of the feedback capacitor decreases with frequency, requiring less output voltage to balance the voltages of the inverting and noninverting inputs.

The *multiple-feedback* circuit in Fig 3.69B results in a band-pass response while using only resistors and capacitors. This circuit is just one of many different types of active filters. Active filters are discussed in the **RF and AF Filters** chapter.

RECTIFIERS AND PEAK DETECTORS

The high open-loop gain of the op amp can also be used to simulate the I-V characteristics of an ideal diode. A *precision rectifier* circuit is shown in **Fig 3.70** along with the I-V characteristics of a real (dashed lines) and ideal (solid line) diode. The high gain of the op amp compensates for the V_F forward voltage drop of the real diode in its feedback loop with an output voltage equal to the input voltage plus V_F. Remember that the op amp's output increases until its input voltages are balanced. When the input voltage is negative, which would reverse-bias the diode, the op amp's output can't balance the input because the diode blocks any current flow through the feedback loop. The resistor at the output holds the voltage at zero until the input voltage is positive once again. Precision half-wave and full-wave rectifier circuits are shown in **Fig 3.71** and their operation is described in many reference texts.

One application of the precision rectifier circuit useful in radio is the *peak detector*, shown in **Fig 3.72**. A precision rectifier is used to charge the output capacitor which

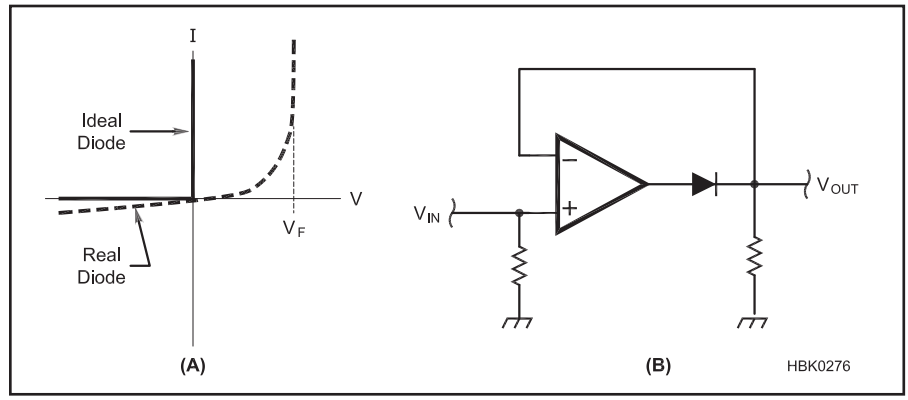

Fig 3.70 — Ideal and real diode I-V characteristics are shown at (A). The op amp precision rectifier circuit is shown at (B).

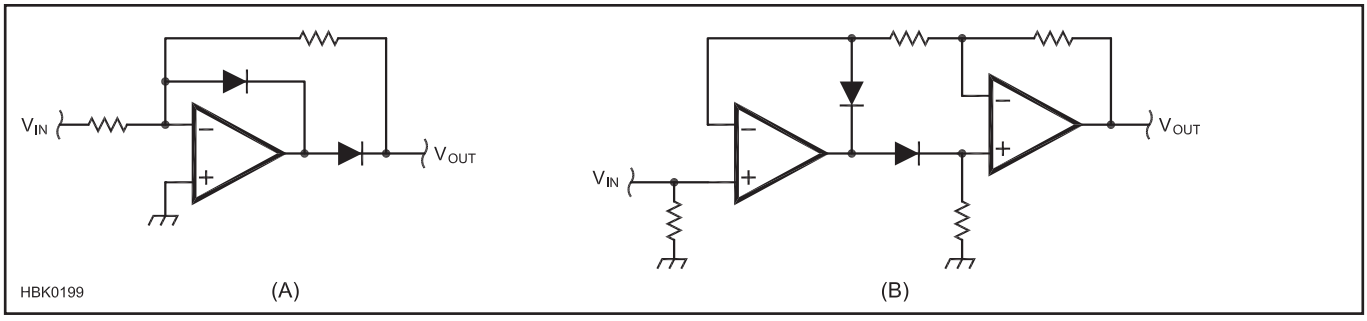

Fig 3.71 — Half-wave precision rectifier (A). The extra diode at the output of the op amp prevents the op amp from saturating on negative half-cycles and improves response time. The precision full-wave rectifier circuit at (B) reproduces both halves of the input waveform.

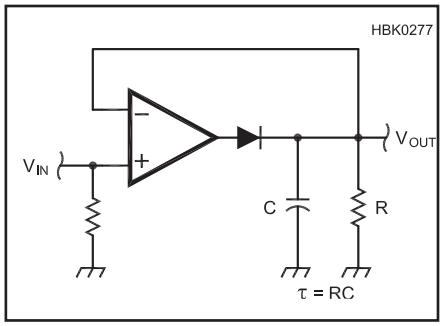

Fig 3.72 — Peak detector. Coupling a precision diode with a capacitor to store charge creates a peak detector. The capacitor will charge to the peak value of the input voltage. R discharges the capacitor with a time constant of $\tau = RC$ and can be omitted if it is desired for the output voltage to remain nearly constant.

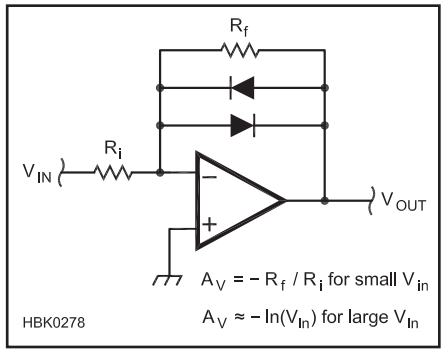

Fig 3.73 — Log amplifier. At low voltages, the gain of the circuit is $-R_f/R_i$, but as the diodes begin to conduct for higher-voltage signals, the gain changes to $-\ln(V_{in})$ in because of the diode's exponential current as described in the Fundamental Diode Equation.

holds the peak voltage. The output resistor sets the time constant at which the capacitor discharges. The resistor can also be replaced by a transistor acting as a switch to reset the detector. This circuit is used in AGC loops, spectrum analyzers, and other instruments that measure the peak value of ac waveforms.

LOG AMPLIFIER

There are a number of applications in radio in which it is useful for the gain of an amplifier to be higher for small input signals than for large input signals. For example, an audio compressor circuit is used to reduce the variations in a speech signal's amplitude so that the average power output of an AM or SSB transmitter is increased. A *log amplifier* circuit whose gain for large signals is proportional to the logarithm of the input signal's amplitude is shown in **Fig 3.73**. The log amp circuit is used in compressors and limiter circuits.

At signal levels that are too small to cause significant current flow through the diodes, the gain is set as in a regular inverting amplifier, $A_V = -R_f / R_i$. As the signal level increases, however, more current flows through the diode according to the Fundamental Diode Equation (equation 4) given earlier in this chapter. That means the op amp output voltage has to increase less (lower gain) to cause enough current to flow through R_i such that the input voltages balance. The larger the input voltage, the more the diode conducts and the lower the gain of the circuit. Since the diode's current is exponential in response to voltage, the gain of the circuit for large input signals is logarithmic.

Voltage-Current Converters

Another pair of useful op amp circuits convert voltage into current and current into voltage. These are frequently used to convert currents from sensors and detectors into voltages that are easier to measure. **Fig 3.74A** shows a voltage-to-current converter in which the output current is actually the current in the feedback loop. Because the op amp's high open-loop gain insures that its input voltages are equal, the current $I_{R1} = V_{IN} / R1$. Certainly, this could also be achieved with a resistor and Ohm's Law, but the op amp circuit's high input impedance means there is little interaction between the input voltage source and the output current.

Going the other way, Fig 3.74B is a current-to-voltage converter. The op amp's output will change so that the current through the feedback resistor, R1, exactly equals the input current, keeping the inverting terminal at ground potential. The output voltage, $V_O = I_{IN}$ R1. Again, this could be done with just a resistor, but the op amp provides isolation between the source of input current and the output voltage. Fig 3.74C shows an application of a current-to-voltage converter in which the small currents from a photodiode are turned into voltage. This circuit can be used as a detector for amplitude modulated light pulses or waveforms.

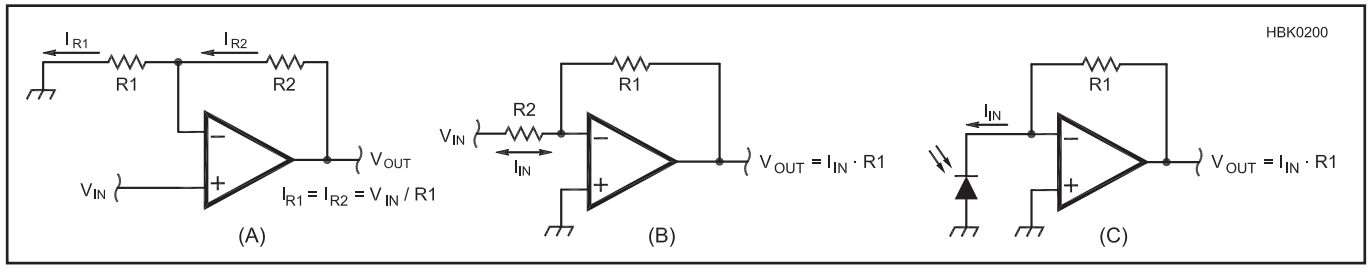

Fig 3.74 — Voltage-current converters. The current through R1 in (A) equals $V_{in}/R1$ because the op amp keeps both input terminals at approximately the same voltage. At (B), input current is balanced by the op amp, resulting in $V_{OUT} = I_{IN}R1$. Current through a photodiode (C) can be converted into a voltage in this way.

3.7 Analog-Digital Conversion

While radio signals are definitely analog entities, much of the electronics associated with radio is digital, operating on binary data representing the radio signals and the information they carry. The interface between the two worlds — analog and digital — is a key element of radio communications systems and the function is performed by analog-digital converters.

This section presents an overview of the different types of analog-digital converters and their key specifications and behaviors. The chapter on **Digital Basics** presents information on interfacing converters to digital circuitry and associated issues. The chapter on **DSP and Software Radio Design** discusses specific applications of analog-digital conversion technology in radio communications systems.

3.7.1 Basic Conversion Properties

Analog-digital conversion consists of taking data in one form, such as digital binary data or an analog ac RF waveform, and creating an equivalent representation of it in the opposite domain. Converters that create a digital representation of analog voltages or currents are called *analog-to-digital converters* (ADC), *analog/digital converters, A/D converters* or *A-to-D converters.* Similarly, converters that create analog voltages or currents from digital quantities are called *digital-to-analog converters* (DAC), *digital/analog converters, D/A converters* or *D-to-A converters.* The word "conversion" in this first section on the properties of converting information between the analog and digital domains will apply equally to analog-to-digital or digital-to-analog conversion.

Converters are typically implemented as integrated circuits that include all of the necessary interfaces and sub-systems to perform the entire conversion process. Schematic symbols for ADCs and DACs are shown in **Fig 3.75**.

Fig 3.76 shows two different representations of the same physical phenomenon; an analog voltage changing from 0 to 1 V. In the analog world, the voltage is continuous and can be represented by any real number between 0 and 1. In the digital world, the number of possible values that can represent any phenomenon is limited by the number of bits contained in each value.

In Fig 3.76, there are only four two-bit digital values 00, 01, 10, and 11, each corresponding to the analog voltage being within a specific range of voltages. If the analog voltage is anywhere in the range 0 to 0.25 V, the digital value representing the analog voltage will be 00, no matter whether the voltage is 0.0001 or 0.24999 V. The range 0.25 to 0.5 V is represented by the digital value 01, and so forth.

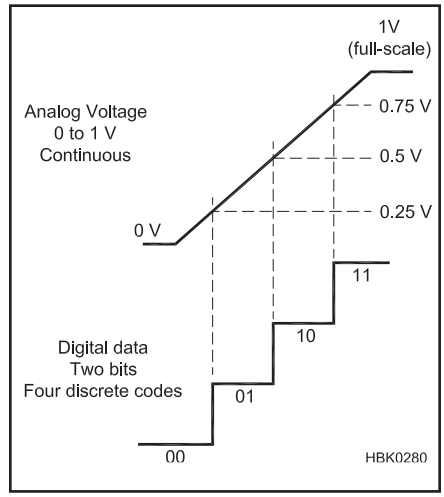

Fig 3.76 — The analog voltage varies continuously between 0 and 1 V, but the two-bit digital system only has four values to represent the analog voltage, so representation of the analog voltage is coarse.

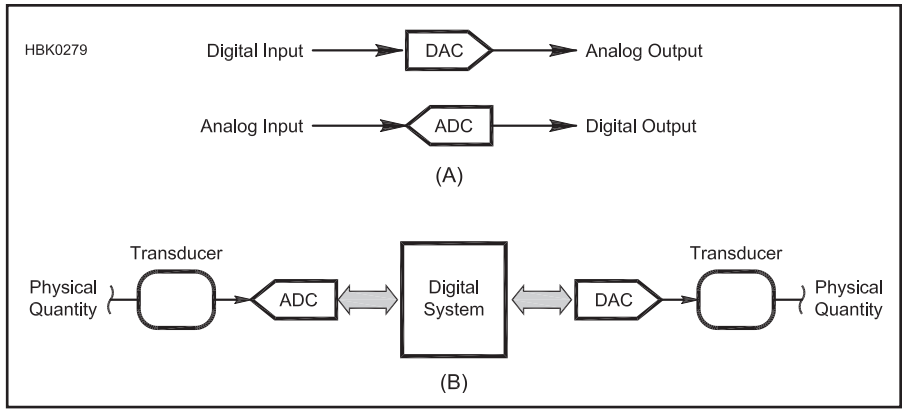

Fig 3.75 — Schematic symbols (A) for digital-to-analog converters (DAC) and analog-to-digital converters (ADC). The general block diagram of a system (B) that digitizes an analog signal, operates on it as digital data, then converts it back to analog form.

The process of converting a continuous range of possible values to a limited number of discrete values is called *digitization* and each discrete value is called a *code* or a *quantization code*. The number of possible codes that can represent an analog quantity is 2^N, where N is the number of bits in the code. A two-bit number can have four codes as shown in Fig 3.76, a four-bit number can have sixteen codes, an eight-bit number 256 codes, and so forth. Assuming that the smallest change in code values is one bit, that value is called the *least significant bit* (LSB) regardless of its position in the format used to represent digital numbers.

RESOLUTION AND RANGE

The *resolution* or *step size* of the conversion is the smallest change in the analog value that the conversion can represent. The *range* of the conversion is the total span of analog values that the conversion can process. The maximum value in the range is called the *full-scale* (F.S.) value. In Fig 3.76, the conversion range is 1 V. The resolution of the conversion is

$$\text{resolution} = \frac{\text{range}}{2^N}$$

In Fig 3.76, the conversion resolution is $\frac{1}{4} \times 1\text{ V} = 0.25\text{ V}$ in the figure. If each code had four bits instead, it would have a resolution of $\frac{1}{2}^4 \times 1\text{ V} = 0.083\text{ V}$. Conversion range does not necessarily have zero as one end point. For example, a conversion range of 5 V may span 0 to 5 V, –5 to 0 V, 10 to 15 V, –2.5 to 2.5 V, and so on.

Analog-digital conversion can have a range that is *unipolar* or *bipolar*. Unipolar means a conversion range that is entirely positive or negative, usually referring to voltage. Bipolar means the range can take on both positive and negative values.

Confusingly, the format in which the bits are organized is also called a code. The binary code represents digital values as binary numbers with the least significant bit on the right. *Binary-coded-decimal* (BCD) is a code in which groups of four bits represent individual decimal values of 0-9. In the *hexadecimal* code, groups of four bits represent decimal values of 0-15. There are many such codes. Be careful in interpreting the word "code" to be sure the correct meaning is used or understood.

To avoid having to know the conversion range to specify resolution, *percentage resolution* is used instead.

$$\%\,\text{resolution} = \frac{\text{resolution}}{\text{full scale}} \times 100\% \qquad (69)$$

In Fig 3.76, the conversion's percentage resolution is

$$\%\,\text{resolution} = \frac{0.25\ \text{V}}{1.0\ \text{V}} \times 100\% = 25\%$$

Because each code represents a range of possible analog values, the limited number of available codes creates *quantization error*. This is the maximum variation in analog values that can be represented by the same code. In Fig 3.76, any value from 0.25 through 0.50 V could be represented by the same code: 01. The quantization error in this case is 0.25 V. Quantization error can also be specified as % full-scale by substituting the value of error for resolution in equation 69.

Resolution can also be defined by the number of bits in the conversion. The higher the number of bits, the smaller the resolution as demonstrated above. Since many converters have variable ranges set by external components or voltages, referring to percent resolution or as a number of bits is preferred. The conversions between percent resolution and number of bits are as follows:

$$\%\,\text{resolution} = \frac{1}{2^N} \times 100\% \qquad (70)$$

and

$$N = \frac{\log\left(\frac{100\%}{\%\,\text{resolution}}\right)}{\log 2} \qquad (71)$$

Quantization error can also be specified as a number of least significant bits (LSB) where each bit is equivalent to the conversion's resolution.

When applied to receivers, dynamic range is more useful than percent resolution. For each additional bit of resolution, the resolution becomes two times greater, or 6.02 dB.

$$\text{Dynamic range (dB)} = N \times 6.02\ \text{dB} \qquad (72)$$

For example, a 16-bit conversion has a dynamic range of 16 × 6.02 = 96.32 dB. This is the conversion's theoretical dynamic range with no noise present in the system. If noise is present, some number of the smallest codes will contain only noise, reducing the dynamic range available to represent a signal. The noise inherent in a particular conversion process can be specified as an analog value (mV, μA and so on) or as a number of bits, meaning the number of codes that only represent noise. For example, if a conversion has five bits of noise, any value represented by the smallest five codes is considered to be noise. The conversion's *effective number of bits* (ENOB) describes the number of bits available to contain information about the signal.

ACCURACY

A companion to resolution, *accuracy* refers to the ability of the converter to either assign the correct code to an analog value or create the true analog value from a specific code. As with resolution, it is most convenient to refer to accuracy as either a percentage of full scale or in bits. *Full-scale error* is the maximum deviation of the code's value or the analog quantity's value as a percentage of the full scale value. If a converter's accuracy is given as 0.02% F.S. and the conversion range is 5 V, the conversion can be in error by as much as 0.02% × 5 V = 1 mV from the correct or expected value. *Offset* has the same meaning in conversion as it does in analog electronics — a consistent shift in the value of the conversion from the ideal value.

Linearity error represents the maximum deviation in step size from the ideal step size. This is also called *integral nonlinearity* (INL). In the converter of Fig 3.76, ideal step size is 0.25 V. If the linearity error for the conversion was given as 0.05% F.S., any actual step size could be in error by as much as 0.05% × 5 V = 2.5 mV. The amount of error is based on the full-scale value, not the step-size value. *Differential nonlinearity* is a measure of how much any two adjacent step sizes deviate from the ideal step size. Errors can be represented as a number of bits, usually assumed to be least significant bits, or LSB, with one bit representing the same range as the conversion resolution.

Accuracy and resolution are particularly important in conversions for radio applications because they represent distortion in the signal being processed. An analog receiver that distorts the received signal creates undesired spurious signals that interfere with the desired signal. While the process is not exactly the same, distortion created by the signal conversion circuitry of a digital receiver also causes degradation in performance.

DISTORTION AND NOISE

Distortion and noise in a conversion are characterized by several parameters all related to linearity and accuracy. THD+N (*Total Harmonic Distortion + Noise*) is a measure of how much distortion and noise is introduced by the conversion. THD+N can be specified in percent or in dB. Smaller values are better. SINAD (*Signal to Noise and Distortion Ratio*) is related to THD+N, generally specified along with a desired signal level to show what signal level is required to achieve a certain level of SINAD or the highest signal level at which a certain level of SINAD can be maintained. *Spurious-free Dynamic Range* (SFDR) is the difference between the amplitude of the desired signal and the highest unwanted signal. SFDR is generally specified in dB and a higher number indicates better performance.

CONVERSION RATE AND BANDWIDTH

Another important parameter of the conversion is the *conversion rate* or its reciprocal, *conversion speed*. A code that represents an analog value at a specific time is called a *sample*, so conversion rate, f_S, is specified in *samples per second* (sps) and conversion speed as some period of time per sample, such as 1 μsec/s. Because of the mechanics by which conversion is performed, conversion speed can also be specified as a number of cycles of clock signal used by the digital system performing the conversion. Conversion rate then depends on the frequency of the clock. Conversion speed may be smaller than the reciprocal of conversion rate if the system controlling the conversion introduces a delay between conversions. For example, if a conversion can take place in 1 μs, but the system only performs a conversion once per ms, the conversion rate is 1 ksps, not the reciprocal of 1 μs/s = 1 Msps.

The time accuracy of the digital clock that controls the conversion process can also affect the accuracy of the conversion. An error in long-term frequency will cause all conversions to have the same frequency error. Short-term variations in clock period are called *jitter* and they add noise to the conversion.

According to the *Nyquist Sampling Theorem*, a conversion must occur at a rate twice the highest frequency present in the analog signal. This minimum rate is the *Nyquist rate* and the maximum frequency allowed in the analog signal is the *Nyquist frequency*. In this way, the converter *bandwidth* is limited to one-half the conversion rate.

Referring to the process of converting analog signals to digital samples, if a lower rate is used, called *undersampling*, false signals called *aliases* will be created in the digital representation of the input signal at frequencies related to the difference between the Nyquist sampling rate and f_S. This is called *aliasing*. Sampling faster than the Nyquist rate is called *oversampling*.

Because conversions occur at some maximum rate, there is always the possibility of signals greater than the Nyquist frequency being present in an analog signal undergoing conversion or that is being created from digital values. These signals would result in aliases and must be removed by *band-limiting filters* that remove them prior to conversion. The mechanics of the sampling process are discussed further in the chapter **DSP and Software Radio Design**.

3.7.2 Analog-to-Digital Converters (ADC)

There are a number of methods by which the conversion from an analog quantity to a set of digital samples can be performed. Each has its strong points — simplicity, speed,

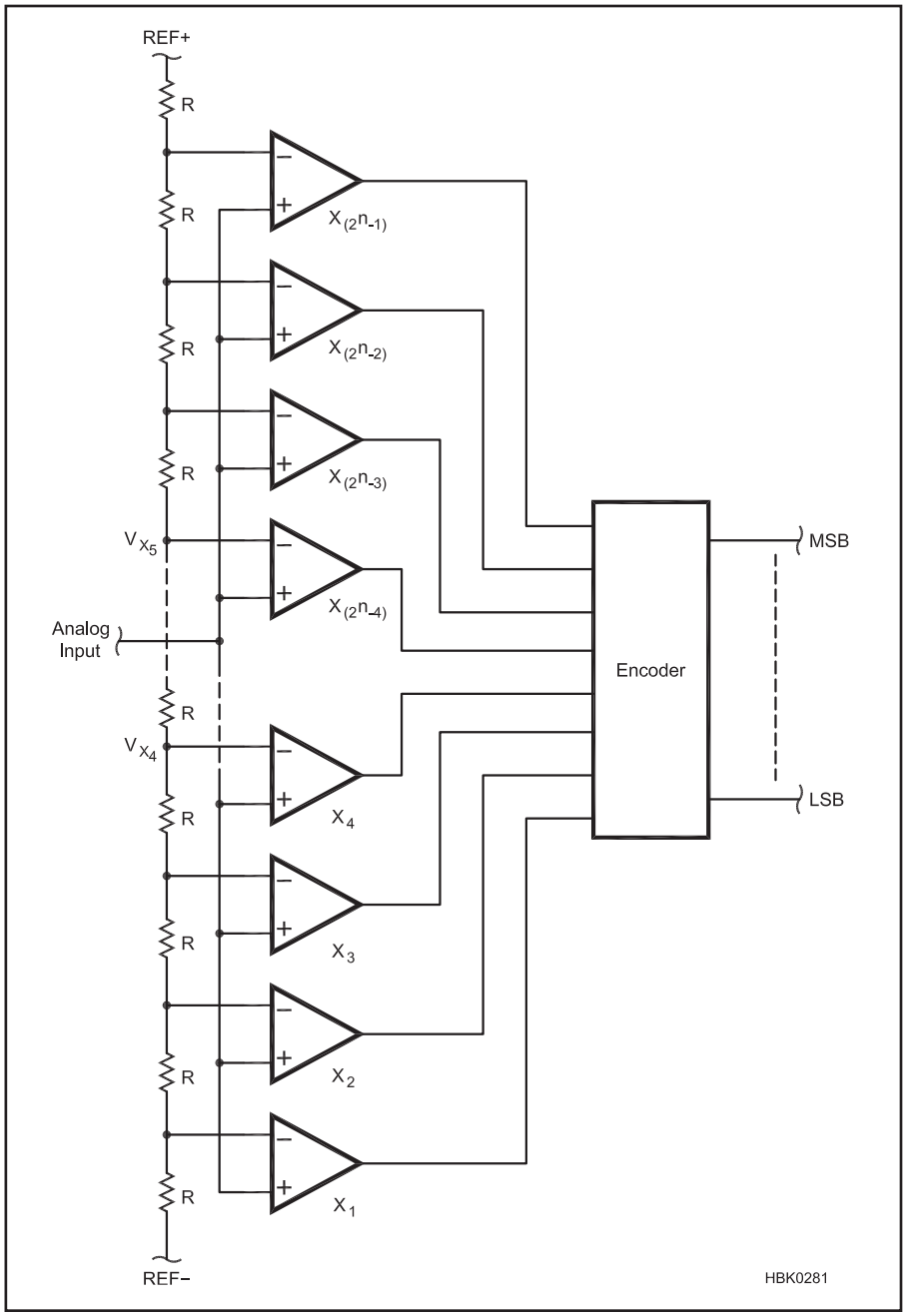

Fig 3.77 — The comparators of the flash converter are always switching state depending on the input signal's voltage. The decoder section converts the array of converter output to a single digital word.

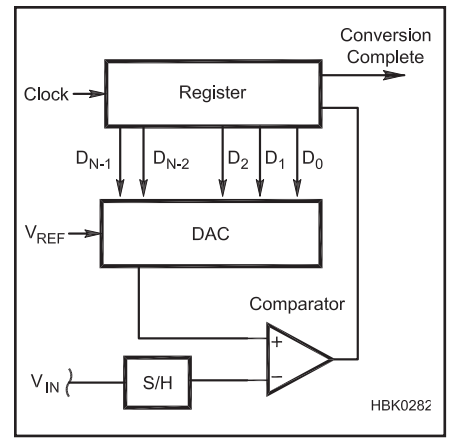

Fig 3.78 — The successive-approximation converter creates a digital word as it varies the DAC signal in order to keep the comparator's noninverting terminal close to the input voltage. A sample-and-hold circuit (S/H) holds the input signal steady while the measurement is being made.

resolution, accuracy — all affect the decision of which method to use for a particular application. In order to pick the right type of ADC, it is important to decide which of these criteria most strongly affect the performance of your application.

FLASH CONVERTER

The simplest type of ADC is the *flash converter*, shown in **Fig 3.77**, also called a *direct-conversion ADC*. It continually generates a digital representation of the analog signal at its input. The flash converter uses an array of comparators that compare the amplitude of the input signal to a set of reference voltages. There is one reference voltage for each step.

The outputs of the comparator array represent a digital value in which each bit indicates whether the input signal is greater (1) or less than (0) the reference voltage for that comparator. A digital logic *priority encoder* then converts the array of bits into a digital output code. Each successive conversion is available as quickly as the comparators can respond and the priority encoder can create the output code.

Flash converters are the fastest of all ADCs (conversion speeds can be in the ns range) but do not have high resolution because of the number of comparators and reference voltages required.

SUCCESSIVE-APPROXIMATION CONVERTER

The *successive-approximation converter* (SAC) is one of the most widely-used types of converters. As shown in **Fig 3.78**, it uses a single comparator and DAC (digital-to-analog converter) to arrive at the value of the input voltage by comparing it to successive analog values generated by the DAC. This type of converter offers a good compromise of conversion speed and resolution.

The DAC control logic begins a conversion by setting the output of the DAC to ½ of the conversion range. If the DAC output is greater than the analog input value, the output of the comparator is 0 and the most significant bit of the digital value is set to 0. The DAC output then either increases or decreases by ¼ of the range, depending on whether the value of the first comparison was 1 or 0. One test is made for each bit in digital output code and the result accumulated in a storage register. The process is then repeated, forming a series of approximations (thus the name of the converter), until a test has been made for all bits in the code.

While the digital circuitry to implement the SAC is somewhat complex, it is less expensive to build and calibrate than the array of comparators and precision resistors of the flash converter. Each conversion also takes

a known and fixed number of clock cycles. The higher the number of bits in the output code, the longer the conversion takes, all other things being equal.

DUAL-SLOPE INTEGRATING CONVERTERS

The *dual-slope integrating ADC* is shown in **Fig 3.79**. It makes a conversion by measuring the time it takes for a capacitor to charge and discharge to a voltage level proportional to the analog quantity.

A constant-current source charges a capacitor (external to the converter IC) until the comparator output indicates that the capacitor voltage is equal to the analog input signal. The capacitor is then discharged by the constant-current source until the capacitor is discharged. This process is repeated continuously. The frequency of the charge-discharge cycle is determined by the values of R and C in Fig 3.79. The frequency is then measured by frequency counter circuitry and converted to the digital output code.

Dual-slope ADCs are low-cost and relatively immune to noise and temperature variations. Due to the slow speed of the conversion (measured in tens of ms) these converters are generally only used in test instruments, such as multimeters.

DELTA-ENCODED CONVERTERS

Instead of charging and discharging a capacitor from 0 V to the level of the input signal, and then back to 0 V, the *delta-encoded ADC* in **Fig 3.80** continually compares the output of a DAC to the input signal using a comparator. Whenever the signal changes, the DAC is adjusted until its output is equal to the input signal. Digital counter circuits keep track of the DAC value and generate the digital output code. Delta-encoded counters are available with wide conversion ranges and high resolution.

SIGMA-DELTA CONVERTERS

The *sigma-delta converter* also uses a DAC and a comparator in a feedback loop to generate a digital signal as shown in **Fig 3.81**. An integrator stores the sum of the input signal and the DAC output. (This is "sigma" or sum in the converter's name.) The comparator output indicates whether the integrator output is above or below the reference voltage and that signal is used to adjust the DAC's output so that the integrator output stays close to the reference voltage. (This is the "delta" in the name.) The stream of 0s and 1s from the comparator forms a high-speed digital bit stream that is digitally-filtered to form the output code.

Sigma-delta converters are used where high resolution (16 to 24 bits) is required at low sampling rates of a few kHz. The digital filtering in the converter also reduces the need for external band-limiting filters on the input signal.

ADC SUBSYSTEMS

Sample-and-Hold

ADCs that use a sequence of operations to create the digital output code must have a means of holding the input signal steady while the measurements are being made. This function is performed by the circuit of **Fig 3.82**. A high input-impedance buffer drives the external storage capacitor, C_{HOLD}, so that its voltage is the same as the input signal. Another high input-impedance buffer

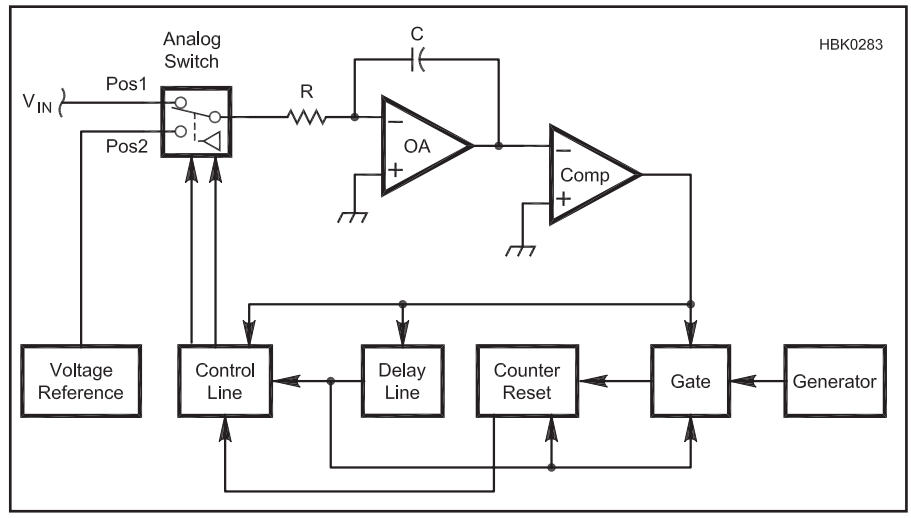

Fig 3.79 — Dual-slope integrating converter. By using a constant-current source to continually charge a capacitor to a known reference voltage then discharge it, the resulting frequency is directly proportional to the resistor value.

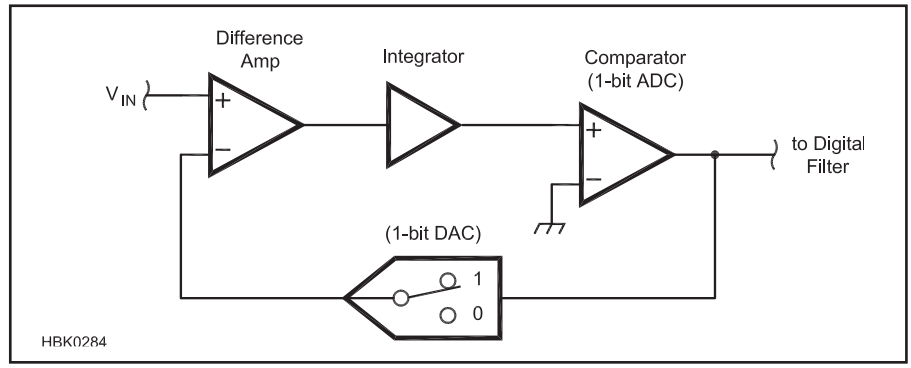

Fig 3.80 — Delta-encoded converter. The 1-bit DAC is operated in such a way that the bit stream out of the comparator represents the value of the input voltage.

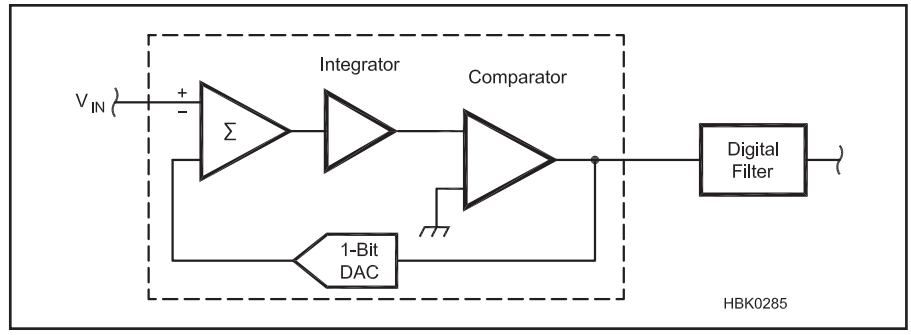

Fig 3.81 — Sigma-delta converter. Similar to the delta-encoded converter (Fig 3.80), the converter runs much faster than the output samples and uses a digital filter to derive the actual output value.

is used to provide a replica of the voltage on C_{HOLD} to the conversion circuitry.

When a conversion is started, a digital control signal opens the input switch, closes the output switch, and the capacitor's voltage is measured by the converter. It is important that the capacitor used for C_{HOLD} have low *leakage* so that while the measurement is being made, the voltage stays constant for the few ms required. This is of particular important in high-precision conversion.

Single-Ended and Differential Inputs

The input of most ADCs is *single-ended*, in which the input signal is measured between the input pin and a common ground. Shown in **Fig 3.83A**, this is acceptable for most applications, but if the voltage to be measured is small or is the difference between two nonzero voltages, an ADC with *differential inputs* should be used as in Fig 3.83B. Differential inputs are also useful when measuring current as the voltage across a small resistor in series with the current. In that case, neither side of the resistor is likely to be at ground, so a differential input is very useful. Differential inputs also help avoid the issue of noise contamination as discussed below.

Input Buffering and Filtering

The input impedance of most ADCs is high enough that the source of the input signal is unaffected. However, to protect the ADC input and reduce loading on the input source, an external buffer stage can be used. **Fig 3.84** shows a typical buffer arrangement with clamping diodes to protect against electrostatic discharge (ESD) and an RC-filter to prevent RF signals from affecting the input signal. In addition, to attenuate higher-frequency signals that might cause aliases, the input filters can also act as band-limiting filters.

Analog and Digital "Ground"

By definition, ADCs straddle the analog and digital domain. In principle, the signals remain separate and isolated from each other. In practice, however, voltages and currents from the analog and digital circuitry can be mixed together. This can result in the contamination of an analog signal with components of digital signals, and rarely, vice versa. Mostly, this is a problem when trying to measure small voltages in the presence of large power or RF signals.

The usual problem is that currents from high-speed digital circuitry find their way into analog signal paths and create transients and other artifacts that affect the measurement of the analog signal. Thus, it is important to have separate current paths for the two types of signals. The manufacturer of the converter will provide guidance for the proper use of the converter either in the device's data sheet or as application notes. Look for separate pins on

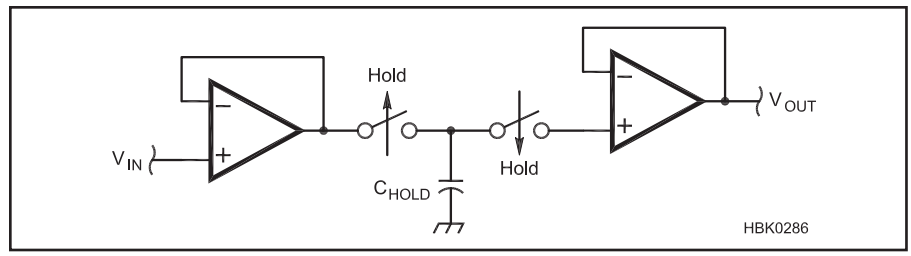

Fig 3.82 — Sample-and-hold (S/H). An input buffer isolates the sampled voltage from the input signal by charging the capacitor C_{HOLD} to that voltage with the input switch closed and the output switch open. When a measurement is being taken, the input switch is open to prevent the input signal from changing the capacitor voltage, and the output switch is closed so that the output buffer can generate a steady voltage at its output.

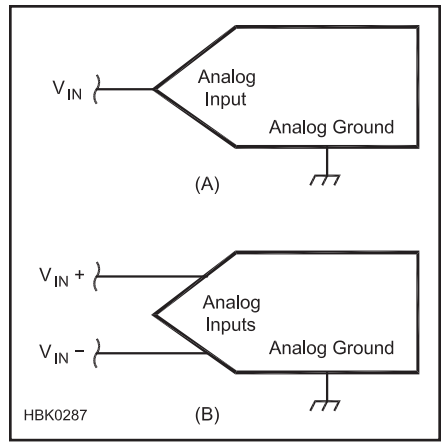

Fig 3.83 — Single-ended ADC inputs have a single active line and a ground or return line (A). Single-ended ADC input are often susceptible to noise and common mode signals or any kind of disturbance on their ground rails. In (B), the differential inputs used help the circuit "ignore" offsets and shifts in the input signal.

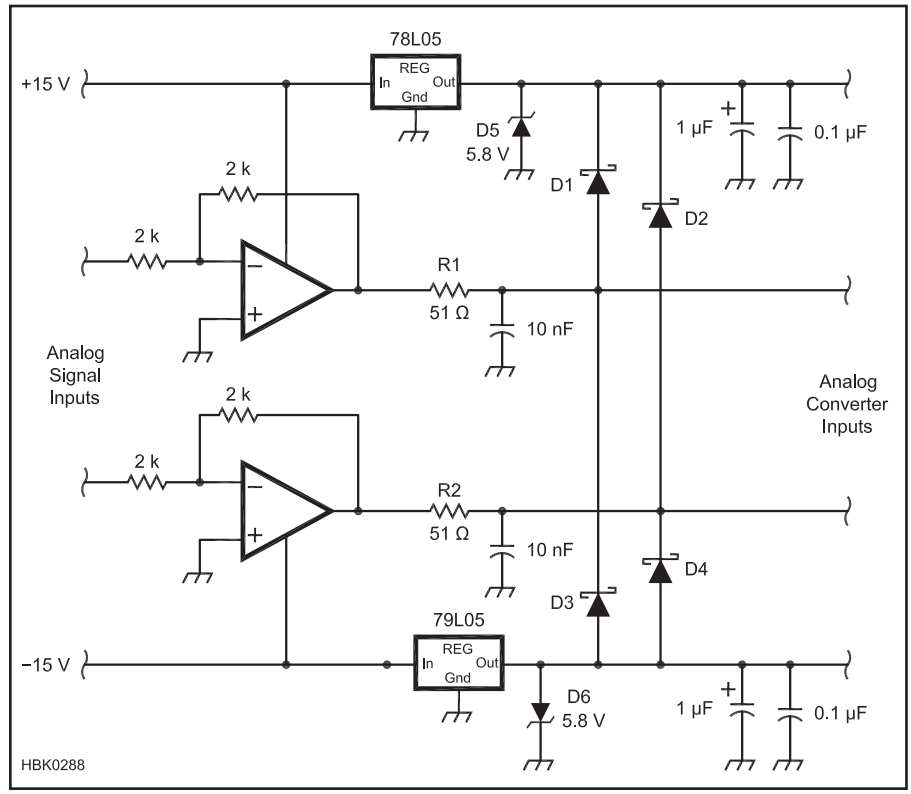

Fig 3.84 — Typical ADC input buffer-filter circuit. Unity-gain voltage followers help isolate the ADC from the input source. RC filters following the buffers act as band-limiting filters to prevent aliasing. Zener diodes are used to clamp the transient voltage and route the energy of transient into the power supply system.

the converter, such as "AGND" or "DGND" that indicate how the two types of signal return paths should be connected.

3.7.3 Digital-to-Analog Converters (DAC)

Converting a digital value to an analog quantity is considerably simpler than the reverse, but there are several issues primarily associated with DACs that affect the selection of a particular converter.

As each new digital value is converted to analog, the output of the DAC makes an abrupt *step change*. Even if very small, the response of the DAC output does not respond perfectly or instantaneously. *Settling time* is the amount of time required for the DAC's output to stabilize within a certain amount of the final value. It is specified by the manufacturer and can be degraded if the load connected to the DAC is too heavy or if it is highly reactive. In these cases, using a buffer amplifier is recommended.

Monotonicity is another aspect of characterizing the DAC's accuracy. A DAC is monotonic if in increasing the digital input value linearly across the conversion range, the output of the DAC increases with every step. Because of errors in the internal conversion circuitry, it is possible for there to be some steps that are too small or too large, leading to output values that seem "out of order." These are usually quite small, but if used in a precision application, monotonicity is important.

SUMMING DAC

A *summing DAC*, shown in **Fig 3.85**, is a summing amplifier with all of the inputs connected to a single reference voltage through switches. The digital value to be converted controls which switches are closed. The larger the digital value, the more switches are closed. Higher current causes the summing amplifier's output voltage to be higher, as well. Most summing DACs have a digital signal interface to hold the digital value between successive conversions, but not all, so be sure before you select a particular DAC.

The input resistors are *binary weighted* so that the summing network resistors representing the more significant bit values inject more current into the op amp's summing junction. Each resistor differs from its neighboring resistors in the amount of current it injects into the summing node by a factor of two, recreating the effect of each digital bit inthe output voltage. At high resolutions, this becomes a problem because of the wide spread in resistor values — a 12-bit DAC would require a spread of 2048 betweenthe largest and smallest resistor values. Summing DACs are generally only available with low resolution for that reason.

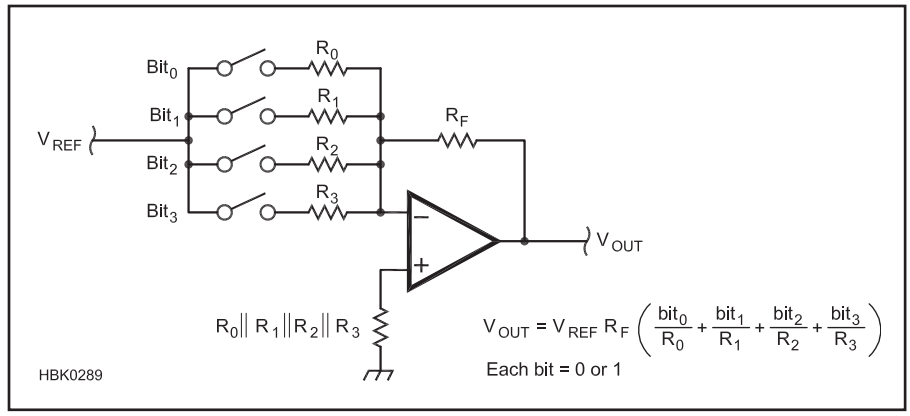

Fig 3.85 — Summing DAC. The output voltage is the inverted, weighted sum of the inputs to each summing resistor at the input. Digital data at the input controls the current into the summing resistors and thus, the output voltage.

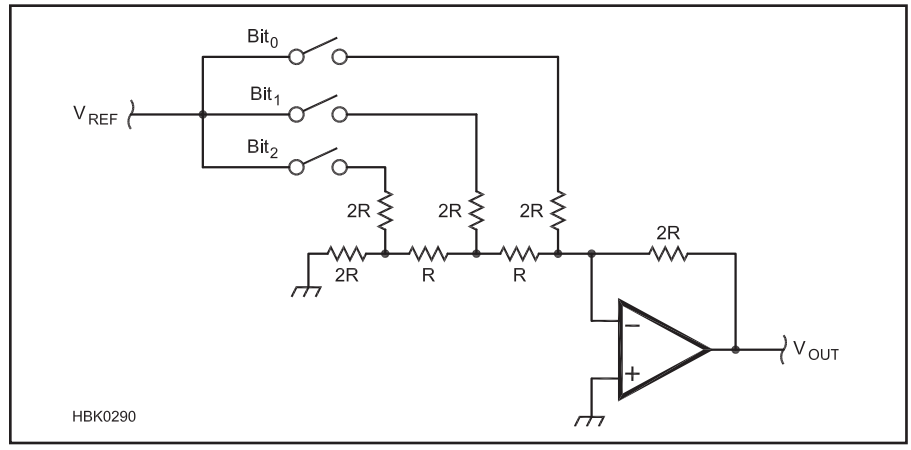

Fig 3.86 — R-2R Ladder DAC. This is the most common form of DAC because all of the resistor values are similar, making it easier to manufacture. The similarity in resistor values also means that there will be less variation of the comparator with temperature and other effects that affect all resistors similarly.

The *current output DAC* functions identically to the summing DAC, but does not have an op amp to convert current in the digitally-controlled resistor network to voltage. It consists only of the resistor network, so an external current-to-voltage circuit (discussed in the previous section on op amps) is required to change the current to a voltage. In some applications, the conversion to voltage is not required or it is already provided by some other circuit.

R-2R LADDER DAC

The summing and current output DACs both used binary weighted resistors to convert the binary digital value into the analog output. The practical limitation of this design is the large difference in value between the smallest and largest resistor. For example, in a 12-bit DAC, the smallest and largest resistors differ by a factor of 2048 (which is 2^{12-1}). This can be difficult to fabricate in an IC without expensive *trimming* processes that adjust each resistor to the correct value and that maintain the powers-of-two relationship over wide temperature variations. For DACs with high-resolutions of 8 bits or more, the R-2R ladder DAC of **Fig 3.86** is a better design.

If the resistances are fairly close in value, the problems of manufacturing are greatly reduced. By using the *R-2R ladder* shown in the figure, the same method of varying current injected into an op amp circuit's summing junction can be used with resistors of only two values, R and 2R. In fact, since the op amp feedback resistor is also one of the IC resistors, the absolute value of the resistance R is unimportant, as long as the ratio of R:2R is maintained. This simplifies manufacturing greatly and is an example of IC design being based on ratios instead of absolute values. For this reason, most DACs use the R-2R ladder design and the performance differences lie mostly in their speed and accuracy.

3.7.4 Choosing a Converter

From the point of view of performance, choosing a converter, either an ADC or a DAC, comes down to resolution, accuracy and speed. Begin by determining the percent resolution or the dynamic range of the converter. Use the equations in the preceding section to determine the number of bits the converter must have. Select from converters with the next highest number of bits. For example, if you determine that you need 7 bits of resolution, use an 8-bit converter.

Next, consider accuracy. If the converter is needed for test instrumentation, you'll need to perform an *error budget* on the instrument's conversion processes, include errors in the analog circuitry. Once you have calculated percent errors, you can determine the requirements for FS error, offset error, and nonlinearities. If an ADC is going to be used for receiving applications, the spur-free dynamic range may be more important than high precision. (The chapter on **DSP and Software Radio Design** goes into more detail about the performance requirements for ADCs used in these applications.)

The remaining performance criterion is the speed and rate at which the converter can operate. Conversions should be able to be made at a minimum of twice the highest frequency of signal you wish to reproduce. (The signal is assumed to be a sinusoid. If a complex waveform is to be converted, you must account for the higher frequency signals that create the nonsinusoidal shapes.) If the converter will be running near its maximum rate, be sure that the associated digital interface, supporting circuitry, and software can support the required data rates, too!

Having established the conversion performance requirements, the next step is to consider cost, amount of associated circuitry, power requirements, and so forth. For example, a self-contained ADC is easier to use and takes up less PC board space, but is not as flexible as one that allows the designer to use an external voltage reference to set the conversion range. Many DACs with "current output" in their description are actually R-2R DACs with additional analog circuitry to provide the current output. DACs are available with both current and voltage outputs, as well. Other considerations, such as the nature of the required digital interface, as discussed in the next section, can also affect the selection of the converter.

3.8 Miscellaneous Analog ICs

The three main advantages of designing a circuit into an IC are to take advantage of the matched characteristics of its components, to make highly complex circuitry more economical, and to miniaturize the circuit and reduce power consumption. As circuits standardize and become widely used, they are often converted from discrete components to integrated circuits. Along with the op amp described earlier, there are many such classes of linear ICs.

3.8.1 Transistor and Driver Arrays

The most basic form of linear integrated circuit and one of the first to be implemented 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 because each device requires a separate off-chip connection. 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.

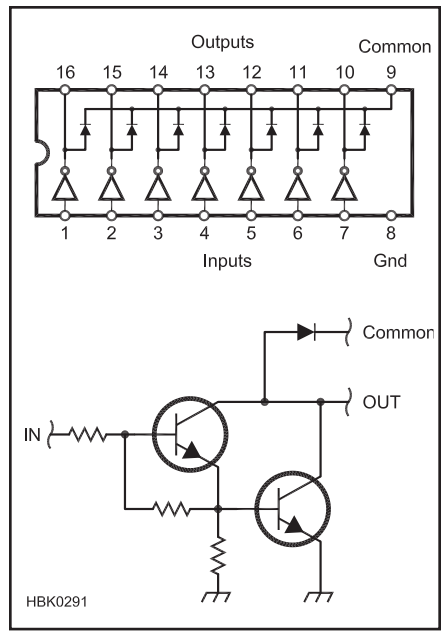

Fig 3.87 — Typical ULN2000-series driver array configuration and internal circuit. The use of driver array ICs is very popular as an interface between microprocessor or other low-power digital circuits and loads such as relays, solenoids or lamps.

Driver arrays, such as the ULN2000-series devices shown in **Fig 3.87** are very useful in creating an interface between low-power circuits such as microprocessors and higher-power loads and indicators. Each driver consists of a Darlington pair switching circuit as described earlier in this chapter. There are different versions with different types and arrangements of resistors and diodes.

Many manufacturers offer driver arrays. They are available with built-in kickback diodes to allow them to drive inductive loads, such as relays, and are heavy enough to source or sink current levels up to 1 A. (All of the drivers in the array can not operate at full load at the same time, however. Read the data sheet carefully to determine what limitations on current and power dissipation may exist.)

3.8.2 Voltage Regulators and References

One of the most popular linear ICs is the voltage regulator. There are two basic types, the three-terminal regulator and the regulator controller. Examples of both are described in the **Power Sources** chapter.

The three-terminal regulator (input, ground, output) is a single package designed to perform all of the voltage regulation functions. The output voltage can be fixed, as in the 7800-series of regulators, or variable, as in the LM317 regulator. It contains a voltage reference, comparator circuits, current and temperature sensing protective circuits, and the main pass element. These ICs are usually contained in the same packages as power transistors and the same techniques of thermal

management are used to remove excess heat.

Regulator controllers, such as the popular 723 device, contain all of the control and voltage reference circuitry, but require external components for the main pass element, current sensing, and to configure some of their control functions.

Voltage references such as the Linear Technology LT1635 are special semiconductor diodes that have a precisely controlled I-V characteristic. A buffer amplifier isolates the sensitive diode and provides a low output impedance for the voltage signal. Voltage references are used as part of power regulators and by analog-digital converter circuits.

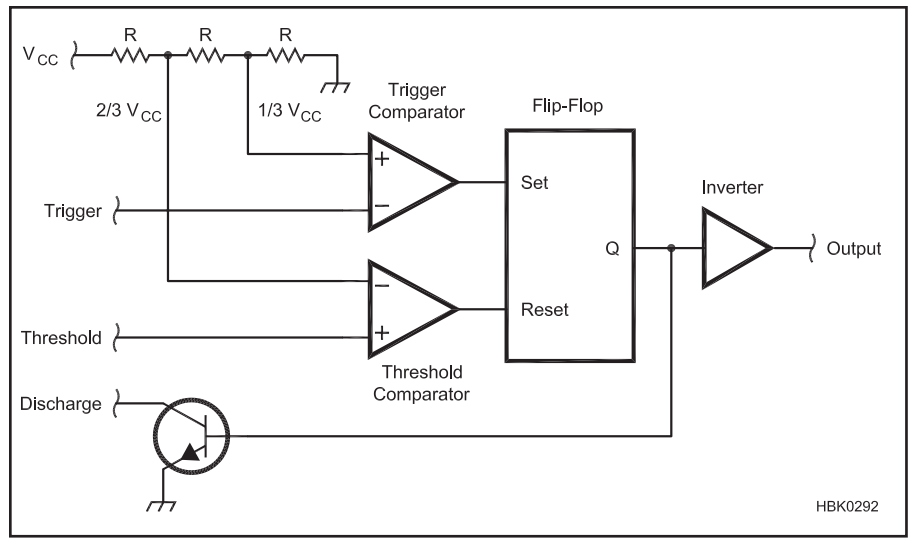

Fig 3.88 — Internal NE555 timer components. This simple array of components combine to make one of the most popular analog ICs. The 555 timer IC uses ratios of internal resistors to generate a precise voltage reference for generating time intervals based on charging and discharging a capacitor.

3.8.3 Timers (Multivibrators)

A *multivibrator* is a circuit that oscillates between two states, usually with a square wave or pulse train output. 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 timer IC (NE555 by Signetics [now Philips] or LM555 by National Semiconductor). This very simple eight-pin 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 *a stable* or *free-running* operation, where the device oscillates continuously. Other applications of a multivibrator include a frequency divider, a delay line, a pulse width modulator and a pulse position modulator. (These can be found in the IC's data sheet or in the reference listed at the end of this chapter.)

Fig 3.88 shows the basic components of a 555. Connected between power input (V_{cc}) and ground, the three resistors labeled "R" at the top left of the figure form a *voltage divider* that divides V_{CC} into two equal steps—one at ⅔ V_{CC} and one at ⅓ V_{CC}. These serve as reference voltages for the rest of the circuit.

Connected to the reference voltages are blocks labeled *trigger comparator* and *threshold comparator*. (Comparators were discussed in a preceding section.) The trigger comparator in the 555 is wired so that its output is high whenever the trigger input is *less* than ⅓ V_{CC} and vice versa. Similarly, the threshold comparator output is high whenever the threshold input is *greater* than ⅔ V_{CC}. These two outputs control a digital *flip-flop* circuit. (Flip-flops are discussed in the **Digital Basics** chapter.)

The flip-flop output, *Q,* changes to high or low when the state of its *set* and *reset* input changes. The Q output stays high or low (it *latches* or *toggles*) until the opposite input changes. When the set input changes from low to high, Q goes low. When reset changes from low to high, Q goes high. The flip-flop ignores any other changes. An inverter makes the 555 output high when Q is low and vice versa — this makes the timer circuit easier to interface with external circuits.

The transistor connected to Q acts as a switch. When Q is high, the transistor is on and acts as a closed switch connected to ground. When Q is low, the transistor is off and the switch is open. These simple building blocks — voltage divider, comparator, flip-flop and switch — build a surprising number of useful circuits.

THE MONOSTABLE OR "ONE-SHOT" TIMER

The simplest 555 circuit is the monostable circuit. This configuration will output one fixed-length pulse when triggered by an input pulse. **Fig 3.89** shows the connections for this circuit.

Starting with capacitor C discharged, the flip-flop output, Q, is high, which keeps the discharge transistor turned on and the voltage on C below ⅔ V_{CC}. The circuit is in its stable state, waiting for a trigger pulse.

When the voltage at the trigger input drops below ⅓ V_{CC}, the trigger comparator output changes from low to high, which causes Q to toggle to the low state. This turns off the transistor (opens the switch) and allows C to begin charging toward V_{CC}.

When C reaches ⅔ V_{CC}, this causes the threshold comparator to switch its output from low to high and that resets the flip-flop. Q returns high, turning on the transistor and discharging C. The circuit has returned to its stable state. The output pulse length for the monostable configuration is:

$$T = 1.1\ R\ C_1 \qquad (73)$$

Notice that the timing is independent of the absolute value of V_{CC} — the output pulse width is the same with a 5 V supply as it is with a 15 V supply. This is because the 555 design is based on ratios and not absolute voltage levels.

THE ASTABLE MULTIVIBRATOR

The complement to the monostable circuit is the astable circuit in **Fig 3.90**. Pins 2, 6 and 7 are configured differently and timing resistor is now split into two resistors, R1 and R2.

Start from the same state as the monostable circuit, with C completely discharged. The monostable circuit requires a trigger pulse to initiate the timing cycle. In the astable circuit, the trigger input is connected directly to the capacitor, so if the capacitor is discharged, then the trigger comparator output must be high. Q is low, turning off the discharge transistor, which allows C to immediately begin charging.

C charges toward V_{CC}, but now through the combination of R1 and R2. As the capacitor voltage passes ⅔ V_{CC}, the threshold comparator output changes from low to high, resetting Q to high. This turns on the discharge transistor and the capacitor starts to discharge through R2. When the capacitor is discharged below ⅓ V_{CC}, the trigger comparator changes from high to low and the cycle begins again, automatically. This happens over and over, causing a train of pulses at the output while C charges and discharges between ⅓ and ⅔ V_{CC} as seen in the figure.

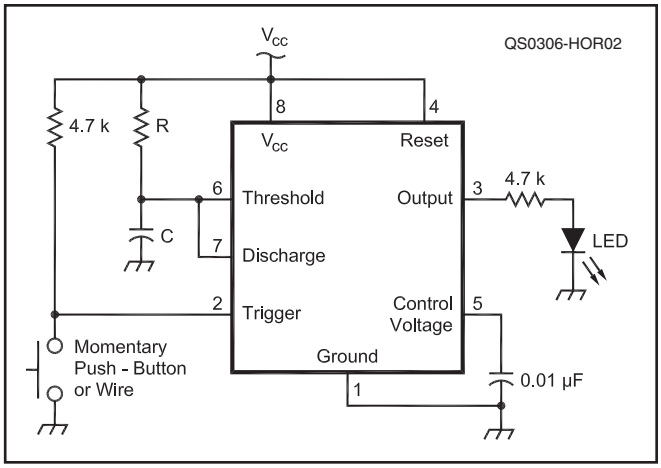

Fig 3.89 — Monostable timer. The timing capacitor is discharged until a trigger pulse initiates the charging process and turns the output on. When the capacitor has charged to 2/3 V_{CC}, the output is turned off, the capacitor is discharged and the timer awaits the next trigger pulse.

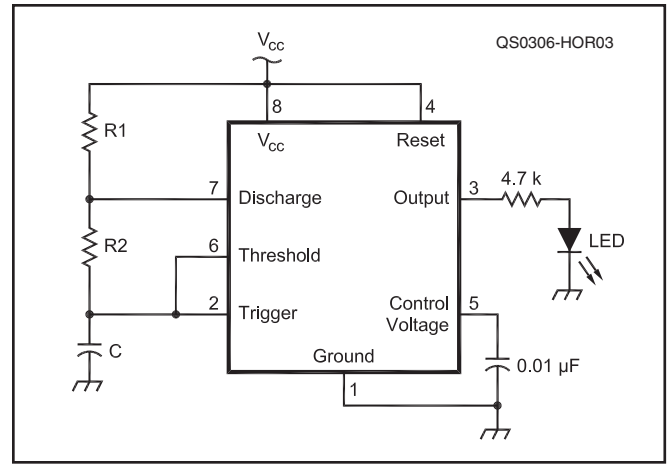

Fig 3.90 — Astable timer. If the capacitor discharge process initiates the next charge cycle, the timer will output a pulse train continuously.

The total time it takes for one complete cycle is the charge time, T_c, plus the discharge time, T_d:

$$T = T_c + T_d = 0.693\,(R_1 + R_2)\,C + 0.693\,R_2C$$
$$= 0.693\,(R_1 + 2R_2)\,C \qquad (74)$$

and the output frequency is:

$$f = \frac{1}{T} = \frac{1.443}{(R_1 + 2\,R_2)\,C} \qquad (75)$$

When using the 555 in an application in or around radios, it is important to block any RF signals from the IC power supply or timing control inputs. Any unwanted signal present on these inputs, especially the Control Voltage input, will upset the timer's operation and cause it to operate improperly. The usual practice is to use a 0.01 µF bypass capacitor (shown on pin 5 in both Fig 3.89 and 3.90) to bypass ac signals such as noise or RF to ground. Abrupt changes in V_{CC} will also cause changes in timing and these may be prevented by connecting filter capacitors at the V_{CC} input to ground.

3.8.4 Analog Switches and Multiplexers

Arrays of analog switches, such as the Maxim MAX312-series, allow routing of audio through lower frequency RF signals without mechanical switches. There are several types of switch arrays. Independent switches have isolated inputs and outputs and are turned on and off independently. Both SPST and SPDT configurations are available. Multiple switches can be wired with common control signals to implement multiple-pole configurations.

Use of analog switches at RF through microwave frequencies requires devices specifically designed for those frequencies. The Analog Devices ADG901 is a switch usable to 2.5 GHz. It absorbs the signal when off, acting as a terminating load. The ADG902 instead reflects the signal as an open circuit when off. Arrays of three switches called "tee-switches" are used when very high isolation between the input and output is required.

Multiplexers or "muxes" are arrays of SPST switches configured to act as a multiposition switch that connects one of four to sixteen input signals to a single output. Demultiplexers ("demuxes") have a single input and multiple outputs. Multiplexer ICs are available as single N-to-1 switches (the MAX4617 is an 8-to-1 mux) or as groups of N-to-1 switches (the MAX4618 is a dual 4-to-1 mux).

Crosspoint switch arrays are arranged so that any of four to sixteen signal inputs can be connected to any of four to sixteen output signal lines. The Analog Devices AD8108 is an 8-by-8 crosspoint switch with eight inputs and eight outputs. These arrays are used when it is necessary to switch multiple signal sources among multiple signal receivers. They are most commonly used in telecommunications.

All analog switches use FET technology as the switching element. To switch ac signals, most analog switches require both positive and negative voltage power supplies. An alternative is to use a single power supply voltage and ground, but bias all inputs and output at one-half the power supply voltage. This requires dc blocking capacitors in all signal paths, both input and output, and loading resistors may be required at the device outputs. The blocking capacitors can also introduce low-frequency roll-off.

The impedance of the switching element varies from a few ohms to more than 100 ohms. Check the switch data sheet to determine the limits for how much power and current the switches can handle. Switch arrays, because of the physical size of the array, can have significant coupling or *crosstalk* between signal paths. Use caution when using analog switches for high-frequency signals as coupling generally increases with frequency and may compromise the isolation required for high selectivity in receivers and other RF signal processing equipment.

3.8.5 Audio Output Amplifiers

While it is possible to use op amps as low power audio output drivers for headphones, they generally have output impedances that are too high for most audio transducers such as speakers and headphones. The LM380 series of audio driver ICs has been used in radio circuits for many years and a simple schematic for a speaker driver is shown in **Fig 3.91**.

The popularity of personal music players has resulted in the creation of many new and inexpensive audio driver ICs, such as the National Semiconductor LM4800- and LM4900-series. Drivers that operate from voltages as low as 1.5 V for battery-powered devices and up to 18 V for use in vehicles are now available.

When choosing an audio driver IC for communications audio, the most important parameters to evaluate are its power requirements and power output capabilities. An overloaded or underpowered driver will result in distortion. Driver ICs intended for music players have frequency responses well in excess of the 3000 Hz required for communications. This can lead to annoying and fatiguing hiss unless steps are taken to reduce

the circuit's frequency response.

Audio power amplifiers should also be carefully decoupled from the power supply and the manufacturer may recommend specific circuit layouts to prevent oscillation or feedback. Check the device's data sheet for this information.

3.8.6 Temperature Sensors

Active temperature sensors use the temperature-dependent properties of semiconductor devices to create voltages that correspond to absolute temperature in degrees Fahrenheit (LM34) or degrees Celsius (LM35). These sensors (of which many others are available than the two examples given here) are available in small plastic packages, both leaded and surface-mount, that respond quickly to temperature changes. They are available with 1% and better accuracy, requiring only a source of voltage at very low current and ground. Complete application information is available in the manufacturer data sheets. Thermistors, a type of passive temperature sensor, are discussed in the **Electrical Fundamentals** chapter. Temperature sensors are used in radio mostly in cooling and thermal management systems.

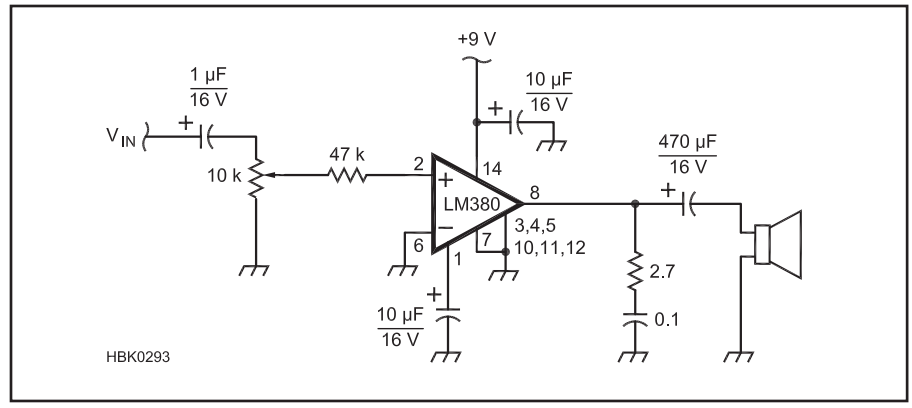

Fig 3.91 — Speaker driver. The LM380-series of audio output drivers are well-suited for low-power audio outputs, such as for headphones and small speakers. When using IC audio output drivers, be sure to refer to the manufacturer's data sheet for layout and power supply guidelines.

3.8.7 Electronic Subsystems

As a particular technology becomes popular, a wave of integrated circuitry is developed to service that technology and reduce its cost of production and service. A good example is the wireless telephony industry. IC manufacturers have developed a large number of devices targeting this industry; receivers, transmitters, couplers, mixers, attenuators, oscillators and so forth. In addition, other wireless technologies such as data transmission provide opportunities for manufacturers to create integrated circuits that implement radio-related functions at low cost. Analog ICs used in the construction of various radio systems and supporting equipment are discussed in the appropriate chapters of this book.

3.9 Analog Glossary

AC ground — A circuit connection point that presents a very low impedance to ac signals.

Accuracy — The ability of an analog-to-digital conversion to assign the correct code to an analog value or create the true analog value from a specific code.

Active — A device that requires power to operate.

Active region — The region in the characteristic curve of an analog device in which it is capable of processing the signal linearly.

Amplification — The process by which amplitude of a signal is increased. Gain is the amount by which the signal is amplified.

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

Analog-to-digital converter (ADC) — Circuit (usually an IC) that generates a digital representation of an analog signal.

Anode — The element of an analog device that accepts electrons or toward which electrons flow.

Attenuation — The process of reducing the amplitude of a signal.

Avalanche breakdown — Current flow through a semiconductor device in response to an applied voltage beyond the device's ability to control or block current flow.

Base — The terminal of a bipolar transistor in which control current flows.

Beta (β) — The dc current gain of a bipolar transistor, also designated h_{FE}.

Biasing — The addition of a dc voltage or current to a signal at the input of an analog device, changing or controlling the position of the device's operating point 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.

Black box — Circuit or equipment that is analyzed only with regards to its external behavior.

Bode plot — Graphs showing amplitude response in dB and phase response in degrees versus frequency on a logarithmic scale.

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

Carrier — (1) Free electrons and holes in semiconductor material. (2) An unmodulated component of a modulated signal.

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

Cathode — The element of an analog device that emits electrons or from which electrons are emitted or repelled.

Characteristic curve — A plot of the relative responses of two or three analog-device parameters, usually of an output with respect to an input. (Also called *I-V* or *V-I curve.*)

Class — For analog amplifiers (Class A, B, AB, C), a categorization of the fraction of the input signal cycle during which the amplifying device is active. For digital or switching amplifiers (Class D and above), a categorization of the method by which the signal is amplified.

Clipping — A nonlinearity in amplification in which the signal's amplitude can no longer be increased, usually resulting in distortion of the waveform. (Also called *clamping* or *limiting.*)

Closed-loop gain — Amplifier gain with an external feedback circuit connected.

Collector — The terminal of a bipolar transistor from which electrons are removed.

Code — One possible digital value representing an analog quantity. (Also called *quantization code*.)

Common — A terminal shared by more than one port of a circuit or network.

Common mode — Signals that appear equally on all terminals of a signal port.

Comparator — A circuit, usually an amplifier, whose output indicates the relative amplitude of two input signals.

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

Compression — Reducing the dynamic range of a signal in order to increase the average power of the signal or prevent excessive signal levels.

Conversion efficiency — The amount of light energy converted to electrical energy by a photoelectric device, expressed in percent.

Conversion rate — The amount of time in which an analog-digital conversion can take place. *Conversion speed* is the reciprocal of conversion rate.

Coupling (ac or dc) — The type of connection between two circuits. DC coupling allows dc current to flow through the connection. AC coupling blocks dc current while allowing ac current to flow.

Cutoff frequency — Frequency at which a circuit's amplitude response is reduced to one-half its mid-band value (also called *half-power* or *corner* frequency).

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.

Degeneration (emitter or source) — Negative feedback from the voltage drop across an emitter or source resistor in order to stabilize a circuit's bias and operating point.

Depletion mode — An FET with a channel that conducts current with zero gate-to-source voltage and whose conductivity is progressively reduced as reverse bias is applied.

Depletion region — The narrow region at a PN junction in which majority carriers have been removed. (Also called *space-charge* or *transition* region.)

Digital-to-analog converter (DAC) — Circuit (usually an IC) that creates an analog signal from a digital representation.

Diode — A two-element semiconductor with a cathode and an anode that conducts current in only one direction.

Drain — The connection at one end of a field-effect-transistor channel from which electrons are removed.

Dynamic range — The range of signal levels over which a circuit operates properly. Usually refers to the range over which signals are processed linearly.

Emitter — The terminal of a bipolar transistor into which electrons are injected.

Enhancement mode — An FET with a channel that does not conduct with zero gate-to-source voltage and whose conductivity is progressively increased as forward bias is applied.

Feedback — Routing a portion of an output signal back to the input of a circuit. Positive feedback causes the input signal to be reinforced. Negative feedback results in partial cancellation of the input signal.

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

Forward bias — Voltage applied across a PN junction in the direction to cause current flow.

Forward voltage — The voltage required to cause forward current to flow through a PN junction.

Free electron — An electron in a semiconductor crystal lattice that is not bound to any atom.

Frequency response — A description of a circuit's gain (or other behavior) with frequency.

Gain — see ***Amplification***.

Gain-bandwidth product — The relationship 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 control electrode of a field-effect transistor.

High-side — A switch or controlling device connecting between a power source and load.

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

Hysteresis — In a comparator circuit, the practice of using positive feedback to shift the input setpoint in such a way as to minimize output changes when the input signal(s) are near the setpoint.

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.

Isolation — Eliminating or reducing electrical contact between one portion of a circuit and another or between pieces of equipment.

Junction FET (JFET) — A field-effect transistor whose gate electrode forms a PN junction with the channel.

Linearity — Processing and combining of analog signals independently of amplitude.

Load line — A line drawn through a family of characteristic curves that shows the operating points of an analog device for a given load or circuit component values.

Loading — The condition that occurs when the output behavior of a circuit is affected by the connection of another circuit to that output.

Low-side — A switch or controlling device connected between a load and ground.

Metal-oxide semiconductor (MOSFET) — A field-effect transistor whose gate is insulated from the channel by an oxide layer. (Also called *insulated gate FET* or *IGFET*)

Multivibrator — A circuit that oscillates between two states.

NMOS — N-channel MOSFET.

N-type impurity — A doping atom with an excess of valence electrons that is added to semiconductor material to act as a source of free electrons.

Network — General name for any type of circuit.

Noise — Any unwanted signal, usually random in nature.

Noise figure (NF) — A measure of the noise added to a signal by an analog processing stage, given in dB. (Also called *noise factor.*)

Open-loop gain — Gain of an amplifier with no feedback connection.

Operating point — Values of a set of circuit parameters that specify a device's operation at a particular time.

Operational amplifier (op amp) — An integrated circuit amplifier with high open-loop gain, high input impedance, and low output impedance.

Optoisolator — A device in which current in a light-emitting diode controls the operation of a phototransistor without a direct electrical connection between them.

Oscillator — A circuit whose output varies continuously and repeatedly, usually at a single frequency.

P-type impurity — A doping atom with a shortage of valence electrons that is added to semiconductor material to create an excess of holes.

Passive — A device that does not require power to operate.

Peak inverse voltage (PIV) — The highest voltage that can be tolerated by a reverse biased PN junction before current is conducted. (See also ***avalanche breakdown***.)

Photoconductivity — Phenomenon in

which light affects the conductivity of semiconductor material.

Photoelectricity — Phenomenon in which light causes current to flow in semiconductor material.

PMOS — P-channel MOSFET.

PN junction — The structure that forms when P-type semiconductor material is placed in contact with N-type semiconductor material.

Pole — Frequency at which a circuit's transfer function becomes infinite.

Port — A pair of terminals through which a signal is applied to or output from a circuit.

Quiescent (Q-) point — Circuit or device's operating point with no input signal applied. (Also called *bias point.*)

Pinch-off — The condition in an FET in which the channel conductivity has been reduced to zero.

Products — Signals produced as the result of a signal processing function.

Rail — Power supply voltage(s) for a circuit.

Range — The total span of analog values that can be processed by an analog-to-digital conversion.

Recombination — The process by which free electrons and holes are combined to produce current flow across a PN junction.

Recovery time — The amount of time required for carriers to be removed from a PN junction device's depletion region, halting current flow.

Rectify — Convert ac to pulsating dc.

Resolution — Smallest change in an analog value that can be represented in a conversion between analog and digital quantities. (Also called *step size*.)

Reverse bias — Voltage applied across a PN junction in the direction that does not cause current flow.

Reverse breakdown — The condition in which reverse bias across a PN junction exceeds the ability of the depletion region to block current flow. (See also ***avalanche breakdown.***)

Roll-off — Change in a circuit's amplitude response per octave or decade of frequency.

Safe operating area (SOA) — The region of a device's characteristic curve in which it can operate without damage.

Sample — A code that represents the value of an analog quantity at a specific time.

Saturation (Region) — The region in the characteristic curve of an analog device in which the output signal can no longer be increased by the input signal. See ***Clamping***.

Schottky barrier — A metal-to-semiconductor junction at which a depletion region is formed, similarly to a PN junction.

Semiconductor — (1) An element such as silicon with bulk conductivity between that of an insulator and a metal. (2) An electronic device whose function is created by a structure of chemically-modified semiconductor materials.

Signal-to-noise ratio (SNR) — The ratio of the strength of the desired signal to that of the unwanted signal (noise), usually expressed in dB.

Slew rate — The maximum rate at which a device can change the amplitude of its output.

Small-signal — Conditions under which the variations in circuit parameters due to the input signal are small compared to the quiescent operating point and the device is operating in its active region.

Source — The connection at one end of the channel of a field-effect transistor into which electrons are injected.

Stage — One of a series of sequential signal processing circuits or devices.

Substrate — Base layer of material on which the structure of a semiconductor device is constructed.

Superposition — Process in which two or more signals are added together linearly.

Total harmonic distortion (THD) — A measure of how much noise and distortion are introduced by a signal processing function.

Thermal runaway — The condition in which increasing device temperature increases device current in a positive feedback cycle.

Transconductance — Ratio of output current to input voltage, with units of Siemens (S).

Transfer characteristics — A set of parameters that describe how a circuit or network behaves at and between its signal interfaces.

Transfer function — A mathematical expression of how a circuit modifies an input signal.

Unipolar transistor — see ***Field-effect transistor (FET)***.

Virtual ground — Point in a circuit maintained at ground potential by the circuit without it actually being connected to ground.

Zener diode — A heavily-doped PN-junction diode with a controlled reverse breakdown voltage, used as a voltage reference or regulator.

Zero — Frequency at which a circuit's transfer function becomes zero.

3.10 References and Bibliography

REFERENCES

1. Ebers, J., and Moll, J., "Large-Signal Behavior of Junction Transistors," *Proceedings of the IRE*, 42, Dec 1954, pp 1761-1772.
2. Getreu, I., *Modeling the Bipolar Transistor* (Elsevier, New York, 1979). Also available from Tektronix, Inc, Beaverton, Oregon, in paperback form. Must be ordered as Part Number 062-2841-00.

FURTHER READING

Alexander and Sadiku, *Fundamentals of Electric Circuits* (McGraw-Hill)

Hayward, W., *Introduction to Radio Frequency Design* (ARRL, 2004)

Hayward, Campbell and Larkin, *Experimental Methods in RF Desig*n (ARRL, 2009)

Millman and Grabel, *Microelectronics: Digital and Analog Circuits and Systems* (McGraw-Hill, 1988)

Mims, F., *Timer, Op Amp & Optoelectronic Circuits & Projects* (Master Publishing, 2004)

Jung, W. *IC Op Amp Cookbook* (Prentice-Hall, 1986)

Hill and Horowitz, *The Art of Electronics* (Cambridge University Press, 1989)

Analog-Digital Conversion Handbook (by the staff of Analog Devices, published by Prentice-Hall)

Safe-Operating Area for Power Semiconductors (ON Semi), **www.onsemi.com/pub_link/Collateral/AN875-D.PDF**

"Selecting the Right CMOS Analog Switch", Maxim Semiconductor, **www.maxim-ic.com/appnotes.cfm/an_pk/638**

Hyperphysics Op-Amp Circuit Tutorials, **hyperphysics.phy-astr.gsu.edu/Hbase/Electronic/opampvar.html#c**2