Chapter 5

Electrical Signals and Components

Analog Glossary

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gain — see ***Amplification***.

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

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

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

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

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

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

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

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

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

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

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

Noise — Any unwanted signal.

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

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

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

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

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

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

Plate — See anode, usually used with vacuum tubes.

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

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

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

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

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

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

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

Tetrode — A four-element vacuum tube with a cathode, a control grid, a screen grid, and a plate.

Triode — A three-element vacuum tube with a cathode, a grid, and a plate.

Unipolar Transistor — see ***Field-Effect Transistor (FET)***.

Zener Diode — A PN-junction diode with a controlled peak inverse voltage so that it will start conducting current at a preset reverse voltage.

Introduction

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

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

Analog Signal Processing

LINEARITY

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

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

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

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

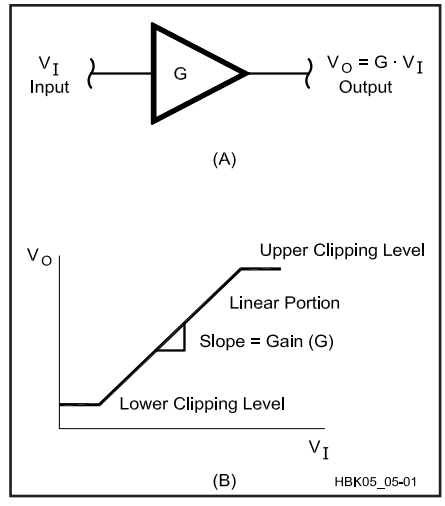


Fig 5.1 — Generic amplifier. (A) Symbol. For the linear amplifier, gain is the constant value, G, and the output voltage is equal to the input voltage times G; (B) Transfer function, input voltage along the x-axis is converted to the output voltage along the y-axis. The linear portion of the response is where the plot is diagonal; its slope is equal to the gain, G. Above and below this range are the clipping limits, where the response is not linear and the output signal is clipped.

Linear Operations

Any operation that modifies a signal and obeys the rules of superposition and scaling is a *linear operation*. The most basic linear operation occurs in an amplifier, a circuit that increases the amplitude of a signal. Schematically, a generic amplifier is signified by a triangular symbol, its input along the left face and its output at the point on the right (see **Fig 5.1**). The linear amplifier multiplies every value of a signal by a constant value. Amplifier gain is often expressed as a multiplication factor (x 5, for example).

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

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

Ideal linear amplifiers have the same gain for all parts of a signal. Thus, a gain of 10 changes 10 V to 100 V, 1 V to 10 V and –1 V to –10 V. Amplifiers are limited by their dynamic range and frequency response, however. An amplifier can only produce output levels that are within the range of its power supply. The power-supply voltages are also called the *rails* of an amplifier. As the amplified output approaches one of the rails, the output will not go beyond a given voltage that is near the rail. The output is limited at the *clipping level* of an amplifier. When an amplifier tries to amplify a signal to be larger than this value, the output remains at this level; this is called output *clipping*. Clipping is a nonlinear effect; an amplifier is considered linear only between its clipping levels. See Fig 5.1.

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

A dc-coupled amplifier equally amplifies signals with frequencies down to dc. An ac-coupled amplifier acts as a high-pass filter, decreasing amplification as the frequency decreases toward dc.

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

Feedback and Oscillation

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

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

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.

Filtering

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

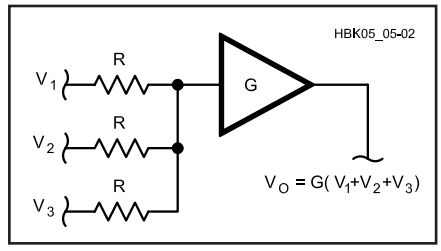


Fig 5.2 — Summing amplifier. The output voltage is equal to the sum of the input voltages times the amplifier gain, G. As long as the resistance values, R, are equal and the amplifier input impedance is much higher, the actual value of R does not affect the output signal.

Summing Amplifiers

In a linear system, nature does most of the work for us when it comes to adding signals; placing two signals together naturally causes them to add. When processing signals, we would like to control the summing operation so the signals do not distort. If two signals come from separate stages and they are connected, the stages may interact, causing both stages to distort their signals. Summing amplifiers generally use a resistor in series with each stage, so the resistors connect to the common input of the following stage. **Fig 5.2** illustrates the resistors connecting to a summing amplifier. Ideally, any time we wanted to combine signals (for example, combining an audio signal with a PL tone in a 2-m FM transmitter prior to modulating the RF signal) we could use a summing amplifier.

Buffering

It is often necessary to isolate the stages of an analog circuit. This isolation reduces the loading, coupling and feedback between stages. An intervening stage, called a *buffer*, is often used for this purpose. A buffer is a linear circuit that is a type of amplifier. It is often necessary to change the characteristic impedance of a circuit

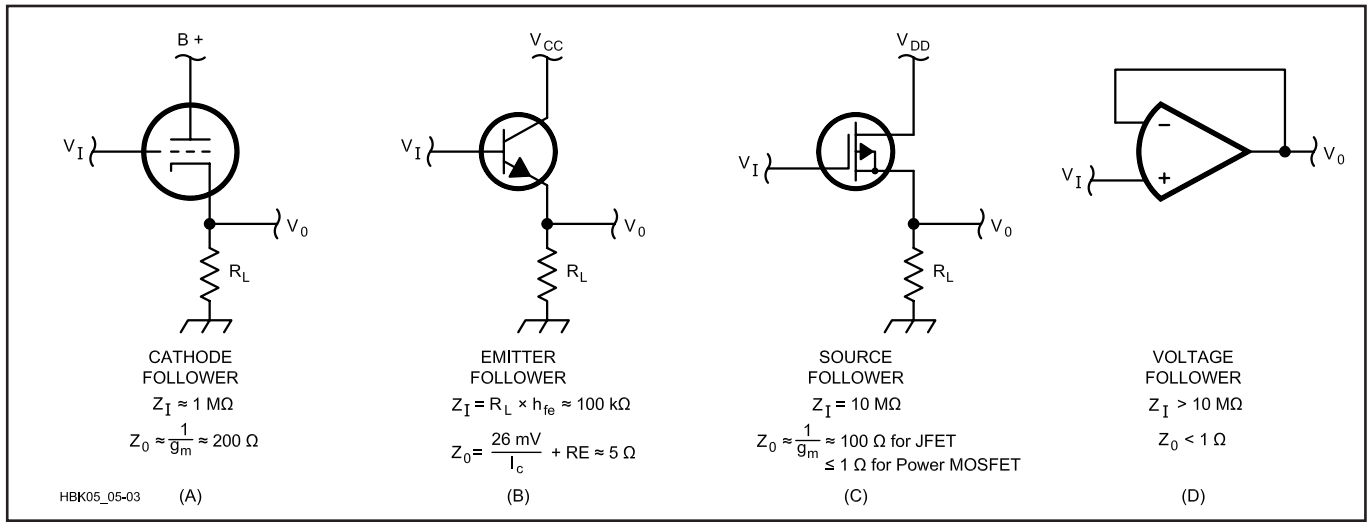


Fig 5.3 — Common buffer stages and some typical input (Z_I) and output (Z_O) impedances. (A) Cathode follower, made with triode tube; (B) Emitter follower, made with NPN bipolar transistor; (C) Source follower, made with FET; and (D) Voltage follower, made with operational amplifier. All of these buffers are terminated with a load resistance, R_L, and have an output voltage that is approximately equal to the input voltage (gain ≈ 1).

between stages. Buffers can have high values of amplification but this is unusual. A buffer performs impedance transformations most efficiently when it has a low or unity gain. **Fig 5.3** shows common forms of buffers with low-impedance outputs: the cathode follower using a triode tube, the emitter follower using a bipolar transistor, the source follower using a field-effect transistor and the voltage follower, using an operational amplifier.

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

Amplitude Modulation/ Demodulation

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

In the linear model of amplitude modulation, the signal to be modulated (such as the audio signal in an AM transmitter) is shifted in frequency by multiplying it with the carrier. The modulated waveform is considered to be a linear function of the signal. The carrier is considered to be part of a time-varying linear system and not a second signal.

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

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

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

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

NONLINEAR OPERATORS

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

Clipping and Rectification

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

Logarithmic Amplification

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

ANALOG BUILDING BLOCKS

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

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

An analog network modifies a signal in a specific way that can be described mathematically. The output is related to the input by a *transfer function*. The mathematical operation that combines a signal with a transfer function is pictured symbolically in **Fig 5.4**. The output signal, w(t), has a

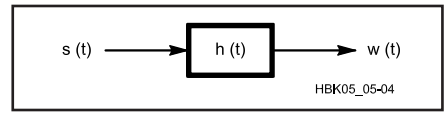


Fig 5.4 — Linear function block. The output signal, w(t) is produced by the action of the transfer function, h(t) on the input signal s(t).

value that changes with time. The output signal is created by the action of an analog transfer function, h(t), on the input signal, g(t).

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

Cascading Stages

If an analog circuit can be described with a transfer function, a combination of analog circuits can also be described similarly. This description of the combined circuits depends upon the relationship between the transfer functions of the parts and that of the combined circuits. In many cases this relationship allows us to predict the behavior of large and complex circuits from what we know about the parts 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 when there is no interaction between stages. In other words, the action of the first stage is unchanged, regardless of whether or not the second stage follows it. Just as the signal entering the first stage is modified by the action of the first transfer function, the ideal cascading of analog circuits results in changes produced only by the individual transfer functions. For any number of stages that are cascaded, the combination of their transfer functions results in a new transfer function. The signal that enters the circuit is changed by the composite transfer function, to produce the signal that exits the cascaded circuits.

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 (oscillation or negative feedback).

Cascaded Buffers

Buffer stages that are made with single active devices can be more effective if cascaded. Two types of such buffers are in common use. The *Darlington pair* is a cascade of two common-collector transistors as shown in **Fig 5.5.** (The various amplifier configurations will be described later in this chapter.) The input impedance of the Darlington pair is equal to the load impedance times the current gain, h_{FE}. The current gain of the Darlington pair is the product of the current gains for the two transistors.

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

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

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

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

A common-emitter amplifier followed by a common-base amplifier is called a *cascode buffer* (see **Fig 5.6**). Cascodes are also made with FETs by following a common-source amplifier by a common-gate configuration. The input impedance and current gain of the cascode are approximately the same as those of the first stage. The output impedance is much higher than that of a single stage. Cascode amplifiers have excellent input/output isolation (very low unwanted feedback) and this can provide high gain with good stability. An example of a cascode buffer made with bipolar transistors has moderate input impedance, Z_I = 1 kΩ, high current gain, h_{FE} = 50 and high output impedance, Z_O = 1 MΩ. There is very little reverse internal feedback in the cascode design, making it very stable, and the amplifier design has little effect on external tuning components. Cascode circuits are often used in tuned amplifier designs for these reasons.

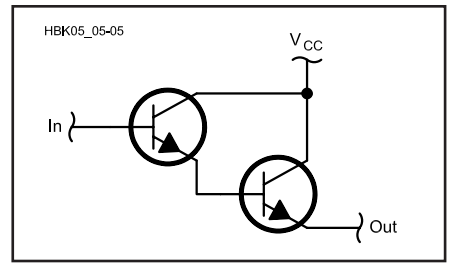


Fig 5.5 — Darlington pair made with two emitter followers. Input impedance, Z_I, is far higher than for a single transistor and output impedance, Z_O, is nearly the same as for a single transistor. DC biasing has been omitted for simplicity.

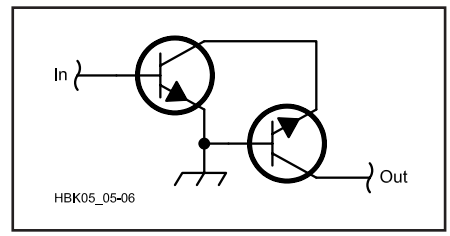


Fig 5.6 — Cascode pair made with two NPN bipolar transistors has a medium input impedance and high output impedance. DC biasing has been omitted for simplicity.

Interstage Loading and Impedance Matching

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

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

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

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

Noise

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

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

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

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

where:

P_{NO} = total noise output power,

A = amplification gain, and

$P_{N\ TH}$ = input thermal noise power.

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

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

Analog Devices

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

TERMINOLOGY

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

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

Amplifier Types

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

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

Characteristic Curves

Analog devices are described most completely with their *characteristic curves*. Almost all devices 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. The characteristic curve is a plot of the interrelationships between two or three variables. The vertical (y) axis parameter is the output, or result of the device being operated with an input parameter on the horizontal (x) axis. Often the output is the result of two input values. The first input parameter is represented along the x axis and the second input parameter by several curves, each for a different value. For example, a vacuum tube characteristic curve may have the plate current along the y axis, the grid voltage along the x axis and several curves, each representing a different value of the plate bias voltage (see **Fig 5.7**).

The parameters plotted in the characteristic curve depend on how the device will be used. The common amplifier configuration defines the input and output leads, and their relationship is diagrammed by the curves. Device parameters are usually derived from the characteristic curve. To calculate a gain,

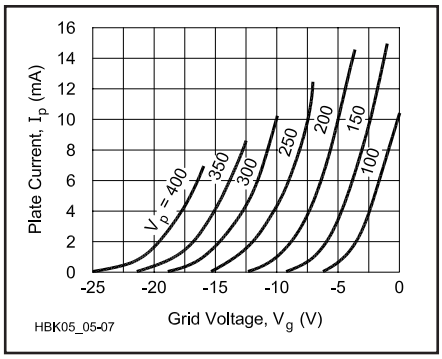


Fig 5.7 — Tube characteristic curve. Input signal is the grid voltage, V_g, along the x-axis, and the output signal is the plate current, I_p, along the y-axis. Different curves are plotted for various values of plate bias voltage, V_p (also called B+).

the operating region of the curve is specified, usually a straight portion of the curve if linear operation is desired. Two points along that portion of the curve are selected, each defined by its location along the x and y axes. If the two points are defined by (x_1,y_1) and (x_2,y_2), the slope, m, of the curve, which can be a gain, a resistance or a conductance, is calculated as:

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

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

BIASING

The operation of an analog signal-processing device is greatly affected by which portion of the characteristic curve is used to do the processing. As an example, consider the vacuum tube characteristic curves in **Fig 5.8** and **Fig 5.9**.

The relationship between the input and the output of a tube amplifier is illustrated in Fig 5.8. The input signal (a sine wave in this example) is plotted in the vertical direction and below the graph. For a grid bias level of –5 V, the sine wave causes the grid voltage, V_g, to deviate between –3 and –7 V. These values correspond to a range of plate currents, I_p, between 1.4 and 2.6 mA. With a plate bias of 200 V and a load resistance, R_p, of 50 kΩ, the corresponding change in plate voltage, V_p, is between 70 and 130 V. Thus, this triode amplifier configuration changes a range of 4 V at the input to 60 V at the output. Also there is a change of output-signal voltage polarity; this amplifier both amplifies the signal magnitude 15 times and shifts the phase of the signal by 180°.

In the previous example the signal was biased so that it fell on a linear (straight) portion of the characteristic curve. If a different bias voltage is selected so that the signal does not fall on a linear portion of the curve, the output signal will be a distorted version of the input signal. This is illustrated in Fig 5.9. The input signal is amplified within a curved region of the characteristic curve. The positive part of the signal is amplified more than the negative part of the signal. Proper biasing is crucial to ensure amplifier linearity.

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

Manufacturers' Data Sheets

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

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

Rather than plotting the characteristic curves for each device, the manufacturer often selects key operating parameters that describe the device operation for the

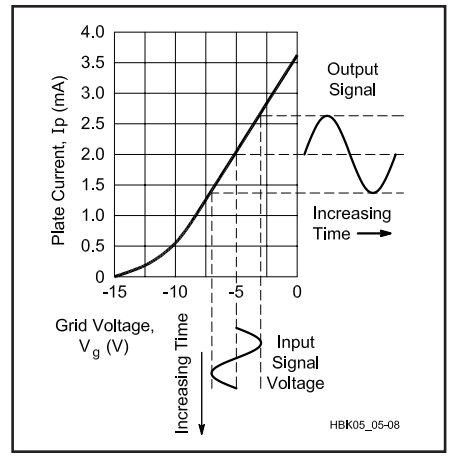


Fig 5.8 — Determination of output signal (to the right of the plot) for a given input signal (below the plot, turned on its side) with a tube characteristic curve plotted for a given plate bias. Note that the grid bias voltage, –5 V, causes the entire range of the input signal to be mapped onto the linear (diagonal straight line) portion of the characteristic curve. The output signal has the same shape as the input signal except that it is larger in amplitude.

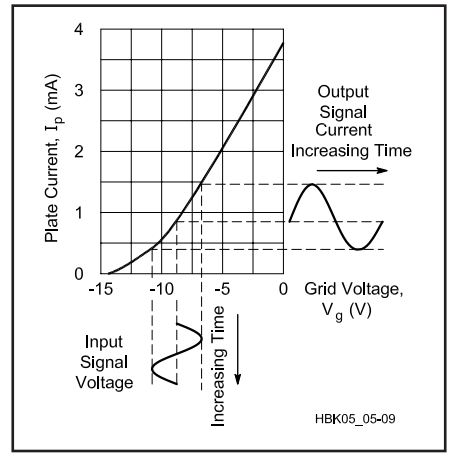


Fig 5.9 — Same characteristic curve and input signal as in Fig 5.8 except the grid bias voltage is now about –8.75 V. The input signal falls on the curved (non-linear) portion of the plot and causes distortion in the output signal. Note how the upper portion of the output sine wave was amplified more than the lower portion.

configurations and parameter ranges that are most commonly used. For example, a bipolar transistor data sheet might include an OPERATING PARAMETERS section. Parameters are listed in an OFF CHARACTERISTICS subsection and an ON CHARACTERISTICS subsection that describe the conduction properties of the device for dc voltages. The SMALL-SIGNAL CHARACTERISTICS section

often contains the guaranteed minimum Gain-Bandwidth Product (f_T), the guaranteed maximum output capacitance, the guaranteed maximum input capacitance and the guaranteed range of the transfer parameters applicable to a given device. Finally, the SWITCHING CHARACTERISTICS section lists absolute maximum ratings for Delay Time (t_d), Rise Time (t_r), Storage Time (t_s) and Fall Time (t_f). Other types of devices list characteristics important to operation of that specific device.

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

VACUUM TUBES

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

Thermionic Theory

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

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

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

Components of a Vacuum Tube

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

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

A vacuum tube that contains only a cathode and a plate is called a *diode tube* (di- for two components). See **Fig 5.10**. The diode tube is similar to a semiconductor diode since it allows current to pass in only one direction; it is used as a rectifier. When the plate voltage becomes

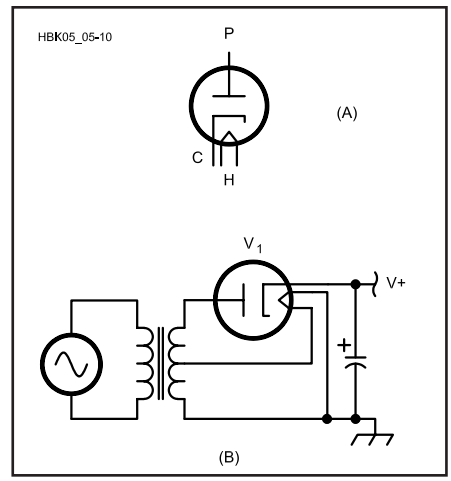


Fig 5.10 — Vacuum tube diode. (A) Schematic symbol detailing heater (H), cathode (C) and plate (P). (B) Power supply circuit using diode as a half wave rectifier.

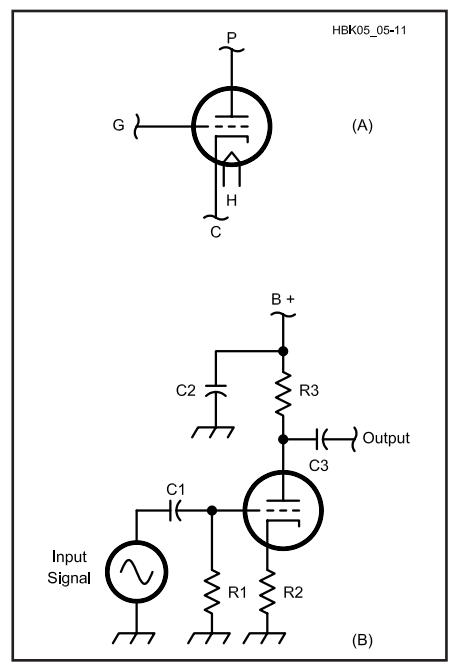


Fig 5.11 — Vacuum tube triode. (A) Schematic symbol detailing heater (H), cathode (C), grid (G) and plate (P). (B) Audio amplifier circuit using a triode. C1 and C3 are dc blocking capacitors for the input and output signals to isolate the grid and plate bias voltages. C2 is a bypass filter capacitor to decrease noise in the plate bias voltage, B+. R1 is the grid bias resistor, R2 is the cathode bias resistor and R3 is the plate bias resistor. Note that although the cathode and grid bias voltages are positive with respect to ground, they are still negative with respect to the plate.

negative, the electrical field that is set up repels electrons, preventing them from being emitted from the cathode.

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

Vacuum tubes containing a cathode, a grid and a plate are called *triode* tubes (tri- for three components). See **Fig 5.11**. They are generally used as amplifiers, particularly at frequencies in the HF range and below. Characteristic curves for triodes normally relate grid bias voltage and plate bias voltage to plate current for the triode (Fig 5.7). There are three descriptors of a tube's performance that can be derived from the characteristic curves. The *plate resistance*, r_p, describes the resistance to the flow of electrons from cathode to plate. The r_p is calculated by selecting a vertical line in the characteristic curve and dividing the change in plate-to-cathode voltage (ΔV_p) of two of the lines by the corresponding change in plate current (ΔI_p).

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

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

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

Triode amplification factors range from 10 to about 100.

The plate current flows to the plate bias supply, so the output from a triode amplifier is often expressed as the voltage that is developed as this current passes through a load resistor. The value of the load resistance affects the tube amplification, as illustrated by the dynamic characteristic curves in **Fig 5.12**, so the tube μ does not fully describe its action as an amplifier. *Grid-plate transconductance* (g_m) takes into account the change of amplification due to load resistance. The slope of the lines in the characteristic curve represents g_m. (Since the various lines are nearly parallel in the linear operating region, they have about the same slope.)

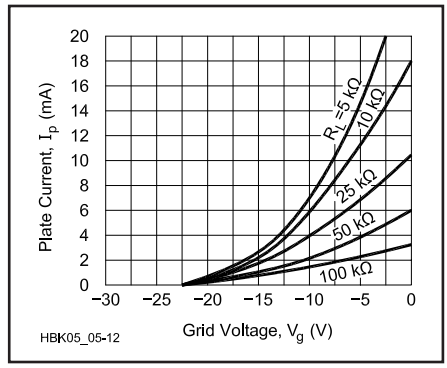


Fig 5.12 — Vacuum tube dynamic characteristic curve. This corresponds to the V_p = 300 line in Fig 5.7 with different values of load resistance. This shows how the tube will behave when cascaded to circuits with different input impedances.

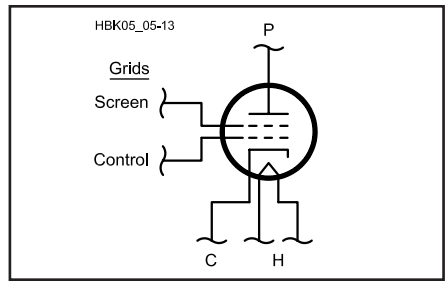


Fig 5.13 — Vacuum tube tetrode. Schematic symbol detailing heater (H), cathode (C), the two grids: control and screen, and the plate (P).

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

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

The input impedance of a vacuum tube amplifier is directly related to the grid current. Grid current varies with grid voltage, increasing as the voltage becomes more positive. The normal operation uses a negative grid-bias voltage, and the input impedance can be in the megohm range for very negative grid bias values. This is, however, limited by the desired operating point on the characteristic curve as illustrated in Figs 5.8 and 5.9. The output impedance of the amplifier is a function of the plate resistance, r_p, in parallel with the output capacitance. Typical output impedance is on the order of hundreds of ohms.

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

The grid-to-plate capacitance is the chief source of unwanted signal feedback. A special form of vacuum tube has been developed to deal with the grid-to-plate capacitance. A second grid, called a *screen grid*, is inserted between the original grid (now called a *control grid*) and the plate. The additional tube component leads to the name for this new tube — *tetrode* (tetra- for four components). See **Fig 5.13**. The screen grid reduces the capacitance between the control grid and the plate, but it also reduces the electrical field from the plate that attracts electrons from the cathode. Like the control grid, the screen grid is made of a wire mesh and electrons pass through the spaces between the wires to get to the plate. The bias of the screen grid is positive with respect to the cathode, in order to enhance the attraction of electrons from the cathode. The electrons accelerate toward the screen grid and most of them pass through the spaces and continue to accelerate until they reach the plate. The presence of the screen grid adversely affects the overall efficiency of the tube, since some of the electrons strike the grid wires. A bypass capacitor with a low reactance at the frequency being amplified by the vacuum tube is generally connected between the screen grid and the cathode.

A special form of tetrode concentrates the electrons flowing between the cathode and the plate into a tight beam. The decreased electron-beam area increases the

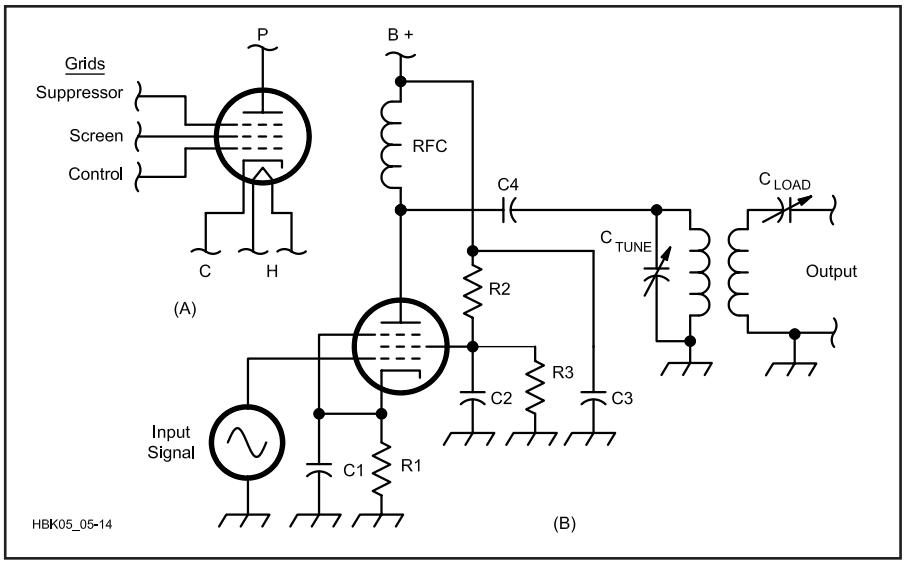


Fig 5.14 — Vacuum tube pentode. (A) Schematic symbol detailing heater (H), cathode (C), the three grids: control, screen and suppressor, and plate (P). (B) RF amplifier circuit using a pentode. C1, C2 and C3 are bypass (filter) capacitors and C4 is a dc blocking capacitor to isolate the plate voltage from the tank circuit components. R1 is the cathode bias resistor. R2 and R3 comprise a screen-dropping voltage divider. The plate tank circuit is tuned to the desired frequency bandpass. As is common, the heater circuit is not shown.

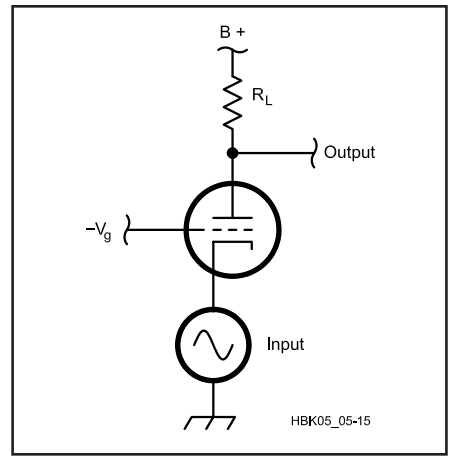


Fig 5.15 — Grounded grid amplifier schematic. The input signal is connected to the cathode, the grid is biased to the appropriate operating point by a dc bias voltage, $-V_G$, and the output voltage is obtained by the voltage drop through R_L that is developed by the plate current, I_P.

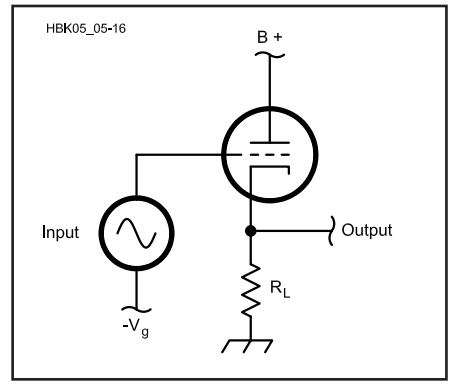


Fig 5.16 — Cathode follower schematic. The input signal is biased by $-V_G$ and fed into the grid. The plate bias, B+ is fed directly into the plate terminal. The output is derived by the cathode current (which is equal to the plate current, I_P) dropping the voltage through the load resistor, R_L.

efficiency of the tube. *Beam tetrodes* permit higher plate currents with lower plate voltages and large power outputs with smaller grid driving power. RF power amplifiers are usually made with this type of vacuum tube.

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

A third grid, called the *suppressor grid*, can be added between the screen grid and the plate. This overcomes the effects of secondary emission in tetrodes. A vacuum tube with three grids is called a *pentode* (penta for five components). See **Fig 5.14**. The suppressor grid is negatively biased with respect to the screen grid and the plate. In some tube designs it is internally connected to the cathode. The suppressor grid repels the secondary emission electrons back to the plate.

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

Types of Vacuum Tube Amplifiers

The descriptions of vacuum tube amplifiers up to this point have been for only one configuration, the common cathode, where the cathode is connected to the signal reference point, the grid is the input and the plate is the output. Although this is the most common configuration of the vacuum tube as an amplifier, other configurations exist. If the signal is introduced into the cathode and the grid is at a reference level (still negatively biased but with no ac component), with the output at the plate, the amplifier is called a *grounded-grid* (**Fig 5.15**). This amplifier is characterized by a very low input impedance, on the order of a few hundred ohms, and a low output impedance, that is mainly determined by the plate resistance of the tube.

The third configuration is called the *cathode follower* (**Fig 5.16**). The plate is the common element, the grid is the input and the cathode is the output. This type of amplifier is often used as a buffer stage due to its high input impedance, similar to that of the common cathode amplifier, and its very low output impedance. The output impedance (Z_O) can be calculated from the tube characteristics as:

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

where:

r_p = tube plate resistance

μ = tube amplification factor.

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

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

Other Types of Tubes

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

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

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

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

A variation of the vacuum tube that is widely used in oscilloscopes and television monitors is the *cathode ray tube (CRT)*, diagrammed in **Fig 5.17**. The CRT has a cathode and grid much like a triode tube. The plate, usually referred to as the *anode* in this device, is designed to accelerate the electrons to very high velocities, with anode voltages that can be as high as tens of thousands of volts. The anode of the CRT differs from the plates of other vacuum tubes, since it is designed as a set of plates that are parallel to the electron beam. The anode voltage accelerates the electrons but does not absorb them. The electron beam passes by the anode and continues to the face of the tube. The cathode, grid and anode are all located in the neck of the CRT and are collectively referred to as the *electron gun*.

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

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

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

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

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

Standard vacuum tubes work well for frequencies up to hundreds of megahertz. At frequencies higher than this, the

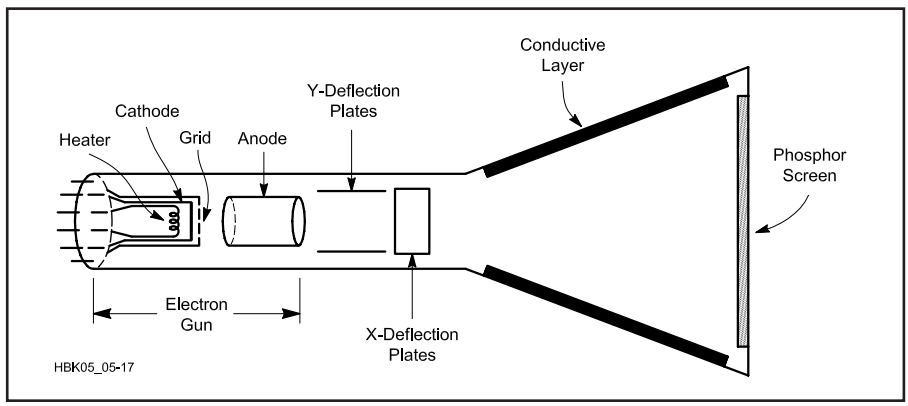


Fig 5.17 — Cross section of CRT. The electron gun generates a stream of electrons and is made up of a heater, cathode, grid and anode (plate). The electron beam passes by two pairs of deflection plates that deviate the path of the beam in the vertical (y) direction and then the horizontal (x) direction. The deflected electron beam strikes a phosphor screen and causes it to glow at that spot. Any electrons that bounce off the screen are absorbed by the conductive layer along the sides of the tube, preventing spurious luminescence.

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

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

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

PHYSICAL ELECTRONICS OF SEMICONDUCTORS

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

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

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

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

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

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

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

The conductivity of an extrinsic semiconductor depends on the charge density (in other words, the concentration of free electrons in N-type, and holes in P-type, semiconductor material). As the energy in the semiconductor increases, the charge density also increases. This is the basis of how all semiconductor devices operate: the major difference is the way in which the energy level is increased. Variations are: The *transistor*, where conductivity is

altered by injecting current into the device via a wire; the *thermistor*, where the level of heat in the device is detected by its conductivity, and the *photoconductor*, where light energy that is absorbed by the semiconductor material increases the conductivity.

The PN Semiconductor Junction

If a piece of N-type semiconductor material is placed against a piece of P-type semiconductor material, the location at which they join is called a *PN semiconductor junction*. The junction has characteristics that make it possible to develop diodes and transistors. The action of the junction is best described by a diode operating as a rectifier. Initially, when the two types of semiconductor material are placed in contact, each type of material will have only its majority carriers: P-type will have only holes and N-type will have only free electrons. The net positive charge of the P-type material attracts free electrons from across the junction and the opposite is true in the N-type material. These attractions lead to diffusion of some of the majority carriers across the junction, which neutralize the carriers immediately on the other side. The region close to the junction is then *depleted* of carriers, and, as such, is named the *depletion region* (or the *space-charge region* or the *transition region*). The width of the depletion region is very small, on the order of 0.5 μm.

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

Diodes are commonly made of silicon or germanium. Although they act similarly, they have slightly different characteristics. The *junction threshold voltage*, or *junction barrier voltage,* is the forward bias voltage at which current begins to pass through the device. This voltage is different for the two kinds of diodes. In the diode response curve of **Fig 5.18**, this value corresponds to the voltage at which the positive portion of the curve begins to rise sharply from the x axis. Most silicon diodes have a junction threshold voltage of about 0.7 V, while the value for germanium diodes typically is 0.3 V. The reverse biased leakage current is much lower for silicon diodes than for germanium diodes. The forward resistance of a diode is typically very low and varies with the amount of forward current.

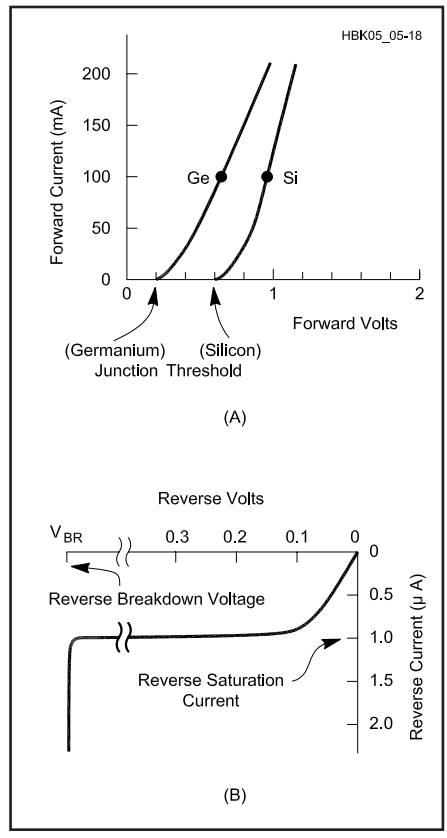


Fig 5.18 — Semiconductor diode (PN junction) response curve. (A) Forward biased (anode voltage higher than cathode) response for Germanium (Ge) and Silicon (Si) devices. Each curve breaks away from the x-axis at its junction threshold voltage. The slope of each curve is its forward resistance. (B) Reverse biased response. Very small reverse current increases until it reaches the reverse saturation current (I_0). The reverse current increases suddenly and drastically when the reverse voltage reaches the reverse breakdown voltage, V_{BR}.

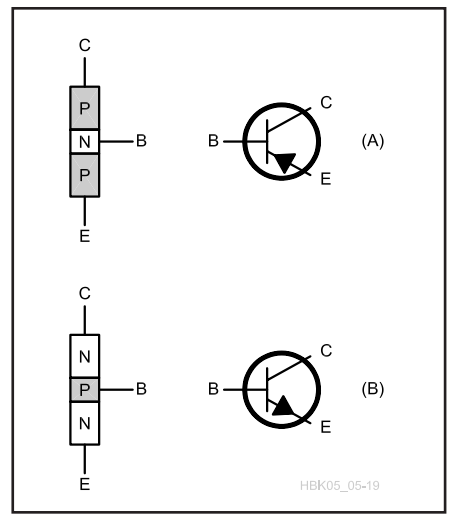


Fig 5.19 — Bipolar transistors. (A) A layer of N-type semiconductor sandwiched between two layers of P-type semiconductor makes a PNP device. The schematic symbol has three leads: collector (C), base (B) and emitter (E), with the arrow pointing in toward the base. (B) A layer of P-type semiconductor sandwiched between two layers of N-type semiconductor makes an NPN device. The schematic symbol has three leads: collector (C), base (B) and emitter (E), with the arrow pointing out away from the base.

Multiple Junctions

A bipolar transistor is formed when two PN junctions are placed next to each other. If N-type material is surrounded by P-type material, the result is a PNP transistor. Alternatively, if P-type material is in the middle of two layers of N-type material, the NPN transistor is formed (**Fig 5.19**).

Physically, we can think of the transistor as two PN junctions back-to-back, such as two diodes connected at their *anodes* (the positive terminal) for an NPN transistor or two diodes connected at their *cathodes* (the negative terminal) for a PNP transistor. The connection point is the base of the transistor. (You can't actually *make* a transistor this way.) A transistor conducts when the base-emitter junction is forward biased and the base-collector is reverse biased. Under these conditions, the emitter region emits majority carriers into the base region, where they are minority carriers because the materials of the emitter and base regions have opposite polarity. The excess minority carriers in the base are attracted across the base-collector junction, where they are collected and are once again considered majority carriers. The flow of majority carriers from emitter to collector can be modified by the application of a bias current to the base terminal. If the bias current has the same polarity as the base material (for example holes flowing into a P-type base) the emitter-collector current increases. A transistor allows a small base current to control a much larger collector current.

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

PNPN Diode

If four alternate layers of P-type and N-type material are placed together, a PNPN (usually pronounced like *pinpin*) diode with three junctions is obtained (see **Fig 5.20**). This device, when the anode is at a higher potential than the cathode, has its first and third junctions forward biased and its center junction reverse biased. In this state, there is little current, just as in the reverse biased diode. As the forward bias voltage is increased, the current through the device increases slowly until the *breakover (or firing) voltage*, V_{BO}, is reached and the flow of current abruptly increases. The PNPN diode is often considered to be a switch that is off below V_{BO} and on above it.

Bilateral Diode Switch

A semiconductor device similar to two PNPN diodes facing in opposite directions and attached in parallel is the *bilateral diode switch* or *diac*. This device has the characteristic curve of the PNPN diode for both positive and negative bias voltages. Its construction, schematic symbol and characteristic curve are shown in **Fig 5.21**.

Silicon Controlled Rectifier

Another device with four alternate layers of P-type and N-type semiconductor is the *silicon controlled rectifier (SCR)*, or *thyristor*. In addition to the connections to the outer two layers, two other terminals can be brought out for the inner two layers. The connection to the P-type material near the cathode is called the *cathode gate* and the N-type material near the anode is called the *anode gate*. In nearly all commercially available SCRs, only the cathode gate is connected (**Fig 5.22**).

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

Triac

A five-layered semiconductor whose operation is similar to a bidirectional SCR is the *triac* (**Fig 5.23**). This is also similar to a bidirectional diode switch with a bias control gate. The gate terminal of the triac can control both positive and negative breakover voltages and the devices can pass both polarities of voltage.

SCRs and triacs are often used to modify ac power sources. A sine wave with a given RMS value can be switched on and off at preset points during the cycle to decrease the RMS voltage. When conduction is delayed until after the peak (as **Fig 5.24** shows) the peak-to-peak voltage is reduced. If conduction starts before the

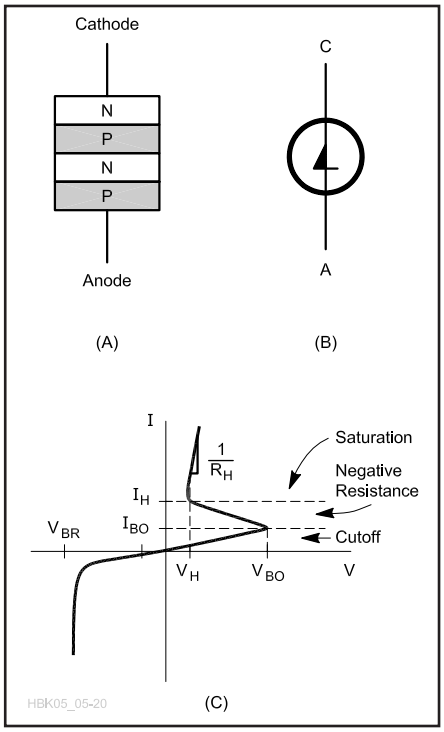


Fig 5.20 — PNPN diode. (A) Alternating layers of P-type and N-type semiconductor. (B) Schematic symbol with cathode (C) and anode (A) leads. (C) Voltage-current response curve. Reverse biased response is the same as normal PN junction diodes. Forward biased response acts as a hysteresis switch. Resistance is very high until the bias voltage reaches V_{BO} 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.

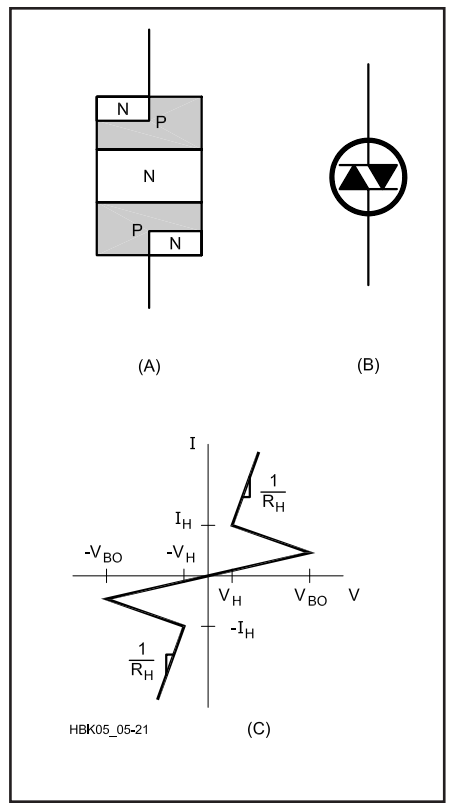


Fig 5.21 — Bilateral switch. (A) Alternating layers of P-type and N-type semiconductor. (B) Schematic symbol. (C) Voltage-current response curve. The right-hand side of the curve is identical to the PNPN diode response in Fig 5.20. The device responds identically for both forward and reverse bias so the left-hand side of the curve is symmetrical with the right-hand side.

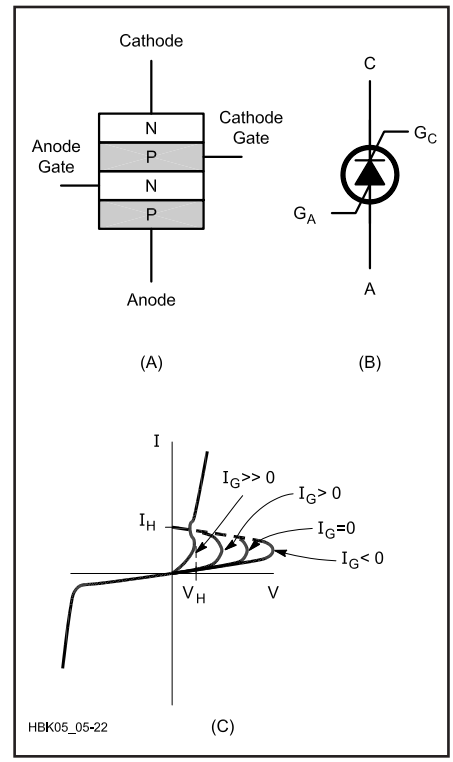


Fig 5.22 — SCR. (A) Alternating layers of P-type and N-type semiconductor. This is similar to a PNPN diode with gate terminals attached to the interior layers. (B) Schematic symbol with anode (A), cathode (C), anode gate (G_A) and cathode gate (G_C). Many devices are constructed without G_A. (C) Voltage-current response curve with different responses for various gate currents. I_G = 0 has the same response as the PNPN diode.

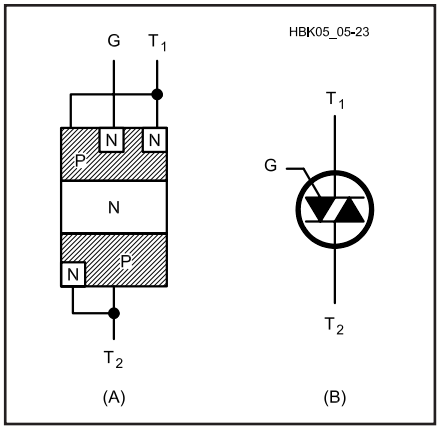


Fig 5.23 — Triac. (A) Alternating layers of P-type and N-type semiconductor. This behaves as two SCR devices facing in opposite directions with the anode of one connected to the cathode of the other and the cathode gates connected together. (B) Schematic symbol.

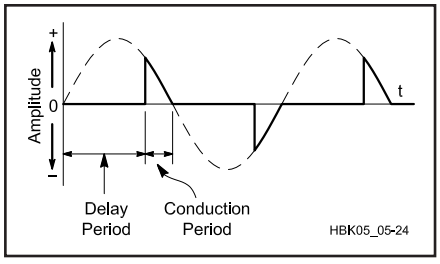


Fig 5.24 — Triac operation on sine wave. The dashed line is the original sine wave and the solid line is the portion that conducts through the triac. The relative delay and conduction period times are controlled by the amount or timing of gate current, I_G. The response of an SCR is the same as this for positive voltages (above the x-axis) and with no conduction for negative voltages.

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

FIELD-EFFECT TRANSISTORS

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

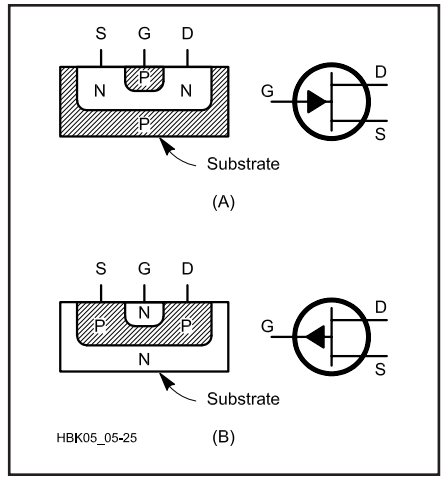


Fig 5.25 — JFET devices with terminals labeled: source (S), gate (G) and drain (D). A) Pictorial of N-type channel embedded in P-type substrate and schematic symbol. B) P-channel embedded in N-type substrate and schematic symbol.

Current moves within the FET in a channel, from the source connection to the drain connection. A gate terminal generates an electric field that controls the current (see **Fig 5.25**). The channel is made of either N-type or P-type semiconductor material; an FET is specified as either an N-channel or P-channel device. Majority carriers flow from source to drain. In N-channel devices, electrons flow so the drain potential must be higher than that of the source ($V_{DS} > 0$). In P-channel devices, the flow of holes requires that $V_{DS} < 0$. The polarity of the electric field that controls current in the channel is determined by the majority carriers of the channel, ordinarily positive for P-channel FETs and negative for N-channel FETs.

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

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

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

JFET

There are two basic types of FET. In the *junction FET (JFET)*, the gate material is made of the opposite polarity semiconductor to the channel material (for a P-channel FET the gate is made of N-type semiconductor material). The gate-channel junction is similar to a diode's PN junction. As with the diode, current is high if the junction is forward biased and is extremely small when the junction is reverse biased. The latter case is the way that JFETs are used, since any current in the gate is undesirable. The magnitude of the reverse bias at the junction is proportional to the size of the electric field that "pinches" the channel. Thus, the current in the channel is reduced for higher reverse gate bias voltage.

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

MOSFET

Placing an insulating layer between the gate and the channel allows for a wider

range of control (gate) voltages and further decreases the gate current (and thus increases the device input resistance). The insulator is typically made of an oxide (such as silicon dioxide, SiO_2). This type of device is called a *metal-oxide-semiconductor FET (MOSFET)* or *insulated-gate FET (IGFET)*. The substrate is often connected to the source internally. The insulated gate is on the opposite side of the channel from the substrate (see **Fig 5.26**). The bias voltage on the gate terminal either attracts or repels the majority carriers of the substrate across the PN junction with the channel. This narrows (depletes) or widens (enhances) the channel, respectively, as V_{GS} changes polarity. For N-channel MOSFETs, positive gate voltages with respect to the substrate and the source ($V_{GS} > 0$) repel holes from the channel into the substrate, thereby widening the channel and decreasing channel resistance. Conversely, $V_{GS} < 0$ causes holes to be attracted from the substrate, narrowing the channel and increasing the channel resistance. Once again, the polarities discussed in this example are reversed for P-channel devices. The common abbreviation for an N-channel MOSFET is *NMOS*, and for a P-channel MOSFET, *PMOS*.

Because of the insulating layer next to the gate, input resistance of a MOSFET is usually greater than 10^{12} Ω (a million megohms). Since MOSFETs can both deplete the channel, like the JFET, and also enhance it, the construction of MOSFET devices differs based on the channel size in the resting state, $V_{GS} = 0$. A *depletion mode* device (also called a *normally on MOSFET*) has a channel in resting state that gets smaller as a reverse bias is applied; this device conducts current with no bias applied (see Fig 5.26 A and B). An *enhancement mode* device (also called a *normally off MOSFET*) is built without a channel and does not conduct current when $V_{GS} = 0$; increasing forward bias forms a channel that conducts current (see Fig 5.26 C and D).

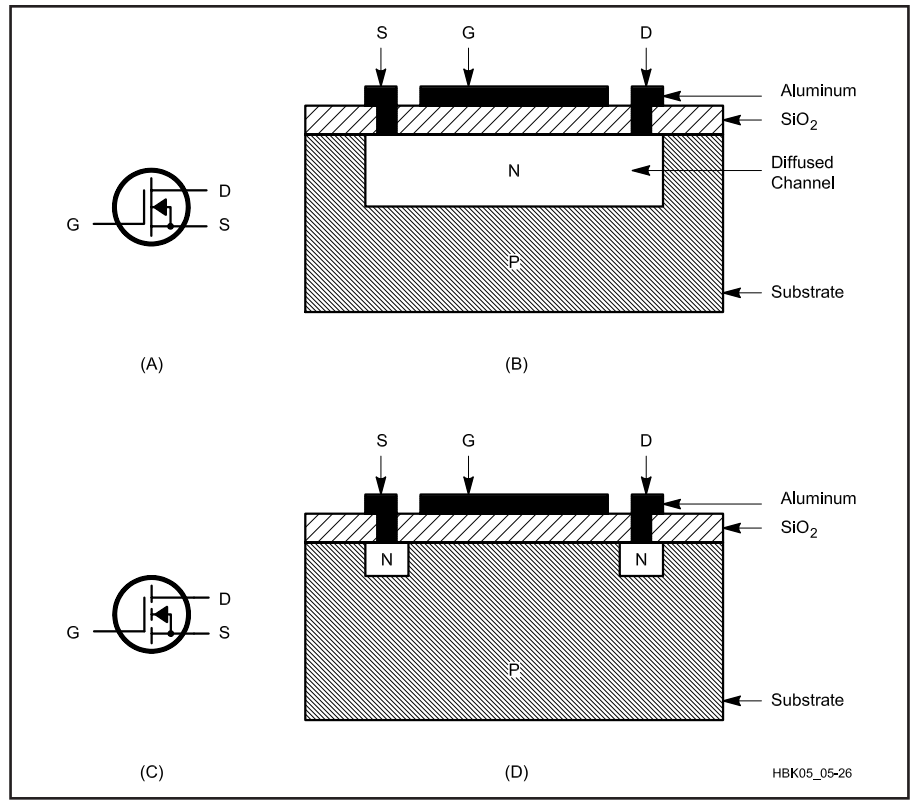


Fig 5.26 — MOSFET devices with terminals labeled: source (S), gate (G) and drain (D). N-channel devices are pictured. P-channel devices have the arrows reversed in the schematic symbols and the opposite type semiconductor material for each of the layers. (A) N-channel depletion mode device schematic symbol and (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.

Semiconductor Temperature Effects

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

Semiconductor Failure

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

Impurities are introduced into intrinsic semiconductors by diffusion, the same physical property that lets you smell cookies baking from several rooms away. Smells diffuse through air much faster than molecules diffuse through solids. Once the impurities diffuse into the semiconductor, they tend to stay in place. Rates of diffusion are proportional to temperature, and semiconductors are doped with

impurities at high temperature to save time. Once the doped semiconductor material is cooled, the rate of diffusion of the impurities is so low that they are essentially immobile for many years to come.

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

Thermistors

A *thermistor* is an intrinsic (no N or P doping) semiconductor metal-oxide compound device that has a large negative temperature coefficient (NTC) of resistance. Thermally generated free electrons and holes become available as current carriers in thermistors. Metal oxides, such as nickel-oxide (NiO), dimanganese-trioxide (Mn_2O_3) and cobalt-trioxide (Co_2O_3), are chosen for their stable electrical properties. Silicon and Germanium are not used as thermistors because their temperature properties are very sensitive to impurities.

A related device, the *sensistor*, uses large amounts of doping materials to achieve a large positive temperature coefficient (PTC) of resistance, usually over some restricted temperature range.

Practical Semiconductors

SEMICONDUCTOR DIODES

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

The diode symbol is shown in **Fig 5.27**. Current passes most easily from anode to cathode, in the direction of the arrow. This is often referred to as the *forward* direction and the opposite is the *reverse* direction. Remember that *current* refers to the flow of electricity from higher to lower potentials and is in the opposite direction to the flow of electrons (current moves from anode to cathode and electrons flow from cathode to anode, as based on the definitions of the words, *anode* and *cathode*). The anode of a semiconductor junction diode is made of P-type material and the cathode is made of N-type material, as indicated in Fig 5.27. Most diodes are marked with a band on the cathode end (Fig 5.27). The ideal diode would have zero resistance in the forward direction and infinite resistance in the reverse direction. This is not the case for actual devices, which behave as shown in the plot of a diode response in Fig 5.18. Note that the scales of the two parts of the graph are drastically different. The inverse of the slope of the line (the change in voltage between two points on a straight portion of the line divided by the corresponding change in current) on the upper right is the resistance of the diode in the forward direction. The range of voltages is small and the range of currents is large since the forward resistance is very small (in this example, about 2 Ω). The lower left portion of the curve illustrates a much higher resistance that increases from tens of kilohms to thousands of megohms as the reverse voltage gets larger, and then decreases to near zero (a nearly vertical line) very suddenly at the peak inverse voltage (PIV = 100 V in this example).

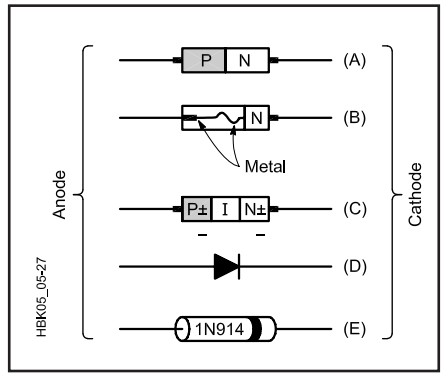


Fig 5.27 — Practical semiconductor diodes. All devices are aligned with anode on the left and cathode on the right. (A) Standard PN junction diode. (B) Point-contact or "cat's whisker" diode. (C) PIN diode formed with heavily doped P-type (P^+), undoped (intrinsic) and heavily doped N-type (N^+) semiconductor material. (D) Diode schematic symbol. (E) Diode package with marking stripe on the cathode end.

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

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

Fast Diodes

The speed of a diode affects the frequencies on which it can act. The diode response in Fig 5.18 is a steady state response, showing how that diode will act at dc. As the frequency increases, the diode may not be able to keep up with the changing polarity of the signal and its response will not be as expected. Diode speed mainly depends on charge storage in the depletion region. Under reverse bias, excess charges move away from the junction, forming a larger space-charge region that is the equivalent of a dielectric. The diode thus exhibits capacitance, which is inversely proportional to the width of the dielectric and directly proportional to the cross-sectional surface area of the junction.

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

The PIN diode, shown in Fig 5.27C is a *slow response* diode that is capable of passing microwave signals when it is forward biased. This device is constructed with a layer of intrinsic (undoped) semi-

conductor placed between very highly doped P-type and N-type material (called P^+-type and N^+-type material to indicate the high level of doping), creating a PIN junction. These devices provide very effective switches for RF signals and are often used in TR switches in transceivers. PIN diodes have longer than normal carrier lifetimes, resulting in a slow switching process that causes them to act more like resistors than diodes at high radio frequencies.

Varactors

If the PN junction capacitance is controlled rather than reduced, a diode can be made to act as a variable capacitor. As the reverse bias voltage on a diode increases, the width of the junction increases, which decreases its capacitance. A *varactor* is a diode whose junction is specially formulated to have a relatively large range of capacitance values for a modest range of reverse bias voltages (**Fig 5.28**). Although special forms of varactors are available from manufacturers, other types of diodes may be used as inexpensive varactor diodes, but the relationship between reverse voltage and capacitance is not always reliable. When designing with varactor diodes, the reverse bias voltage must be absolutely free of noise since any variations in the bias voltage will cause changes in capacitance. Unwanted frequency shifts or instability will result if the reverse bias voltage is noisy. It is possible to frequency modulate a signal by adding the audio signal to the reverse bias on a varactor diode used in the carrier oscillator.

Zener Diodes

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

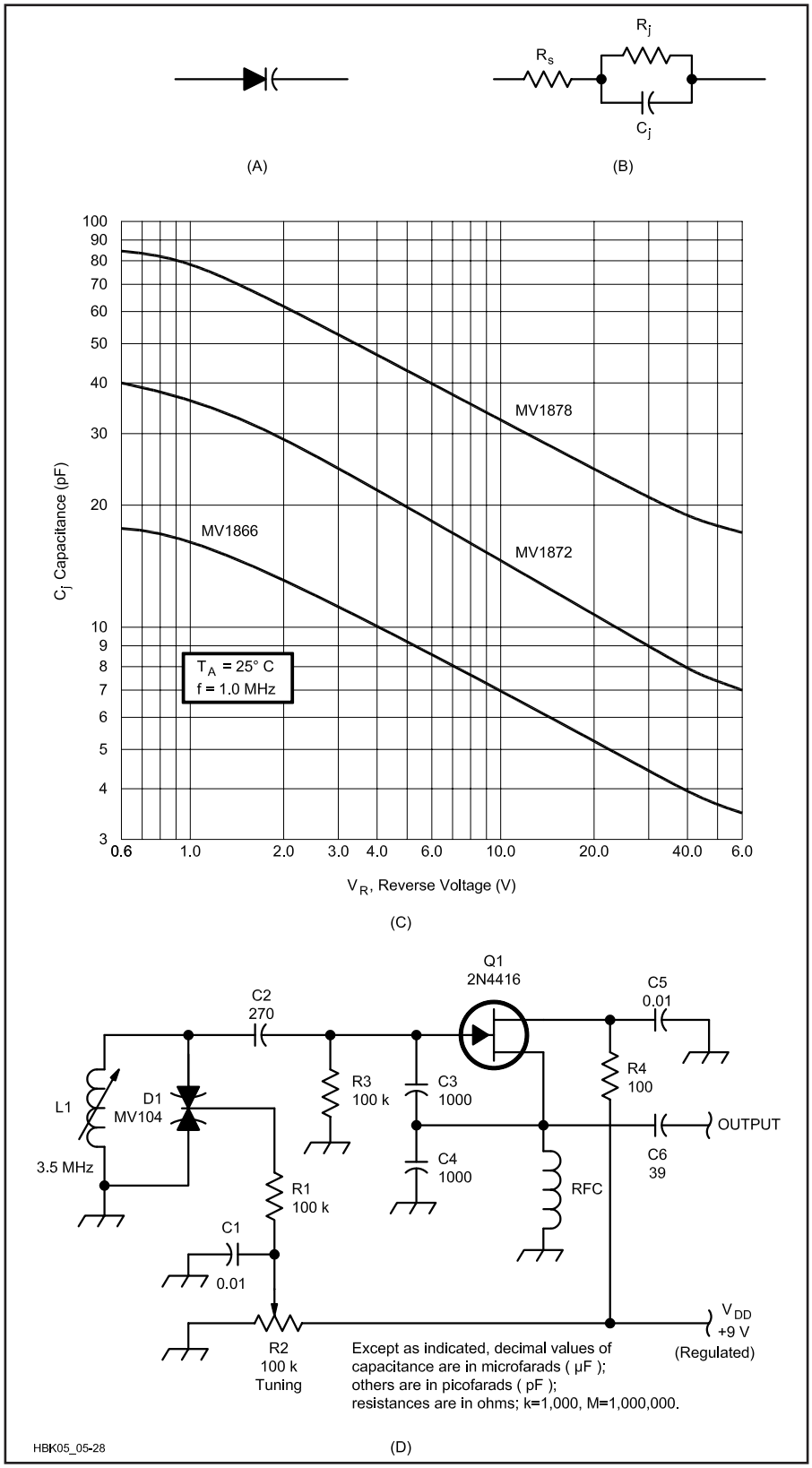


Fig 5.28 — Varactor diode. (A) Schematic symbol. (B) Equivalent circuit of the reverse biased varactor diode. R_S is the junction resistance, R_J is the leakage resistance and C_J is the junction capacitance, which is a function of the magnitude of the reverse bias voltage. (C) Plot of junction capacitance, C_J, as a function of reverse voltage, V_R, for three different varactor devices. Both axes are plotted on a logarithmic scale. (D) Oscillator circuit with varactor tuning. D1-L1 is a tuned circuit with a dual varactor diode that is controlled by the voltage from potentiometer R2. C1 is a filter capacitor to ensure that the varactor bias voltage is clean dc. C2 and C6 are dc blocking capacitors. Q1 is an N-channel JFET in common drain configuration with feedback to the gate through C3. R3 is the gate bias resistor. R4 is the drain voltage resistor with filter capacitor C5.

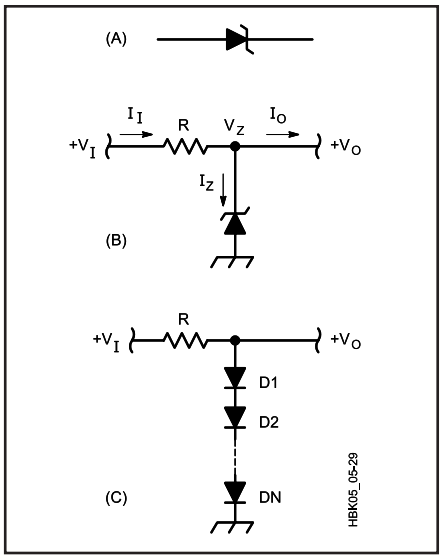


Fig 5.29 — Zener diode. (A) Schematic symbol. (B) Basic voltage regulating circuit. V_Z is the Zener reverse breakdown voltage. The Zener diode draws more current until $V_I - I_IR = V_Z$. The circuit design should select R so that when the maximum current is drawn, $R < (V_I - V_Z) / I_O$. The diode should be capable of passing the same current when there is no output current drawn. (C) For small voltages, several forward biased diodes can be used in place of Zener diodes. Each diode will drop the voltage by about 0.7 V for silicon or 0.3 V for germanium.

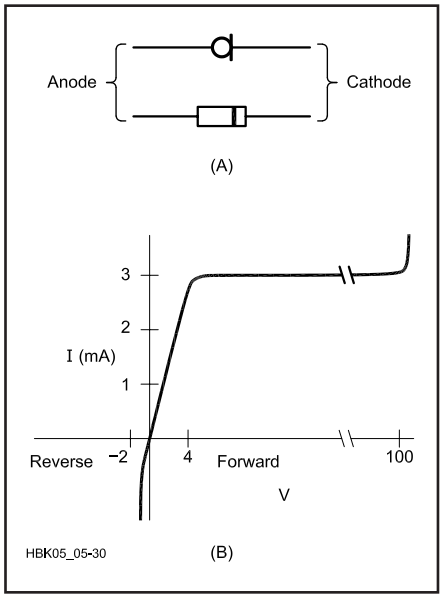


Fig 5.30 — Current regulator diode. (A) Schematic symbol and package with line marking cathode end. (B) Diode characteristic curve (1N5283 device). When forward bias voltage exceeds about 4 V the current passing through the device is held constant regardless of the voltage across the device.

are often used for this purpose in power supply regulators.

When the reverse breakdown voltage is exceeded, the reverse voltage drop across the Zener diode remains constant. With an appropriate current limiting resistor in series with it, the Zener diode provides an accurate voltage reference (**Fig 5.29**). Zener diodes are rated by their reverse-breakdown voltage and their power-handling capacity. The power is a product of the current passing through the reverse-biased Zener diode "in breakdown" (that is, in the breakdown mode of operation) and the breakdown voltage. Since the same current must always pass though the resistor to drop the source voltage down to the reference voltage, with that current divided between the Zener diode and the load, this type of power source is very wasteful of current. The Zener diode does make an excellent and efficient voltage reference in a larger voltage regulating circuit where the load current is provided from another device whose voltage is set by the reference. (See the **Power Supplies** chapter for more information about using Zener diodes as voltage regulators.) The major sources of error in Zener-diode derived voltages are the variation with load current and the variation due to heat. Temperature compensated Zener diodes are available with temperature coefficients as low as 0.0005 % / °C. If this is unacceptable, voltage reference integrated circuits based on Zener diodes have been developed that include additional circuitry to counteract temperature effects.

Constant Current Diodes

A form of diode, called a *field-effect regulator diode*, provides a constant current over a wide range of forward biased voltages. The schematic symbol and characteristic curve for this type of device are shown in **Fig 5.30**. Constant current diodes are very useful in any application where a constant current is desired. Some part numbers are 1N5283 through 1N5314.

Common Diode Applications

Standard semiconductor diodes have many uses in analog circuitry. Several examples of diode circuits are shown in **Fig 5.31**. Rectification has already been described. There are three basic forms of rectification using semiconductor diodes: half wave (1 diode), full-wave center-tapped (2 diodes) and full-wave bridge (4 diodes). These are more fully described in the **Power Supplies** chapter.

Diodes are commonly used to protect circuits. In battery powered devices a forward biased series diode is often used to protect the circuitry from the user inadvertently inserting the batteries backwards. Likewise, when a circuit is powered from an external dc source, a diode is often placed in series with the power connector in the device to prevent incorrectly wired power supplies from destroying the equipment. Diodes are commonly used to protect analog meters from both reverse voltage and over voltage conditions that would destroy the delicate needle movement.

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

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

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

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

Diodes are used as switches for ac coupled signals when a dc bias voltage can be added to the signal to permit or inhibit

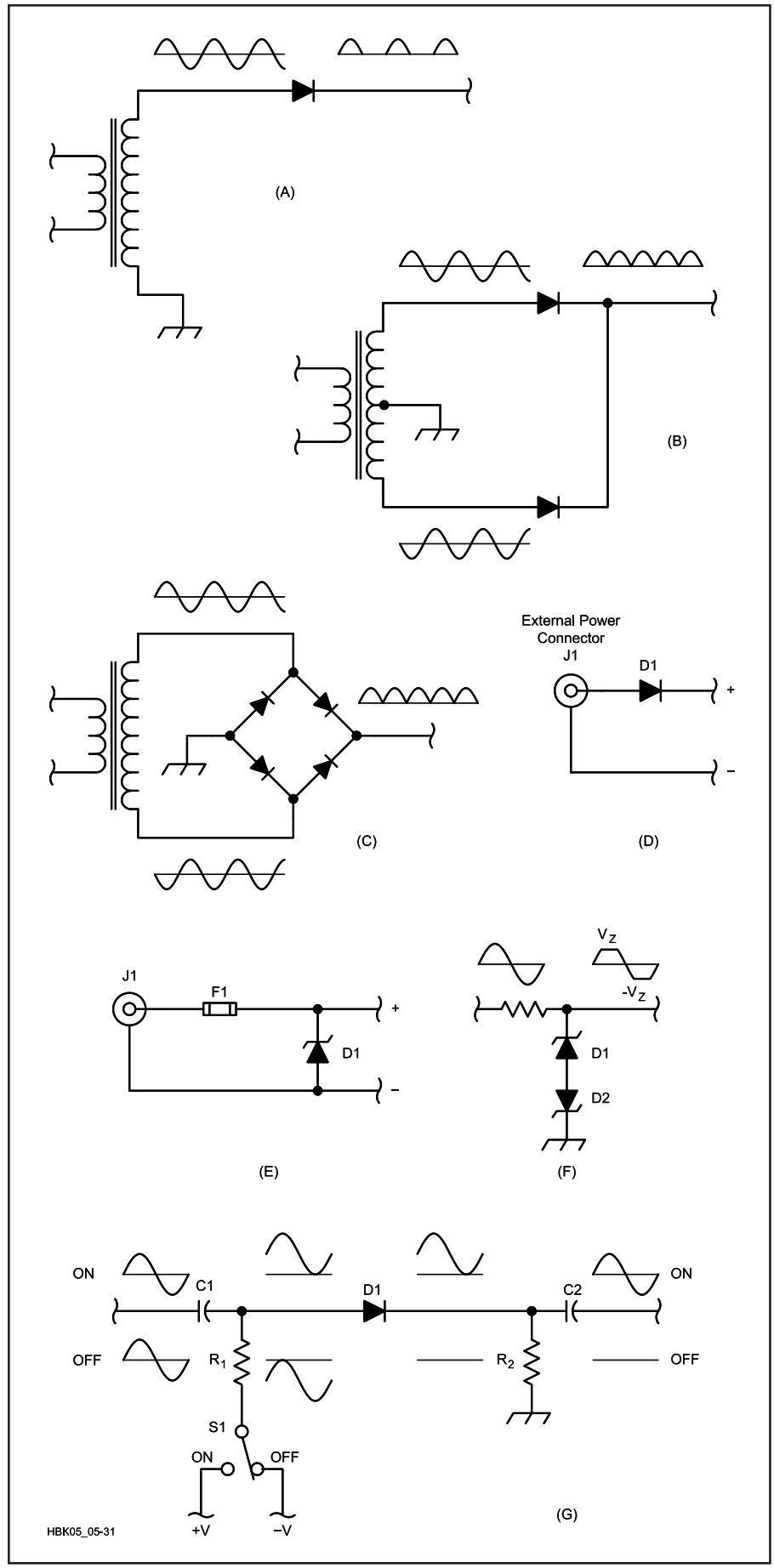


Fig 5.31 — Diode circuits. (A) Half wave rectifier circuit. Only when the ac voltage is positive does current pass through the diode. Current flows only during half of the cycle. (B) Full-wave center-tapped rectifier circuit. Center-tap on the transformer secondary is grounded and the two ends of the secondary are 180° out of phase. During the first half of the cycle the upper diode conducts and during the second half of the cycle the lower diode conducts. There is conduction during the full cycle with only positive voltages appearing at the output. (C) Full-wave bridge rectifier circuit. In each half of the cycle two diodes conduct. (D) Polarity protection for external power connection. J1 is the connector that power is applied to. If polarity is correct, the diode will conduct and if reversed the diode will block current, protecting the circuit that is being powered. (E) Over-voltage protection circuit. If excessive voltage is applied to J1, D1 will conduct current until fuse, F1, is blown. (F) Bipolar voltage clipping circuit. In the positive portion of the cycle, D2 is forward biased, but no current is shunted to ground because D1 is reverse biased. D1 starts to conduct when the voltage exceeds the Zener breakdown voltage, and the positive peak is clipped. When the negative portion of the cycle is reached, D1 is forward biased, but no current is shunted to ground because D2 is reverse biased. When the voltage exceeds the Zener breakdown voltage of D2, it also begins to conduct, and the negative peak is clipped. (G) Diode switch. The signal is ac coupled to the diode by C1 at the input and C2 at the output. R2 provides a reference for the bias voltage. When switch S1 is in the ON position, a positive dc voltage is added to the signal so it is forward biased and is passed through the diode. When S1 is in the OFF position, the negative dc voltage added to the signal reverse biases the diode, and the signal does not get through.

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

BIPOLAR TRANSISTORS

The bipolar transistor is a *current-controlled device*. The current between the emitter and the collector is governed by the current that enters the base. The convention when discussing transistor opera-

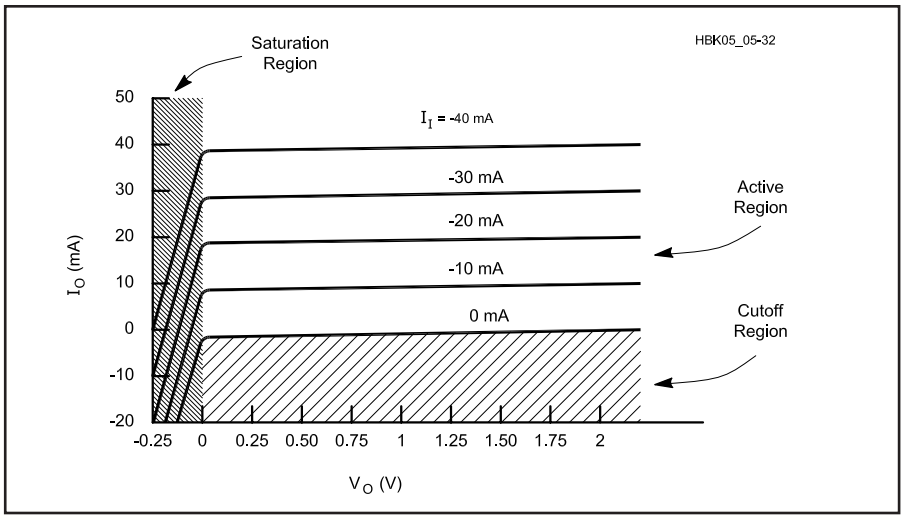


Fig 5.32 — Transistor response curve. The x-axis is the output voltage, and the y-axis is the output current. Different curves are plotted for various values of input current. The three regions of the transistor are its cutoff region, where no current flows in any terminal, its active region, where the output current is nearly independent of the output voltage and there is a linear relationship between the input current and the output current, and the saturation region, where the output current has large changes for small changes in output voltage.

tion is that the three currents into the device are positive (I_c into the collector, I_b into the base and I_e into the emitter). Kirchhoff's current law applies to transistors just as it does to passive electrical networks: the total current entering the device must be zero. Thus, the relationship between the currents into a transistor can be generalized as

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

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

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

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

Current conduction between collector and emitter is described by regions in the common-base response curves of the transistor device (see **Fig 5.32**). The transistor is in its *active 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. The slight slope that does exist is due to base-width modulation (known as the "Early effect"). Under these conditions, there is a linear relationship between the input current (I_I) and I_O. When both the junctions in the transistor are forward biased, the transistor is said to be in its *saturation region*. In this region, V_O is nearly zero and large changes in I_O occur for very small changes in V_O. The *cutoff region* occurs when both junctions in the transistor are reverse biased. Under this condition, there is very little current in the output, only the nanoamperes or microamperes that result from the very small leakage across the input-to-output junction. These descriptions of junction conditions are the basis for the use of transistors. Various configurations of the transistor in circuitry make use of the properties of the junctions to serve different purposes in analog signal processing.

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

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

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

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

and the relationship between α and β is defined as

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

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

The behavior of a transistor can be defined in many ways, depending on which type of amplifier it is wired to be. A complete description of a transistor must include characteristic curves for each configuration. Typically, two sets of characteristic curves are presented: one describing the input behavior and the other describing the output behavior in each amplifier configuration. Different transistor amplifier configurations have different gains, input and output impedances. At low frequencies, where parasitic capacitances aren't a factor, the common emitter configuration has a high current gain (about – 50, with the negative sign indicating a 180° phase shift), medium to high input impedance (about 50 kΩ) and a medium to low output impedance (about 1 kΩ). The common collector has a high current gain (about 50), a high input impedance (about 150 kΩ) and a low output impedance (about 80 Ω). The common base amplifier has a low current gain (about 1), a low input impedance (about 25 Ω) and a very high output impedance (about 2 MΩ). Depending on the intended use of the transistor amplifier in an analog circuit, one configuration will be more appropriate than others. Once the common lead of the transistor amplifier configuration is chosen, the input and output impedance are functions of the device bias levels and circuit loading (**Fig 5.33**). The actual input and output impedances of a transistor amplifier are highly dependent on the input, biasing and load resistors that are used in the circuit.

A typical general-purpose bipolar-transistor data sheet lists important device specifications. Parameters listed in the ABSOLUTE MAXIMUM RATINGS section are the three junction voltages (V_{CEO}, V_{CBO} and V_{EBO}), the continuous collector current (I_C), the total device power dissipation (P_D) and the operating and storage temperature range. In the OPERATING PARAMETERS section, the three guaranteed minimum junction breakdown voltages

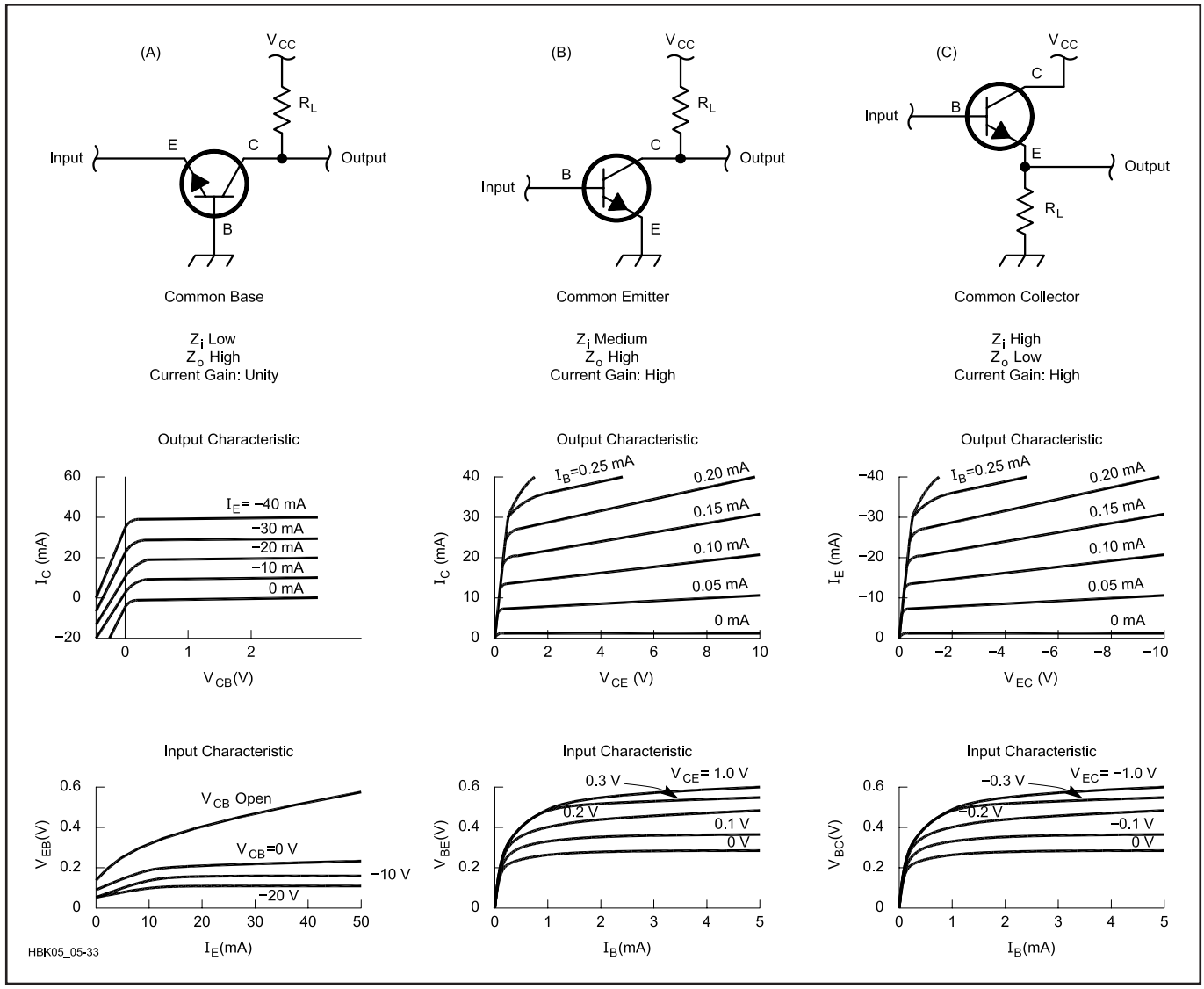


Fig 5.33 — The three configurations of transistor amplifiers. Each has a table of its relative impedance and current gain. The output characteristic curve is plotted for each, with the output voltage along the x-axis, the output current along the y-axis and various curves plotted for different values of input current. The input characteristic curve is plotted for each configuration with input current along the x-axis, input voltage along the y-axis and various curves plotted for different values of output voltage. (A) Common base configuration with input terminal at the emitter and output terminal at the collector. (B) Common emitter configuration with input terminal at the base and output terminal at the collector. (C) Common collector with input terminal at the base and output terminal at the emitter.

are listed $V_{(BR)CEO}$, $V_{(BR)CBO}$ and $V_{(BR)EBO}$ — along with the two guaranteed maximum collector cutoff currents — I_{CEO} and I_{CBO} — under OFF CHARACTERISTICS. Under ON CHARACTERISTICS are the guaranteed minimum dc current gain (h_{FE}), guaranteed maximum collector-emitter saturation voltage — $V_{CE(SAT)}$ — and the guaranteed maximum base-emitter on voltage — $V_{BE(ON)}$. The next section is SMALL-SIGNAL CHARACTERISTICS, where the guaranteed minimum current gain-bandwidth product — f_T, the guaranteed maximum output capacitance — C_{obo}, the guaranteed maximum input capacitance — C_{ibo}, the guaranteed range of input impedance — h_{ie}, the small-signal current gain — h_{fe}, the guaranteed maximum voltage feedback ratio — h_{re} and output admittance — h_{oe} are listed. Finally, the SWITCHING CHARACTERISTICS section lists absolute maximum ratings for delay time — t_d, rise time — t_r, storage time — t_s and fall time — t_f.

Transistor Biasing

Biasing in a transistor adds or subtracts a fixed amount of current from the signal at the input port. This differs from vacuum tube, FET and operational amplifier biasing where a bias *voltage* is added to the input signal. Fixed bias is the simplest form, as shown in **Fig 5.34A**. The operating point is determined by the intersection between the characteristic curves, the load line and the quiescent current bias line (Fig 5.34B). The problem with fixing the bias current is that if the transistor parameters drift due to heat, the operating point will change. The operating point can be stabilized by self biasing, also called emitter biasing, as pictured in Fig 5.34C. If I_C increases due to temperature changes, the current in R_E increases. The larger current through R_E increases the voltage drop across that resistor, causing a decrease in the base current, I_B. This, in turn, leads to a decreasing I_C, minimizing its variation

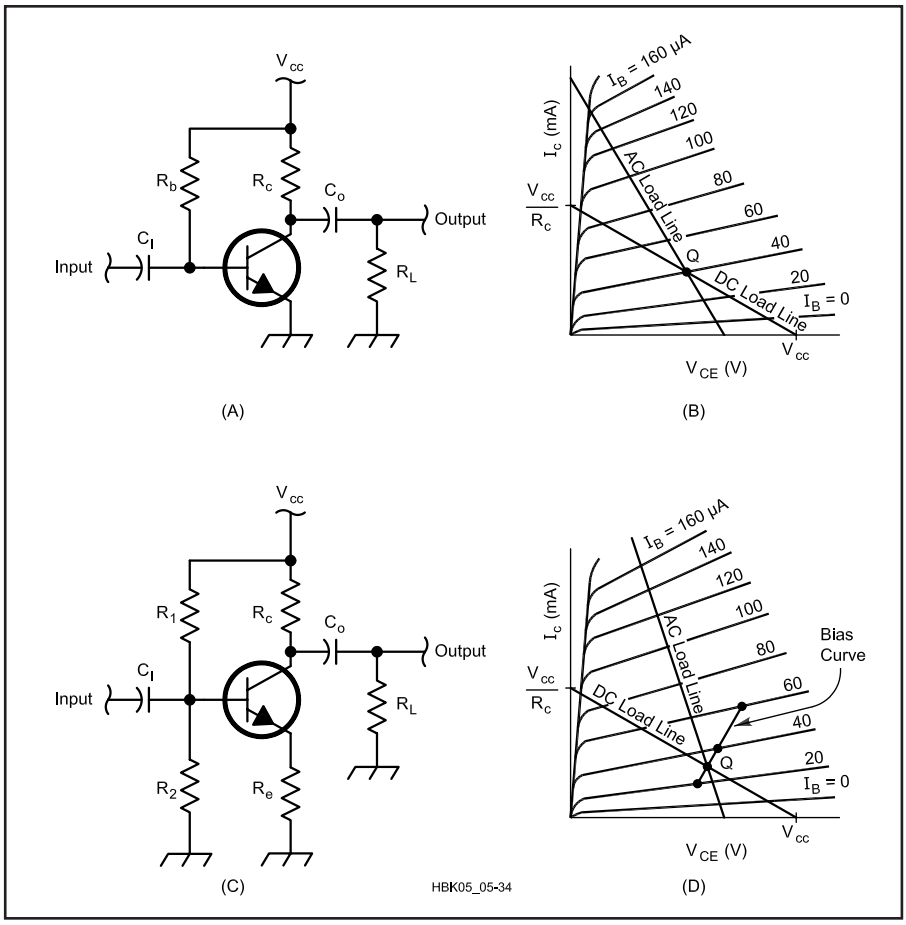


Fig 5.34 — Transistor biasing circuits. (A) Fixed bias. Input signal is ac coupled through C_i. The output has a voltage that is equal to $V_{CC} - I_C \times R_C$. This signal is ac coupled to the load, R_L, through C_O. For dc signals, the entire output voltage is based on the value of R_C. For ac signals, the output voltage is based on the value of R_C in parallel with R_L. (B) Characteristic curve for the transistor amplifier pictured in (A). The slope of the dc load line is equal to $-1 / R_C$. For ac signals, the slope of the ac load line is equal to $-1 / (R_C \parallel R_L)$. The quiescent operating point, Q, is based on the base bias current with no input signal applied and 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.

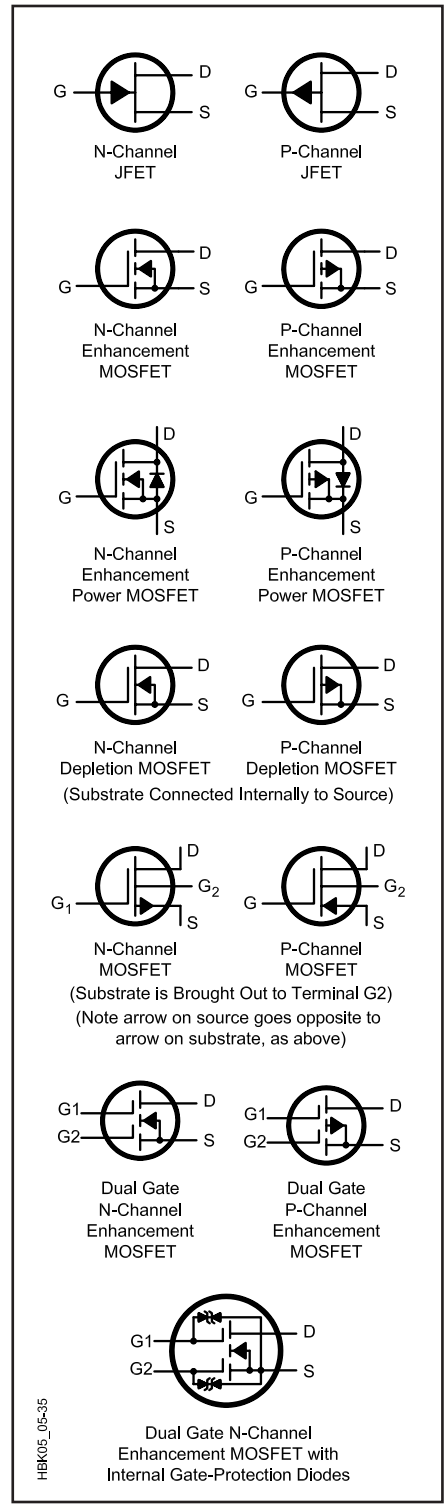


Fig 5.35—FET schematic symbols.

due to heat. The operating point for this type of biasing is plotted in Fig 5.34D.

FIELD-EFFECT TRANSISTORS

FET devices are more closely related to vacuum tubes than are bipolar transistors. Both the vacuum tube and the FET are controlled by the voltage level of the input rather than the input current, as in the bipolar transistor. FETs have three basic terminals, the gate, the source and the drain. These are related to both vacuum tube and bipolar transistor terminals: the gate to the grid and the base, the source to the cathode and the emitter, and the drain to the plate and the collector. Different forms of FET devices are pictured in **Fig 5.35**.

The characteristic curves for FETs are similar to those of vacuum tubes. The two most useful relationships are called the transconductance and output curves (**Fig 5.36**). Transconductance curves give the drain current, I_D, due to different gate-source voltage differences, V_{GS}, for various drain-source voltages, V_{DS}. The same parameters are interrelated in a different way in the output curve. For different values of V_{GS}, I_D is plotted against V_{DS}. In both of these representations, the device output is the drain current and these curves describe the FET in the common-source configuration. The action of the FET channel is so nearly ideal that, as long as the JFET gate does not become forward biased, the drain and source currents are virtually identical. For JFETs the gate leakage current, I_G, is a function of V_{GS} and this is often expressed with an input curve (**Fig 5.37**). The point at which there

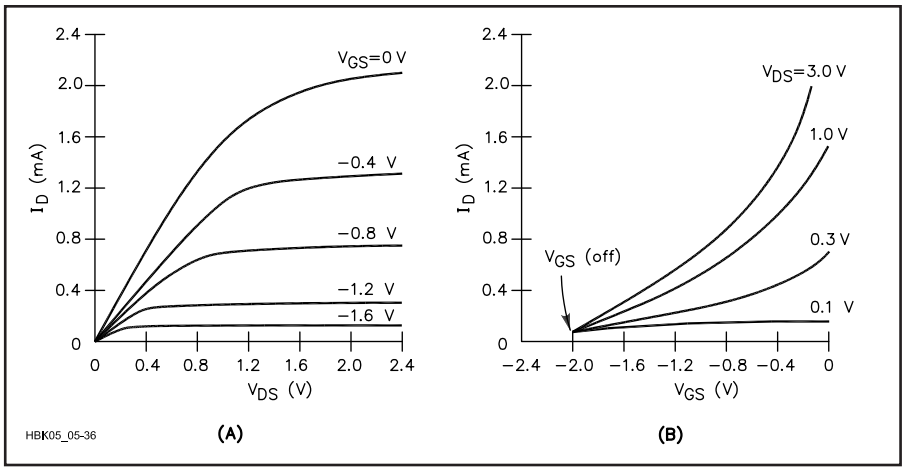


Fig 5.36 — JFET output and transconductance response curves for common source amplifier configuration. (A) Output voltage (V_{DS}) on the x-axis versus output current (I_D) on the y-axis, with different curves plotted for various values of input voltage (V_{GS}). (B) Transconductance curve has the same three variables rearranged, V_{GS} on the x-axis, I_D on the y-axis and curves plotted for different values of V_{DS}.

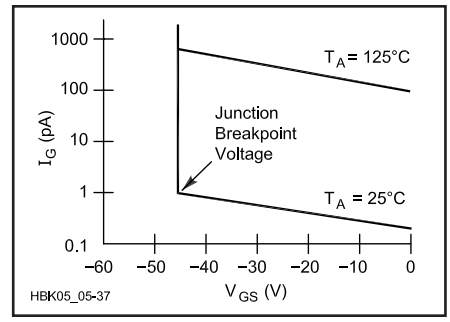


Fig 5.37 — JFET input leakage curves for common source amplifier configuration. Input voltage (V_{GS}) on the x-axis versus input current (I_G) on the y-axis, with two curves plotted for different operating temperatures, 25°C and 125°C. Input current increases greatly when the gate voltage exceeds the junction breakpoint voltage.

is a great increase in I_G is called the *junction break point voltage*. The insulated gates in MOSFET devices do away with any appreciable gate leakage current. MOSFETs do not need input and reverse transconductance curves. Their output curves (**Fig 5.38**) are similar to those of the JFET.

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

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

When 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.

A typical FET data sheet gives ABSOLUTE MAXIMUM RATINGS for V_{DS}, V_{DG}, V_{GS} and I_D, along with the usual device dissipation (P_D) and storage temperature range. The OFF CHARACTERISTICS listed are the gate-source breakdown voltage, $V_{GS(BR)}$, the reverse gate current, I_{GSS} and the gate-source cutoff voltage, $V_{GS(OFF)}$. The ON CHARACTERISTIC is the zero-gate-voltage drain current (I_{DSS}). The SMALL SIGNAL CHARACTERISTICS include the forward transfer admittance, y_{fs}, the output admittance, y_{os}, the static drain-source on resistance, $r_{ds(on)}$ and various capacitances such as input capacitance, C_{iss}, reverse transfer capacitance, C_{rss}, the drain-substrate capacitance, $C_{d(sub)}$. FUNCTIONAL CHARACTERISTICS include the noise figure, NF and the common source power gain G_{ps}.

The relatively flat regions in the MOSFET output curves are often used to provide a constant current source. As is plotted in these curves, the drain current,

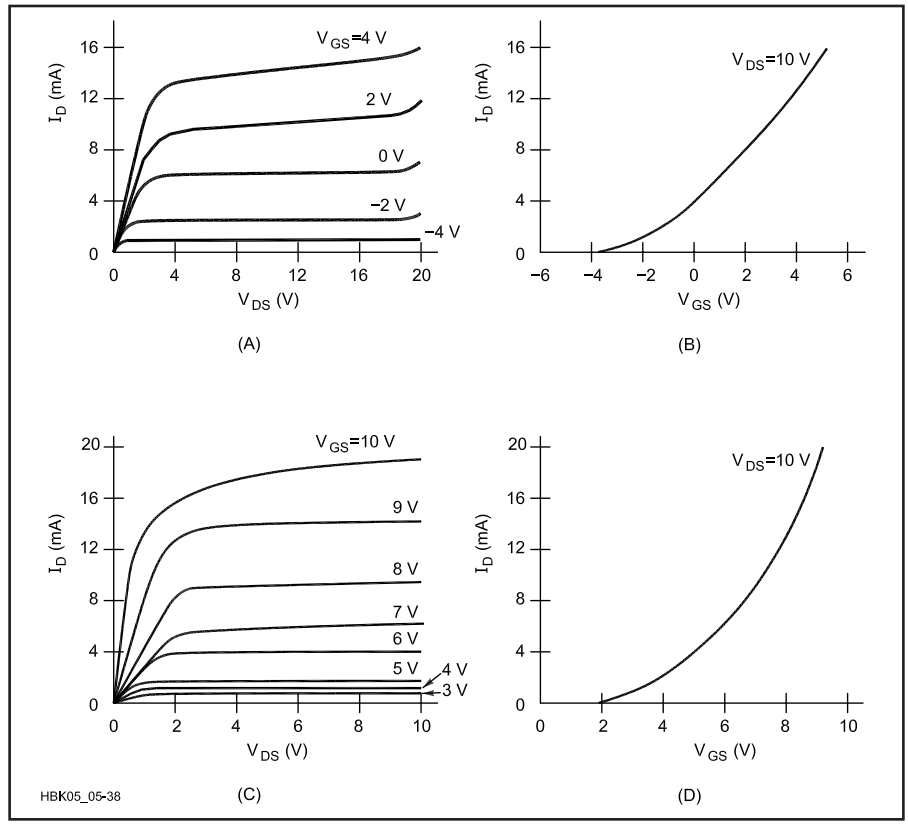


Fig 5.38 — MOSFET output [(A) and (C)] and transconductance [(B) and (D)] response curves. Plots (A) and (B) are for an N-channel depletion mode device. Note that V_{GS} varies from negative to positive values. Plots (C) and (D) are for an N-channel enhancement mode device. V_{GS} has only positive values.

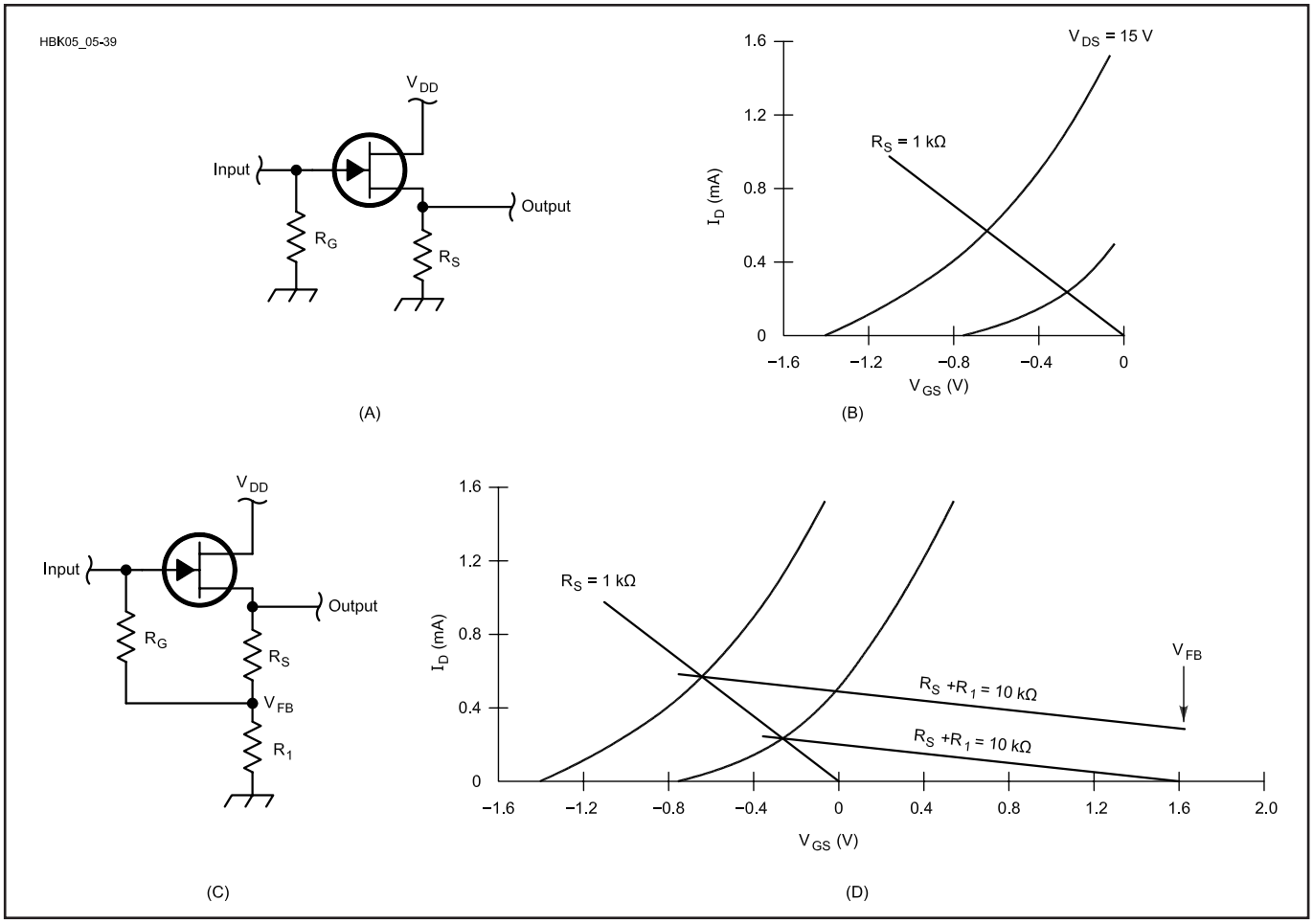


Fig 5.39 — FET biasing circuits. (A) Self-biased common drain JFET circuit. (B) Transconductance curve for self biased JFET in (A). Gate bias is determined by current through R_S. Load line has a slope of – 1 / R_S, and gate bias voltage can vary between where the load line crosses the characteristic curves. (C) Feedback bias common drain JFET circuit.

I_D, changes very little as the drain-source voltage, V_{DS}, varies in this portion of the curve. Thus, for a fixed gate-source voltage, V_{GS}, the drain current can be considered to be constant over a wide range of drain-source voltages.

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

FET Biasing

There are two ways to bias an FET, with and without feedback. Source self biasing for an N-channel JFET is pictured in **Fig 5.39A**. In this common-drain amplifier circuit, bias level is determined by the current through R_S, since I_G is very small and there is essentially no voltage drop across R_G. The characteristic curve for this configuration is plotted in Fig 5.39B. The operating points of the amplifier are where the load line intersects the curves. An example of feedback biasing is shown in Fig 5.39C. R_1 is generally much larger than R_S and the load line is determined by the sum of these resistors, as shown in Fig 5.39D. Feedback biasing increases the input impedance of the amplifier, but is rarely required, since input resistance (R_G) can be made very large.

MOSFET Gate Protection

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

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

Transistor Amplifier Design — a Practical Approach

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

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

How Much Gain?

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

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

Biasing the Transistor Amplifier

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

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

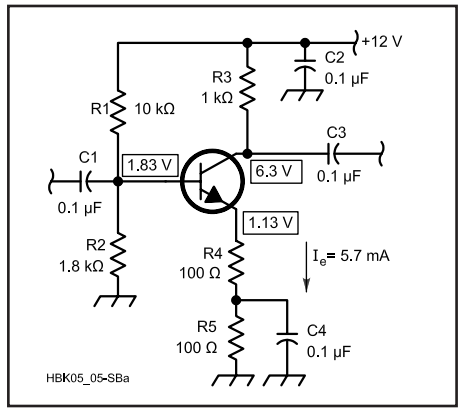


Fig A — Example of a simple low-frequency capacitively coupled transistorized small-signal amplifier. The voltages shown are the preliminary values desired for a collector current of 5 mA. The ac voltage gain is the ratio of the collector load resistor, R3, divided by the unbypassed portion of the emitter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

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

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

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

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

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

Now, let's calculate more accurately the result from using standard values for R1 and R2. The actual

base voltage will be 12 V × [1.8 kΩ / (1.8 kΩ + 10 kΩ)] = 1.83 V, rather than 1.7 V. The resulting emitter voltage is 1.83 V – 0.7 V = 1.13 V, resulting in 1.13 V / 200 Ω = 5.7 mA of collector current, rather than our desired 5 mA. We are close enough — we have finished designing the bias circuitry!

Performance: Voltage Gain

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

Performance: Input Resistance

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

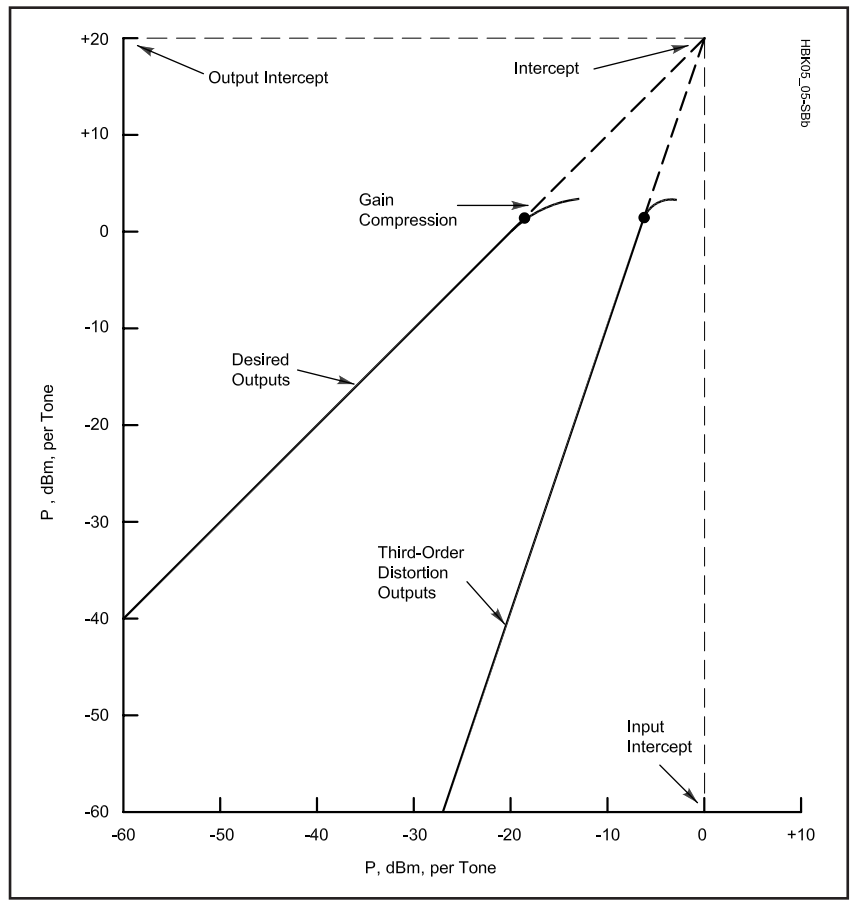


Fig B — Output IMD (intermodulation distortion) as a function of input. In the region below the 1-dB compression point, a decrease in input level of 10 dB results in a drop of IMD products by 30 dB below the level of each output tone in a two-tone signal.

Performance: Overload

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

When a negative-going ac voltage is supplied to the base, the collector current falls, and the collector voltage will rise by the voltage gain of 10. The maximum amount of voltage possible is the 12 V supply voltage, where the transistor is cut off with no collector current. The maximum positive collector swing is from the standing collector dc voltage to the supply voltage: 12 V – 6.3 V = 5.7 V positive swing. This occurs with a peak negative input voltage of 5.7 V / 10 = 0.57 V. Our amplifier will overload rather symmetrically on both negative and positive peaks. This is no accident — we biased it to have a collector voltage halfway between ground and the supply voltage.

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

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

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

forward biased and the MOSFET would not work properly. In some depletion mode MOSFET devices, back-to-back Zener diodes are used to protect the gate.

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

OPTICAL SEMICONDUCTORS

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

Photoconductors

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

A similar effect is used in some diodes and transistors so that their operation can be controlled by light instead of electrical current biasing. These devices are called *photodiodes* and *phototransistors*. The flow of minority carriers across the reverse biased PN junction is increased by light falling on the doped semiconductor material. In the dark, the junction acts the same as any reverse biased PN junction, with a very low current (on the order of 10 μA) that is nearly independent of reverse voltage. The presence of light not only increases the current but also provides a resistance-like relationship (reverse current increases as reverse voltage increases). See **Fig 5.40** for the characteristic response of a photodiode. Even with no reverse voltage applied, the presence of light causes a small reverse current, as indicated by the points at which the lines in Fig 5.40 intersect the left side of the graph. Photoconductors and photodiodes are generally used to produce light-related analog signals that require further processing. The phototransistor can often be used to serve both purposes, acting as an amplifier whose gain varies with the amount of light present. It is also more sensitive to light than the other devices. Phototransistors have lots of gain, but photodiodes normally have less noise, so they make sensitive detectors.

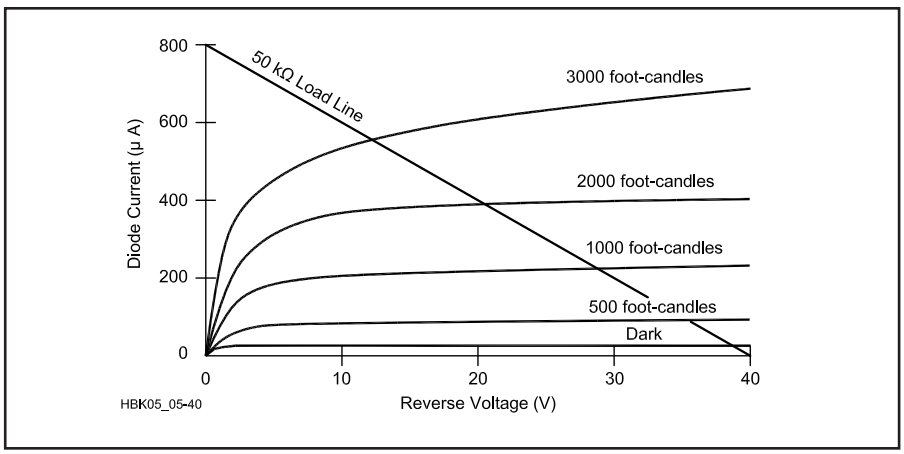


Fig 5.40 — Photodiode response curve. Reverse voltage is plotted on the x-axis and current through diode is plotted on the y-axis. Various response lines are plotted for different illumination. Except for the zero illumination line, the response does not pass through the origin since there is current generated at the PN junction by the light energy. A load line is shown for a 50 kΩ resistor in series with the photodiode.

Photovoltaic Effect

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

Light Emitting Diodes

In the photodiode, energy from light falling on the semiconductor material is absorbed to make additional electron-hole pairs. When the electrons and holes recombine, the same amount of energy is given off. In normal diodes the energy from recombination of carriers is given off as heat. In certain forms of semiconductor material, the recombination energy is given off as light with a mechanism called *electroluminescence*. Unlike the incandescent light bulb, electroluminescence is a cold light source that typically operates with low voltages and currents (such as 1.5 V and 10 mA). Devices made for this purpose are called *light emitting diodes (LEDs)*. They have the advantages of low power requirements, fast switching times (on the order of 10 ns) and narrow spectra (relatively pure color). The LED emits light when it is forward biased and excess carriers are present. As the carriers recombine, light is produced with a color that depends on the properties of the semiconductor material used. Gallium arsenide (GaAs) generates light in the infrared region, gallium phosphide (GaP) gives off red light when doped with oxygen or green light when doped with nitrogen. Orange light is attained with a mixture of GaAs and GaP (GaAsP). Silicon doped with carbon gives off yellow light but does not produce much illumination. Other colors are also possible with different types and concentrations of dopants but usually have lower illumination efficiencies.

The LED is very simple to use. It is connected across a voltage source with a series resistor that limits the current to the

desired level for the amount of light to be generated. The cathode lead is connected to the lower potential, and is usually specially marked (flattening of the lead near the package, a dot of paint next to the lead, and a flat portion of the round device located next to the lead are all common methods).

Optoisolators

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

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

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

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

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

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

Fiber Optics

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

LINEAR INTEGRATED CIRCUITS

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

Hybrid Integrated Circuits

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

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

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

Monolithic Integrated Circuits

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

An integrated circuit (IC) is fabricated in layers. An example of a semiconductor circuit schematic and its implementation in an IC is pictured in **Fig 5.41**. The base layer of the circuit, the *substrate*, is made of P-type semiconductor material. Although less common, the polarity of the substrate can also be N-type material. Since the mobility of electrons is about three times higher than that of holes, bipolar transistors made with N-type collectors and FETs made with N-type channels are capable of higher speeds and power handling. Thus, P-type substrates are far more common. For devices with N-type substrates, all polarities in the ensuing discussion would be reversed. Other substrates have been used, one of the most successful of which is the silicon-on-sapphire (SOS) construction that has been used to increase the bandwidth of integrated circuitry. Its relatively high manufacturing cost has impeded its use, however.

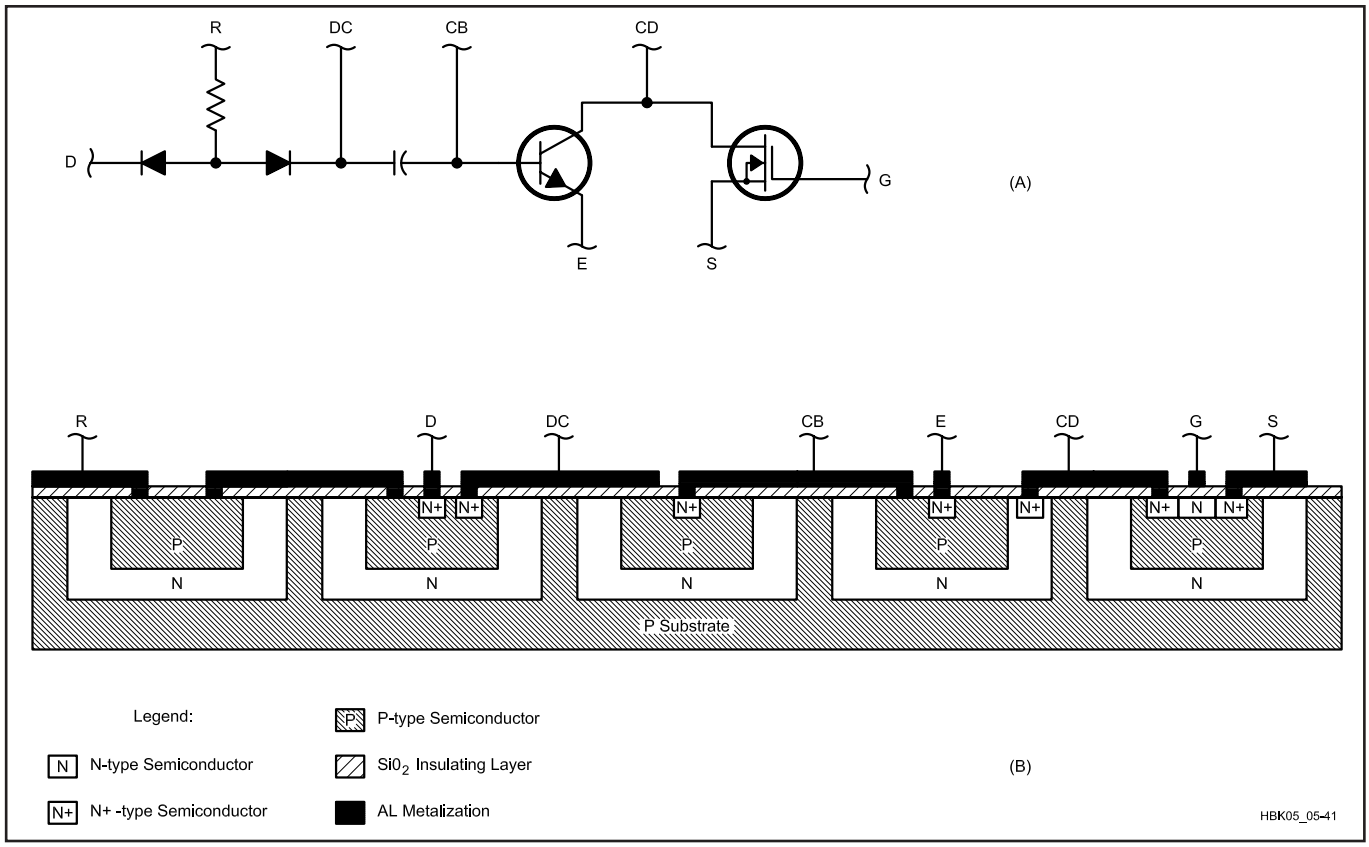


Fig 5.41 — Integrated circuit layout. (A) Circuit containing two diodes, a resistor, a capacitor, an NPN transistor and an N-channel MOSFET. Labeled leads are D for diode, R for resistor, DC for diode-capacitor, E for emitter, S for source, CD for collector-drain and G for gate. (B) Integrated circuit that is identical to circuit in (A). Same leads are labeled for comparison. Circuit is built on a P-type semiconductor substrate with N-type wells diffused into it. An insulating layer of SiO_2 is above the semiconductor and is etched away where aluminum metal contacts are made with the semiconductor. Most metal-to-semiconductor contacts are made with heavily doped N-type material (N^+-type semiconductor).

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 metal (usually aluminum) and is used to make the interconnections between the components.

Different components are made in a single piece of semiconductor material by first diffusing a high concentration of acceptor impurities into the layer of N-type material. This process creates P-type semiconductor — often referred to as P^+-type semiconductor because of its high concentration of acceptor atoms — that isolates regions of N-type material. Each of these regions is then further processed to form single components. A component is produced by the diffusion of a lesser concentration of acceptor atoms into the middle of each isolation region. This results in an N-type *isolation well* that contains P-type material, is surrounded on its sides by P^+-type material and has P-type material (substrate) below it. The cross sectional view in Fig 5.41B 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.

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

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

sectional size of the well. This type of thin film capacitor is also known as a metal oxide semiconductor (MOS) capacitor.

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

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

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

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

Complementary Metal Oxide Semiconductors

Power dissipation in a circuit can be reduced to very small levels (on the order of a few nW) by using the MOSFET devices in complementary pairs (CMOS). Each amplifier is constructed of a series circuit of MOSFET devices, as in **Fig 5.42**. The gates are tied together for the input signal, as are the drains for the output signal. In saturation and cutoff, only one of the devices conducts. The current drawn by the circuit under no load is equal to the OFF leakage current of either device and the voltage drop across the pair is equal to V_{DD}, so the steady state power used by the circuit is always equal to $V_{DD} \times I_{D(OFF)}$. For ac signals, power consumption is proportional to frequency.

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

Integrated Circuit Advantages

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

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

Decreased circuit size and improved reliability are only two of the advantages of monolithic integrated circuitry. The uncertainty of the exact behavior of the integrated components is the same as it is for discrete components, as discussed earlier. The relative properties of the devices on a single chip are very predictable, how-

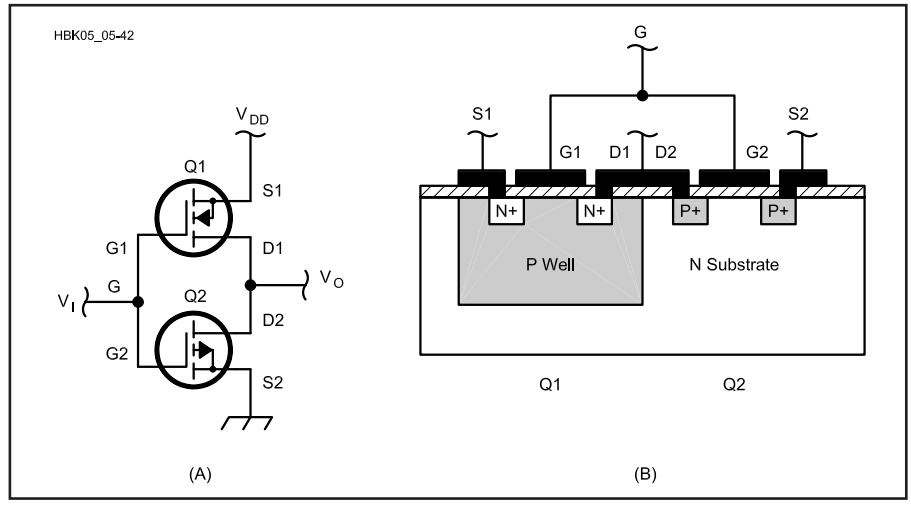


Fig 5.42 — Complementary metal oxide semiconductor (CMOS). (A) CMOS device is made from a pair of enhancement mode MOS transistors. The upper is an N-channel device, and the lower is a P-channel device. When one transistor is biased on, the other is biased off; therefore, there is minimal current from V_{DD} to ground. (B) Implementation of a CMOS pair as an integrated circuit.

ever. Since adjacent components on a semiconductor chip are made simultaneously (the entire N-type layer is grown at once, a single diffusion pass isolates all the wells and another pass fills them), the characteristics of identically formed components on a single chip of silicon should be identical. Even if the exact characteristics of the components are unknown, very often in analog circuit design the major concern is how components interact. For instance, push-pull amplifiers require perfectly matched transistors, and the gain of many amplifier configurations is governed by the ratio between two resistors and not their absolute values of resistance.

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

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

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

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

Integrated Circuit Disadvantages

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

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

Integrated circuits, despite their short interconnection lengths and lower stray inductance, do not have as high a frequency response as similar circuits built with appropriate discrete components. (There are exceptions to this generalization, of course. Monolithic microwave integrated circuits — MMICs — are available for operation on frequencies up through 10 GHz.) The physical architecture of an integrated circuit is the cause of this limitation. Since the substrate and the walls of the isolation wells are made of opposite types of semiconductor material, the PN junction between them must be reverse biased to prevent current from passing into the substrate. Like any other reverse biased PN junction, a capacitance is created at the junction and this limits the frequency response of the devices on the IC. This situation has improved over the years as isolation wells have gotten smaller, thus decreasing the capacitance between the well and the substrate, and techniques have been developed to decrease the PN junction capacitance at the substrate. One such technique has been to create an N^+-type layer between the well and the substrate, which decreases the capacitance of the PN junction as seen by the well. As an example, in the 1970s the LM324 operational amplifier IC package was developed by National Semiconductor and claimed a gain-bandwidth product of 1 MHz. In the 1990s the HFA1102 operational amplifier IC, developed by Harris Semiconductor, was introduced with a gain-bandwidth product of 600 MHz.

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

Common Types of Linear Integrated Circuits

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

Component Arrays

The most basic form of linear integrated circuit is the component array. The most common of these are the resistor, diode

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

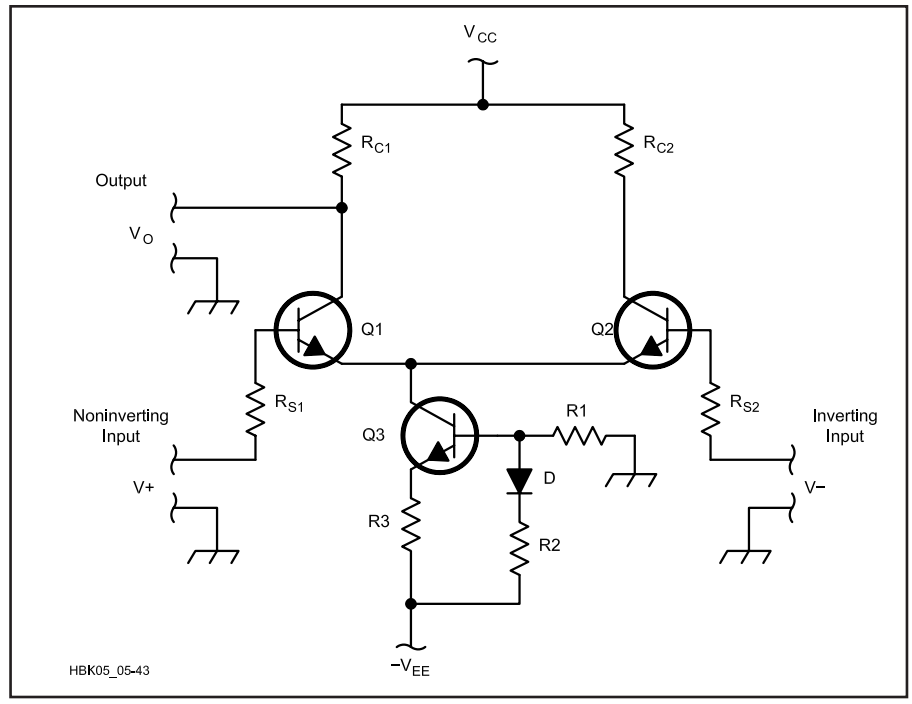


Fig 5.43 — Schematic of the components that make up an operational amplifier. Q1 and Q2 are matched emitter-coupled amplifiers. Q3 provides a constant current source. The symmetry of this device makes the matching of the components critical to its operation. This is why this circuit is usually implemented only in integrated circuitry. This simple op amp design has a large dc offset voltage at the output. Most practical designs include a level-shifting circuit, so the output voltage can exist near ground potential.

Multivibrators

A *multivibrator* is a circuit that oscillates, usually with a square wave output in the audio frequency range. The frequency of oscillation is accurately controlled with the addition of appropriate values of external resistance and capacitance. The most common multivibrator in use today is the 555 (NE555 by Signetics [now Philips] or LM555 by National Semiconductor). This very simple eight-pin DIP device has a frequency range from less than one hertz to several hundred kilohertz. Such a device can also be used in *monostable* operation, where an input pulse generates an output pulse of a different duration, or in *astable* operation, where the device freely oscillates. Some other applications of a multivibrator are as a frequency divider, a delay line, a pulse width modulator and a pulse position modulator.

Operational Amplifiers

An *operational amplifier*, or *op amp*, is one of the most useful linear devices that has been developed in integrated circuitry. While it is possible to build an op amp with discrete components, the symmetry of this circuit requires a close match of many components and is more effective, and much easier, to implement in integrated circuitry. **Fig 5.43** shows a basic op-amp circuit. The op amp approaches a perfect analog circuit building block.

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

The typical op amp has three signal terminals (see **Fig 5.44**). There are two input terminals, the noninverting terminal marked with a + sign and the inverting terminal marked with a – sign. The output of the amplifier has a single terminal and

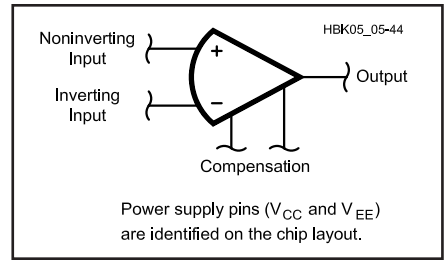


Fig 5.44 — Operational amplifier schematic symbol. The terminal marked with a + sign is the noninverting input. The terminal marked with a – sign is the inverting input. The output is to the right. On some op amps, external compensation is needed and leads are provided, pictured here below the device. Usually, the power supply leads are not shown on the op amp itself but are specified in the data sheet.

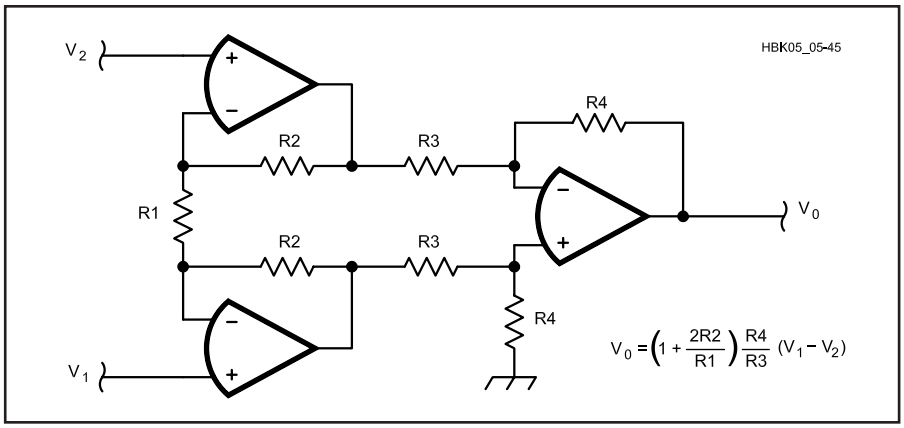


Fig 5.45 — Operational amplifiers arranged as an instrumentation amplifier. The balanced and cascaded series of op amps work together to perform differential amplification with good common-mode rejection and very high input impedance (no load resistor required) on both the inverting (V_1) and noninverting (V_2) inputs.

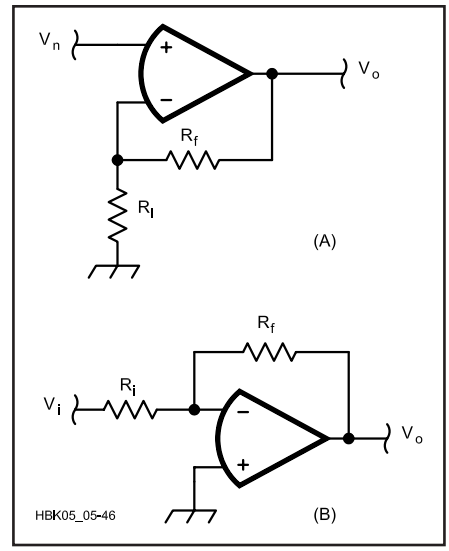


Fig 5.46 — Operational amplifier circuits. (A) Noninverting configuration. (B) Inverting configuration.

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

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

Just as the symmetry of the transistors making up an op amp leads to a device with high values of Z_i, A_v and CMRR and a low value of Z_o, a symmetric combination of op amps is used to further improve these parameters. This circuit, shown in **Fig 5.45**, is called an instrumentation amplifier.

The op amp is capable of amplifying signals to levels limited mainly by the power supplies. Two power supplies are required, thus defining the range of signal voltages that can be processed. In most op amps the signal levels that can be handled are less than the power supply limits (rails), usually one or two diode drops (0.7 V or 1.4 V) away from each rail. Thus, if an op amp has 15 V connected as its upper rail (usually denoted V^+) and ground connected as its lower rail (V^-), input signals can be amplified to be as high as 13.6 V and as low as 1.4 V in most amplifiers. Any values that would be amplified beyond those limits are clamped (output voltages that should be 1.4 V or less appear as 1.4 V and those that should be 13.6 V or more appear as 13.6 V). This clamping action is illustrated in Fig 5.1. Op amps have been developed to handle signals all the way out to the power supply rails (for example, the MAX406, from Maxim Integrated Products).

If a signal is connected to the input terminals of an op amp, it will be amplified as much as the device is able (up to A_v), and will probably grow so large that it clamps, as described above. Even if such large gains are desired, A_v varies from one device to the next and cannot be guaranteed. In most applications the op amp gain is limited to a more reasonable value and this is usually realized by providing a negative feedback path from the output terminal to the inverting input terminal. The *closed loop gain* of an op amp depends solely on the values of the passive components used to form the loop (usually resistors and, for frequency-selective circuits, capacitors). Some examples of different circuit configurations that manipulate the loop gain follow.

The op amp is often used as either an inverting or a noninverting amplifier. Accurate amplification can be achieved with just two resistors: the feedback resistor, R_f, and the input resistor, R_i (see **Fig 5.46**). If connected in the noninverting configuration, the input signal is connected to the noninverting terminal. The feedback resistor is connected between the output and the inverting terminal. The inverting terminal is connected to R_i, which is connected to ground. The gain of this configuration is:

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

where:
V_o is the output voltage, and
V_n is the input voltage to the noninverting terminal.

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

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

where V_i represents the voltage input to R_i.

The negative sign in equation 16 indicates that the signal is inverted. For ac signals, inversion represents a 180° phase shift. The gain of the noninverting op amp can vary from a minimum of × 1 to the

maximum of which the device is capable. The gain of the inverting op amp 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, as indicated by A_v for dc signals, or the gain-bandwidth product for ac signals. Both parameters are usually specified in the manufacturer's data sheet.

A voltage follower is a type of op amp that is commonly used as a buffer stage. The voltage follower has the input connected directly to the noninverting terminal and the output connected directly to the inverting terminal (**Fig 5.47**). This configuration has unity gain and provides the maximum possible input impedance and the minimum possible output impedance of which the device is capable.

A *differential amplifier* is a special application of an operational amplifier (see **Fig 5.48**). 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

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

which, if the ratios

$$\frac{R_i}{R_f} \text{ and } \frac{R_n}{R_g}$$

are equal, simplifies to

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

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

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

The high input impedance of an op amp makes it ideal for use as a *summing amplifier*. In either the inverting or noninverting configuration, the single input signal can be replaced by multiple input signals that are connected together through series resistors, as shown in **Fig 5.49**. For the inverting summing amplifier, the gain of each input signal can be calculated individually using equation 16 and, because of the superposition property, the output becomes the sum of each input signal multiplied by its gain. In the noninverting configuration, the output is the gain times the weighted sum of the m different input signals:

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

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

Other combinations of summing and difference amplification can be realized with a single op amp. The analyses of such circuits use the standard op amp equations coupled with the principle of superposition.

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

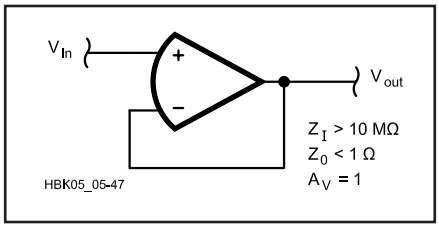


Fig 5.47 — Voltage follower. This operational amplifier circuit makes a nearly ideal buffer with a voltage gain of about one, and with extremely high input impedance and extremely low output impedance.

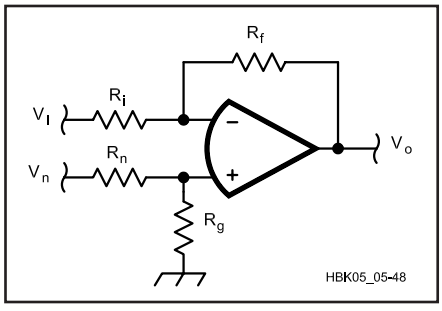


Fig 5.48 — Differential amplifier. This operational amplifier circuit amplifies the difference between the two input signals.

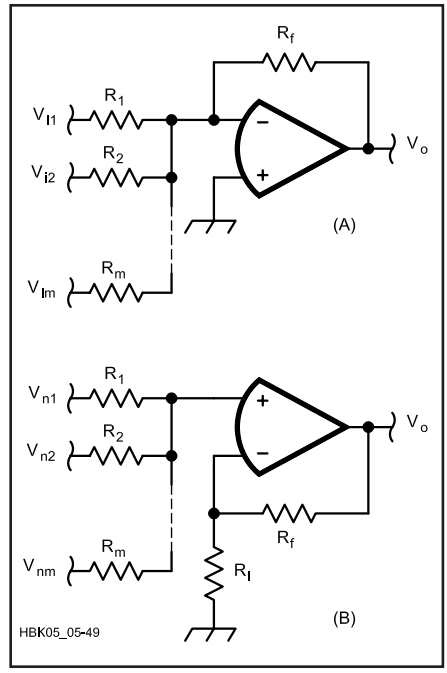


Fig 5.49 — Summing operational amplifier circuits. (A) Inverting configuration. (B) Noninverting configuration.

forms into square waves) and a peak detector.

Charge Coupled Devices

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

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

Balanced Mixers

The *balanced mixer* is a device with many applications in modern radio transceivers (see the **Mixers, Modulators and Demodulators** chapter). Audio signals can be modulated onto a carrier or demodulated from the carrier with a balanced mixer. RF signals can be downconverted to intermediate frequency (IF) or IF can be upconverted to RF with a balanced mixer. This device is made with a bridge of four matched Schottky diodes and the necessary transformers packaged in a small metal, plastic or ceramic container. The consequence of unmatched diodes is poor isolation between the local oscillator (LO) and the two signals. IC mixers often use a "Gilbert cell" to provide LO isolation as high as –30 dB at 500 MHz. The isolation improves with decreasing frequency.

Receiver Subsystems

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

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

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

Older integrated circuits that are subsets of the receiver subsystems are popular. The NE602 contains one double balanced mixer and a local oscillator, along with voltage regulation and buffering (**Fig 5.50**). It contains almost everything required to construct a direct conversion receiver. Its small size, an 8-pin DIP, makes it more desirable for this purpose than using part of an MC3363, which is in a 24-pin DIP and is more expensive. The NE604 contains the IF amplifiers and quadrature detector that, together with two NE602s and an RF amplifier, could almost duplicate the functions of the MC3363 or the NE367.

Transmitter Subsystems

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

Monolithic Microwave Integrated Circuit

A class of bipolar IC that is capable of higher frequency responses is the *mono-*

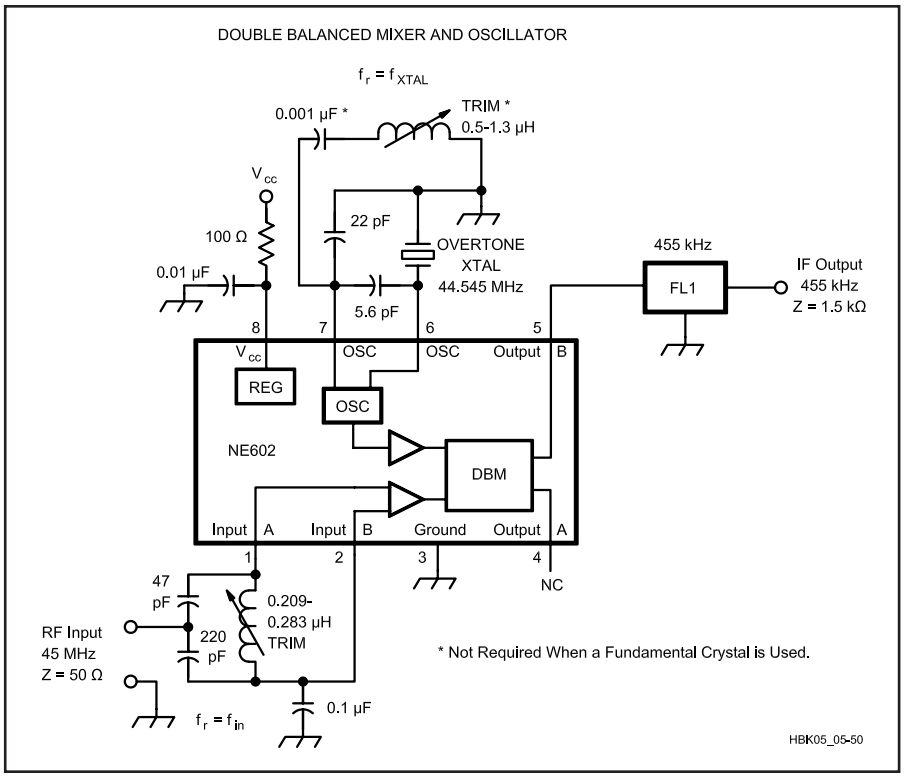


Fig 5.50 — The NE602 functional block diagram in circuit. This device contains a doubly balanced mixer, a local oscillator, buffers and a voltage regulator. This application uses the NE602 to convert an RF signal in a receiver to IF.

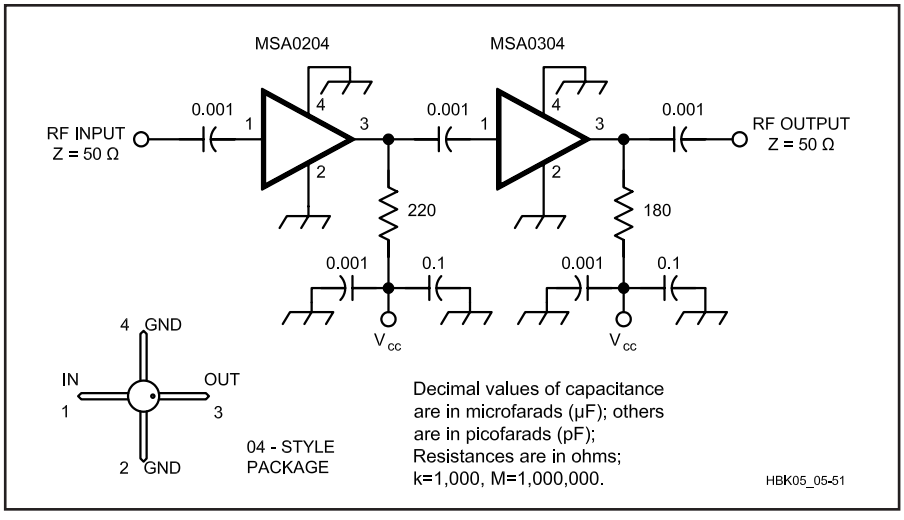


Fig 5.51 — The MSA0204 and MSA0304 MMICs in circuit. Both amplifiers have both input and output impedance of 50 Ω and a bandwidth of more than 2.5 GHz.

lithic microwave integrated circuit (MMIC). There is no formal definition of when an IC amplifier becomes an MMIC and, as the performance of IC devices improves, particularly MOS based devices, the distinction is becoming blurred. MMIC devices typically have predefined operating characteristics and require few external components. An example of an MMIC is a fixed gain amplifier, the MSA0204 (**Fig 5.51**), which can deliver 12 dB of gain up to 1 GHz. More modern MMIC devices are being developed with bandwidths in the tens of GHz.

Comparison of Analog Signal Processing Components

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

Of the four component types, vacuum tubes are physically the largest and require the most operating power. They have more limited life spans, usually because the heater filament burns out just as a light bulb does. Regardless of its use, a vacuum tube always generates heat. Miniaturization is difficult with vacuum tubes both because of their size and because of the need for air space around them for cooling. Vacuum tubes do have advantages, however. They are electrically robust. You need not be as concerned about static charges destroying vacuum tubes. A transmitter with vacuum tube finals usually has a variable matching network built in, and can be loaded into a higher SWR than one with semiconductor finals. Tubes are generally able to withstand the high voltages generated by reflections under high SWR conditions. They are not as easily damaged by short-term overloads or the electromagnetic pulses generated by lightning. The relatively high plate voltages mean that the plate current is lower for a given power output; thus power supplies do not need as high a current handling capability. Vacuum tubes are capable of considerable heat dissipation and many high power applications still use them. Special forms of vacuum tubes are also still used. Most video displays use CRTs, and microwave transmitting tubes are still common.

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

There are many performance advantages to FET devices, particularly MOSFETs. The extremely low gate currents allow the design of analog stages with nearly infinite input resistance. Signal distortion due to loading is minimized in this way. As these characteristics are

improved by technology, we are seeing an increase in FET design at the expense of bipolar transistors.

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

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

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

Digital Fundamentals

Radio Amateurs have been involved with digital technology since the first spark transmitters, a form of pulse-coded transmission, were connected to an "aerial." Modern digital technology use by Radio Amateurs probably arrived first in automatic keyers, where hams learned about flip-flops and gates to replace their semi-automatic mechanical keyers (bugs). Amateur use of digital technology echoes public use of these new abilities, starting with using the first home computers for calculations and later as Teletype and data terminals.

The first PC hobbyists, on the west coast, worked on the idea that technology, and in particular software, should be free and available to all. It is not very surprising to learn that many of these PC pioneers were also Radio Amateurs, who were accustomed to seeing new technical ideas freely distributed in the pages of *QST*

The remainder of this chapter, written by Christine Montgomery, KGØGN, with additional material by Paul Danzer, N1II, presents digital-theory fundamentals and some applications of that theory in Amateur Radio. The fundamentals introduce digital mathematics, including number systems, logic devices and simple digital circuits. Next, the implementation of these simple circuits is explored in integrated circuits, their families and interfacing. Integrated circuits continue with memory chips and microprocessors, culminating in a synthesis of these components in the modern digital computer. Where possible, this section mentions Amateur Radio applications associated with the technologies being discussed, as well as pointers to other chapters that discuss such applications in greater depth.

DIGITAL VS. ANALOG

An essential first step in understanding digital theory is to understand the difference between a *digital* and an *analog* signal. An analog value, a real number, has no end; for example, the number 1/3 is 0.333... where the 3 can be repeated forever, or 3/4 equals 0.7500... with infinite repeated 0s. A digital approximation of an analog number breaks the real number line into discrete steps, for example, the integers. This process of approximating a value with discrete steps either truncates or rounds an analog value to some number of decimal places. For example, rounding 1/3 to an integer gives 0 and rounding 3/4 gives 1.

For a simple physical example, look at your wristwatch. A watch with a face — with the hands of the watch rotating in a continuous, smooth motion — is an analog display. Here, the displayed time has a *continuous* range of values, such as from 12:00 exactly to 12:00 and 1/3 second or any values in between. In contrast, a watch with a digital display is limited to *discrete* states. Here the displayed time jumps from 12:00 and 0 seconds to 12:00 and 1 second, without showing the time in between. (A watch with a second hand that jerks from one second to another could also fit the digital analogy.)

In the digital watch example, time is represented by ten distinct states (0, 1, 2, 3, 4, 5, 6, 7, 8 and 9). Digital electronic signals, however, will usually be much more limited in the number of states allowed.

The digital system used is also called the *binary* system, since only two values are allowed. By using coding, as discussed in the following pages, these two binary values can represent any number of real values. **Fig 5.52** illustrates the contrast of an analog signal (in this case a sine wave) and its digital approximation. Three positive and three negative values are shown as an approximation to the sine wave, but any number of coded value steps can be used as an approximation.

While the focus in this chapter will be on digital theory, many circuits and systems involve *both* digital and analog components. Often, a designer may choose between using digital technology, analog technology or a combination.

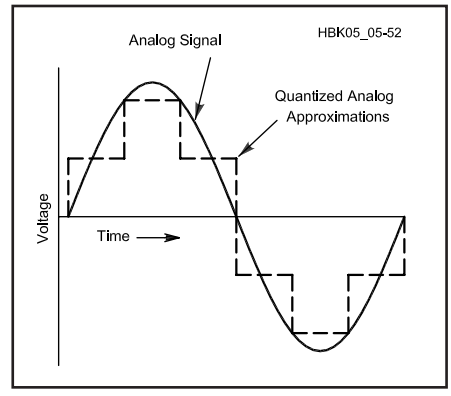


Fig 5.52 — An analog signal and its analog approximation. Note that the analog waveform has continuously varying voltage while the approximated waveform is composed of discrete steps.

Number Systems

In order to understand digital electronics, you must first understand the digital numbering system. Any number system has two distinct characteristics: a set of *symbols* (digits or numerals) and a *base* or radix. A *number* is a collection of these digits, where the left-most digit is the *most significant digit (MSD)* and the right-most digit is the *least significant digit (LSD)*. The value of this number is a weighted sum of its digits. The *weights* are determined by the system's base and the digit's position relative to the decimal point.

While these definitions may seem strange with all the technical terms, they will be more familiar when seen in a decimal system example. This is the "traditional" number system with which we are all familiar.

DECIMAL

The decimal system is a base-10 system, with ten symbols: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}. To count, we start at 0, and then work our way up to the highest single value allowed — 9. Therefore we count 0, 1, 2,3, 4, 5, 6, 7, 8, 9.

Consider 3 digits, represented by XXX. We start at 000, and then fill up the first (least significant) column, on the right:

XXX
000
001
002
003
...
009

We now reset the first column to the lowest possible value, 0, and increase the second column by 1.

010
011
012
013
...
019

We have again filled up the first column, so again reset it to 0, and increase the second column by one.

020
021
022
023
...
029

We repeat this process, until we hit 099. At this point the second column is filled, so we reset the first two columns to 00 and increase the third column by 1, giving us 100. This is how our familiar decimal or 10-digit number system works; number systems working on other bases work the same way.

Each column in a number has a property called weight. As an example, look at the decimal number, 548. The digits are 5, 4, 8, where 5 is the most significant digit since it is positioned to the far left and 8 is the least significant digit since it is positioned to the far right. The value of this number is a weighted sum of its digits, as shown in **Table 5.1**.

The weight of a position is the system's base raised to a power. In this case, for a decimal system the base is 10, so each position is weighted by 10^P with the power determined by the position relative to the decimal. For example, digit 8, immediately to the left of the decimal, is at position 0; therefore, its weight factor is $10^0 = 1$. Similarly, digit 5 is 2 positions to the left of the decimal and has a weight factor $10^2 = 100$. The value of the number is the sum of each digit times its weight.

BINARY

Binary is a *base-2* number system and therefore limited to two symbols: {0, 1}. The weight factors are now powers of 2, like 2^0, 2^1 and 2^2. For example, the decimal number, 163 and its equivalent binary number, 10100011, are shown in **Table 5.2**.

The digits of a binary number are now *bits* (short for binary digit). The MSD is the *most significant bit (MSB)* and the LSD is the *least significant bit (LSB)*. Four bits make a *nibble* and two nibbles, or eight bits, make a *byte*. A *word* can consist of two or four or more bytes. These groupings are useful when converting to hexadecimal notation, which is explained later.

Counting in binary follows the same pattern illustrated for decimal. Consider the three digit binary number XXX. First fill up the right-hand column.

XXX
000
001

The column has been filed, and much quicker then with decimal, since there are only two values instead of 10. But just like decimal, now reset the right-hand column to 0, increase the next column by 1, and continue.

XXX
000
001
010 ←
011

Table 5.1
Decimal Numbers

Example: $5(10^2)$

Digit = 5; Weight = 10; Position = 2

548	=	$5(10^2)$	+	$4(10^1)$	+	$8(10^0)$
	=	$5(100)$	+	$4(10)$	+	$8(1)$
	=	500	+	40	+	8
	=	5		4		8
		MSD				LSD

Table 5.2
Decimal and Binary Number Equivalents

163	=	128	+	0	+	32	+	0	+	0	+0	+	2+	1	decimal
	=	1(128)	+	0(64)	+	1(32)	+	0(16)	+	0(8)	+0(4)	+	1(2)	+1(1)	
	=	$1(2^7)$	+	$0(2^6)$	+	$1(2^5)$	+	$0(2^4)$	+	$0(2^3)$	$+0(2^2)$	+	$1(2^1)$	$+1(2^0)$	
10100011	=	1		0		1		0		0	0		1	1	binary
		MSB												LSB	

Nibble Nibble

Byte = 8 digits

Now the first two columns are full, so reset both back to 0 and increase the next column by 1 and continue:

XXX
000
001
010
011
100 ←
101
110
111 and so on.

Examination of the set of binary numbers from 0 to 15 shows some important characteristics:

Binary Value	*Decimal Value*
0000	0
0001	1
0010	2
0011	3
0100	4
0101	5
0110	6
0111	7
1000	8
1001	9
1010	10
1011	11
1100	12
1101	13
1110	14
1111	15

Notice each column starts with 0. The first (right-most) column alternates; every other value is a 0 and a 1. The second column alternates every two values, that is there are two 0's followed by two 1's. The third column has groups of four 0's and four 1's, and the fourth column has groups of eight 0's and eight 1's. Thus you can make up a binary counting table by simply following this pattern.

HEXADECIMAL

The hexadecimal, or hex, *base-16* number system is widely used in personal computers for its ease in conversion to and from binary numbers and the fact that it is somewhat more human-friendly than long strings of 1's and 0's. A base-16 number requires 16 symbols. Since our normal mathematical number, as set up in the decimal system, has only 10 digits (0 through 9), a set of additional new symbols is required. Hex uses both numbers and characters in its set of sixteen symbols: {0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F}. Here, the letters A to F have the decimal equivalents of 10 to 15 respectively: A=10, B=11, C=12, D=13, E=14 and F=15. Again, the weights are powers of the base, such as 16^0, 16^1 and 16^2.

The four-bit binary listing in the previous paragraph shows that the individual 16 hex digits can be represented by a four-bit binary number. Four binary digits are called a *nibble*.

Since a byte is equal to eight binary digits, two hex digits provide a byte — the equivalent of 8 binary digits. Conversion from binary to hex is therefore simplified. Take a binary number, divide it into groups of four binary digits starting from the right, and convert each of the four binary digits to an individual value.

Conversion from hex to binary is equally convenient; replace each hex digit with its four-bit binary equivalent. As an example, the decimal number 163 is shown in Table 5.2 as binary 10100011. Divide the binary number in groups of four, so 1010 is equivalent to decimal 10 or "A" hex, and 0011 is equivalent to decimal 3, thus decimal number 163 is equivalent to hex A3.

BINARY CODED DECIMAL (BCD)

Scientists have experimented with many devices out of a desire for fast computations. Initially, analog computers were developed and used for many applications, especially military applications. It was not unusual to see analog computers aboard US navy ships as recently as the mid-1960s, where they were used to direct naval gunfire.

Analog computers have a very large disadvantage; they could not be readily reprogrammed. They did have a very great advantage; their output was more human readable than digital computers. Very few humans can get used to either binary or hexadecimal read-outs!

The binary number system representation is the most appropriate form for fast internal computations since there is a direct mathematical relationship for every bit in the number. To interface with a user — who usually wants to see I/O in terms of decimal numbers — other codes are more useful. The *Binary Coded Decimal (BCD)* system is the simplest and most widely used form for inputs and outputs of user-oriented digital systems.

In the Binary Coded Decimal (BCD) system, each decimal digit is expressed as a corresponding 4-bit binary number. In other words, the decimal digits 0 to 9 are encoded as the bit strings 0000 to 1001. To make the number easier to read, a space is left between each 4-bit group. For example, the decimal number 163 is equivalent to the BCD number 0001 0110 0011, as shown in **Table 5.3**.

A generic code could use any n-bit string to represent a piece of information. BCD uses 4 bits because that is the minimum needed to represent a 9. All four bits are always written; even a decimal 0 is written as 0000 in BCD.

The important difference between BCD and the previous number systems is that, starting with decimal 10, BCD loses the standard mathematical relationship of a weighted sum. BCD is simply a cut-off hexadecimal. Instead of using the 4-bit code strings 1010 to 1111 for decimal 10 to 15, BCD uses 0001 0000 to 0001 0101. There are other n-bit decimal codes in use and, even for specifically 4 bits, there are millions of combinations to represent the decimal digits 0-9. BCD is the simplest way to convert between decimal and a binary code; thus it is the ideal form for I/O interfacing. The binary number system, since it maintains the mathematical relationship between bits, is the ideal form for the computer's internal computations.

CONVERSION TECHNIQUES

An easy way to convert a number from decimal to another number system is to do repeated division, recording the remainders in a tower just to the right. The converted number, then, is the remainders, reading up the tower. This technique is illustrated in **Table 5.4** for hexadecimal and binary conversions of the decimal number 163.

For example, to convert decimal 163 to hex, repeated divisions by 16 are performed. The first division gives 163/16 = 10 remainder 3. The remainder 3 is written in a column to the right. The second division gives 10/16 = 0 remainder 10. Since 10 decimal = A hex, A is written in the remainder column to the right. This division gave a divisor of 0 so the process is complete. Reading up the remainders column, the result is A3. The most common mistake in this technique is to forget that the Most Significant Digit ends up at the bottom.

Another technique that should be

Table 5.3
Binary Coded Decimal Number Conversion

		0 0 0 1	0 1 1 0	0 0 1 1	
		⌊ ⌋	⌊ ⌋	⌊ ⌋	**BCD**
	=	$1(2^0)$	$1(2^2) + 1(2^1)$	$1(2^1) + 1(2^0)$	
	=	(1)	(4 + 2)	(2 + 1)	
163	=	1	6	3	**decimal**

Table 5.4
Number System Conversions

Hex		*Remainder*		*Binary*		*Remainder*	
16	|163			2	|163		
	|10	3	LSB		|81	1	LSB
	|0	A	MSB		|40	1	
					|20	0	
					|10	0	
					|5	0	
					|2	1	
					|1	0	
					|0	1	MSB
A3 hex				1010 0011 binary			

briefly mentioned can be even easier: get a calculator with a binary and/or hex mode option. One warning for this technique: this chapter doesn't discuss negative binary numbers. If your calculator does not give you the answer you expected, it may have interpreted the number as negative. This would happen when the number's binary form has a 1 in its MSB, such as the highest (leftmost) bit for the binary mode's default size. To avoid learning about negative binary numbers, always use a leading 0 when you enter a number in binary or hex into your calculator.

Physical Representation of Binary States

STATE LEVELS

Most digital systems use the binary number system because many simple physical systems are most easily described by two state levels (0 and 1). For example, the two states may represent "on" and "off," a punched hole or the absence of a hole in paper tape or a card, or a "mark" and "space" in a communications transmission. In electronic systems, state levels are physically represented by voltages. A typical choice is

state 0 = 0 V
state 1 = 5 V

Since it is unrealistic to obtain these exact voltage values, a more practical choice is a range of values, such as

state 0 = 0.0 to 0.4 V
state 1 = 2.4 to 5.0 V

Fig 5.53 illustrates this representation of states by voltage levels. The undefined region between the two binary states is also known as the *transition region* or *noise margin*.

Transition Time

The gap in **Fig 5.53**, between binary 0 and binary 1, shows that a change in state does not occur instantly. There is a *transition time* between states. This transition time is a result of the time it takes to charge or discharge the stray capacitance in wires and other components because voltage cannot change instantaneously across a capacitor. (Stray inductance in the wires also has an effect because the current through an inductor can't change instantaneously.) The transition from a 0 to a 1 state is called the *rise time*, and is usually specified as the time for the pulse to rise from 10% of its final value to 90% of its final value. Similarly, the transition from a 1 to a 0 state is called the *fall time,* with a similar 10% t0 90% definition. Note that these times need not be the same. **Fig 5.54A** shows an ideal signal, or *pulse*, with zero-time switching. Fig 5.54B shows a typical pulse, as it changes between states in a smooth curve.

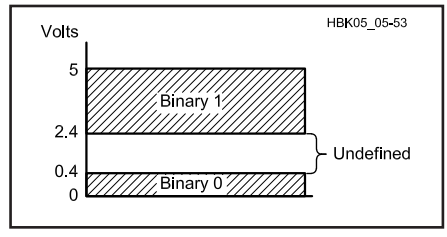


Fig 5.53 — Representation of binary states 1 and 0 by a selected range of voltage levels.

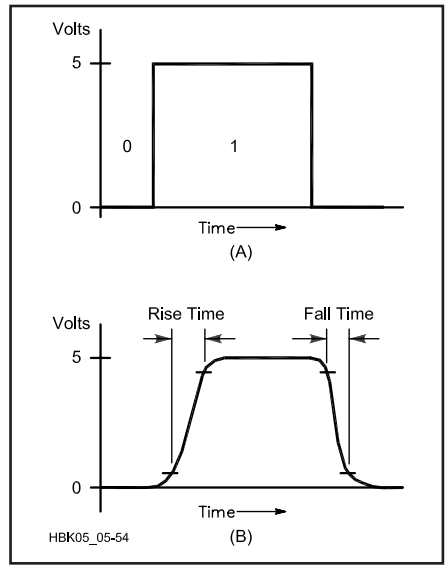


Fig 5.54 — (A) An ideal digital pulse and (B) a typical actual pulse, showing the gradual transition between states.

Rise and fall times vary with the logic family used and the location in a circuit. Typical values of transition time are in the microsecond to nanosecond range. In a circuit, distributed inductances and capacitances in wires or PC-board traces may cause rise and fall times to increase as the pulse moves away from the source.

Propagation Delay

Rise and fall times only describe a relationship within a pulse. For a circuit, a pulse input into the circuit must propagate through the circuit; in other words it must pass through each component in the circuit until eventually it arrives at the circuit output. The time delay between providing an input to a circuit and seeing a response at the output is the *propagation delay* and is illustrated by **Fig 5.55**.

For modern switching logic, typical propagation delay values are in the 1 to 15 nanosecond range. (It is useful to remem-

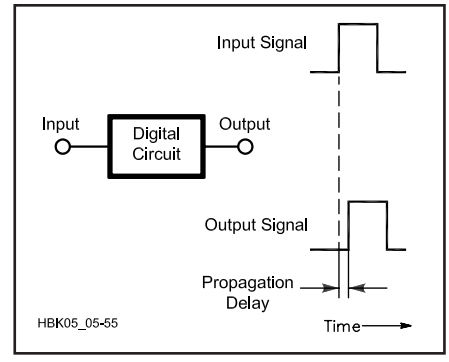


Fig 5.55 — Propagation delay in a digital circuit.

ber that the propagation delay along a wire or printed-circuit-board trace is about 1.0 to 1.5 ns per inch.) Propagation delay is the result of cumulative transition times as well as transistor switching delays, reactive element charging times and the time for signals to travel through wires. In complex circuits, different propagation delays through different paths can cause problems when pulses must arrive somewhere at exactly the same time.

The effect of these delays on digital devices can be seen by looking at the speed of the digital pulses. Most digital devices and all PCs use *clock pulses,* which are one of the more prominently advertised feature of new PCs. A PC having a 2-GHz clock rate translates into each clock pulse, if it is a symmetrical square wave, being at a logic 1 and a logic 0 state for 2.5 ns as a 1 and 2.5 ns as a zero. Therefore if two pulses are supposed to arrive at a logic circuit at the same time, or very close to the same time, the path length for the two signals cannot be any different than two to three inches. This can be a very significant design problem for high-speed logic designs.

Combinational Logic

Having defined a way to use voltage levels to physically represent digital numbers, we can apply digital signal theory to design useful circuits. Digital circuits combine binary inputs to produce a desired binary output or combination of outputs. This simple combination of 0s and 1s can become very powerful, implementing everything from simple switches to powerful computers.

A digital circuit falls into one of two types: combinational logic or sequential logic. In a *combinational logic* circuit, the output depends only on the *present inputs*. (If we ignore propagation delay.) In contrast, in a *sequential logic* circuit, the output depends on the present inputs, the *previous sequence of inputs* and often a clock signal. The next section discusses combinational logic circuits. Later, we will build sequential logic circuits from the basics established here.

BOOLEAN ALGEBRA AND THE BASIC LOGICAL OPERATORS

Combinational circuits are composed of logic gates, which perform binary operations. Logic gates manipulate binary numbers, so you need an understanding of the algebra of binary numbers to understand how logic gates operate. *Boolean algebra* is the mathematical system to describe and design binary digital circuits. It is named after George Boole, the mathematician who developed the system. Standard algebra has a set of basic operations: addition, subtraction, multiplication and division. Similarly, Boolean algebra has a set of basic operations, called *logical operations*: NOT, AND and OR.

The function of these operators can be described by either a Boolean equation or a truth table. A Boolean *equation* describes an operator's function by representing the inputs and the operations performed on them. An equation is of the form "B = A," while an *expression* is of the form "A." In an assignment equation, the inputs and operations appear on the right and the result, or output, is assigned to the variable on the left.

A *truth table* describes an operator's function by listing all possible inputs and the corresponding outputs. Truth tables are sometimes written with Ts and Fs (for true and false) or with their respective equivalents, 1s and 0s. In company databooks (catalogs of logic devices a company manufactures), truth tables are usually written with Hs and Ls (for high and low). In the figures, 1 will mean high and 0 will mean low. This representation is called positive logic. The meaning of different logic types and why they are useful is discussed in a later section.

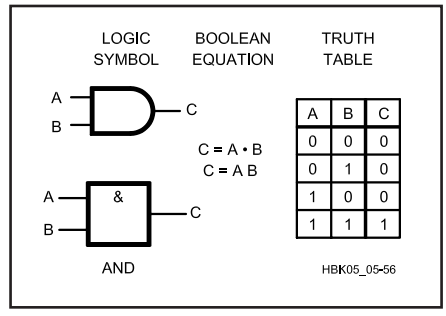


Fig 5.56 — Two-input AND gate.

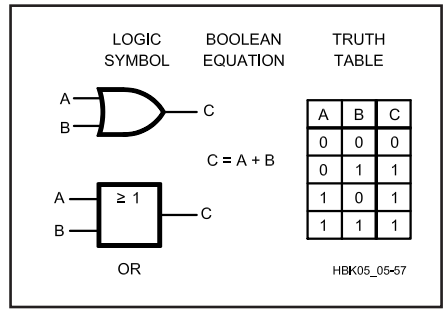


Fig 5.57 — Two-input OR gate.

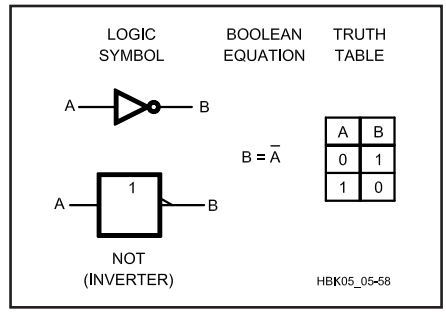


Fig 5.58 — Inverter.

Each Boolean operator also has two circuit symbols associated with it. The traditional symbol — used by ARRL and other US publications — appears on top in each of the figures; for example, the triangle and bubble for the NOT function in **Fig 5.58**. In the traditional symbols, a small circle, or *bubble*, always represents "NOT." (This *bubble* is called a state indicator.) Appearing just below the traditional symbol is the newer ANSI/IEEE Standard symbol. This symbol is always a square box with notations inside it. In these newer symbols, a small flag represents "NOT." The new notation is an attempt to replace the detailed logic drawing of a complex function with a simpler block symbol.

COMMON GATES

Figs **5.56**, **5.57** and 5.58 show the truth tables, Boolean algebra equations and circuit symbols for the three basic Boolean operations: AND, OR and NOT, respectively. All combinational logic functions, no matter how complex, can be described in terms of these three operators. Each truth table can be converted into words. The truth table for the two-input AND gate can be expressed as "the output C is a 1 only when the inputs are both 1's." This can be seen by examining the output column "C" — it remains at a 0 and becomes a 1 only when the input column "A" and the input column "B" are both 1's — the last line of the table.

The NOT operation is also called *inversion*, *negation* or *complement*. The circuit that implements this function is called an *inverter* or *inverting buffer*. The most common notation for NOT is a bar over a variable or expression. For example, NOT A is denoted $\overline{A}$. This is read as either "Not A" or as "A bar." A less common notation

is to denote Not A by A´, which is read as "A prime."

While the inverting buffer and the noninverting buffer covered later have only one input and output, many combinational logic elements can have multiple inputs. When a combinational logic element has two or more inputs and one output, it is called a *gate*. (The term "gate" has many different but specific technical uses. For a clarification of the many definitions of gate, see the section on Synchronicity and Control Signals, later in this chapter.) For simplicity, the figures and truth tables for multiple-input elements will show the operations for only two inputs, the minimum number.

The output of an AND function is 1 only if *all* of the inputs are 1. Therefore, if *any* of the inputs are 0, then the output is 0. The notation for an AND is either a dot (•) between the inputs, as in C = A•B, or nothing between the inputs, as in C = AB. Read these equations as "C equals A AND B."

The OR gate detects if one or more inputs are 1. In other words, if *any* of the inputs are 1, then the output of the OR gate is 1. Since this includes the case where more than one input may be 1, the OR operation is also known as an INCLUSIVE OR. The OR operation detects if *at least one* input is 1. Only if all the inputs are 0, then the output is 0. The notation for an OR is a plus sign (+) between the inputs, as in C = A + B. Read this equation as "C equals A OR B."

Additional Gates

More complex logical functions are derived from combinations of the basic logical operators. These operations — NAND, NOR, XOR and the noninverter or buffer — are illustrated in **Figs 5.59** through **5.62**, respectively. As before, each is described by a truth table, Boolean algebra equation and circuit symbols. Also as before, except for the noninverter, each could have more inputs than the two illustrated.

The NAND gate (short for NOT AND) is equivalent to an AND gate followed by a NOT gate. Thus, its output is the complement of the AND output: The output is a 0 only if all the inputs are 1. If any of the inputs is 0, then the output is a 1.

The NOR gate (short for NOT OR) is equivalent to an OR gate followed by a NOT gate. Thus, its output is the complement of the OR output: If any of the inputs are 1, then the output is a 0. Only if all the inputs are 0, then the output is a 1.

The operations so far enable a designer to determine two general cases: (1) if *all* inputs have a desired state or (2) if *at least one* input has a desired state. The XOR and XNOR gates enable a designer to determine if *one and only one* input of a desired state is present.

The XOR gate (read as EXCLUSIVE OR) has an output of 1 if one and only one of the inputs is a 1 state. The output is 0 otherwise. The symbol for XOR is ⊕. This is easy to remember if you think of the "+" OR symbol enclosed in an "O" for *only one*.

The XOR gate is also known as a "half adder," because in binary arithmetic it does everything but the "carry" operation. The following examples show the possible binary additions for a two-input XOR.

0	0	1	1
0	1	0	1
0	1	1	0

The XNOR gate (read as EXCLUSIVE NOR) is the complement of the XOR gate. The output is 0 if one and only one of the inputs is a 1. The output is 1 either if all inputs are 0 or more than one input is 1.

Buffers

A *noninverter*, also known as a *buffer*, *amplifier* or *driver*, at first glance does not seem to do anything. It simply receives an input and produces the same output. In reality, it is changing other properties of the signal in a useful fashion, such as amplifying the current level. The practical uses of a noninverter include (A) providing sufficient current to drive a number of gates, (B) interfacing between two logic families, (C) obtaining a desired pulse rise time and (D) providing a slight delay to make pulses arrive at the proper time.

Tri-State Gates

Under normal circumstances, a logic element can drive or feed several other logic elements. A typical AND gate might be able to drive or feed 10 other gates. This is known as *fan-out*. However, only one gate output can be connected to a single wire. If you have two possible driving sources to feed one particular wire, some logic network that probably includes a number OR gates must be used. The symbol and truth table for a tri-state gate is in Fig **5.63**.

In each PC data is routed internally on a set of wires called *buses*. A bus consists of a set of wires, and many input logic elements are connected to *listen* on the bus. However the data on the bus can come from many circuits or drivers. To eliminate the need for the network of OR gates to drive each bus wire, as set of gates known as *tri-state* gates are used.

A tri-state gate can be any of the common gates previously described, but with one additional control lead. When this lead is enabled (it can be designed to allow

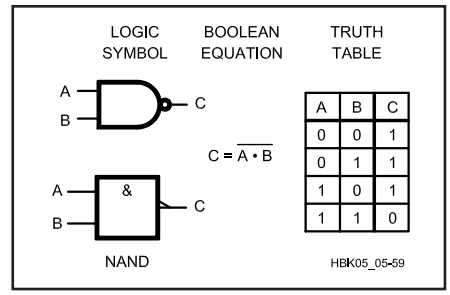


Fig 5.59 — Two-input NAND gate.

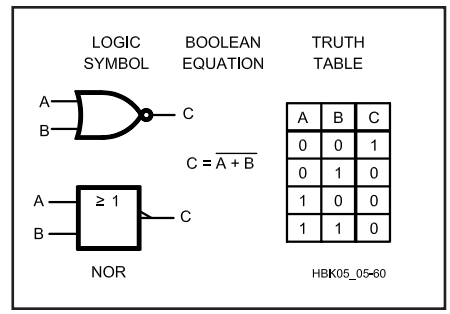


Fig 5.60 — Two-input NOR gate.

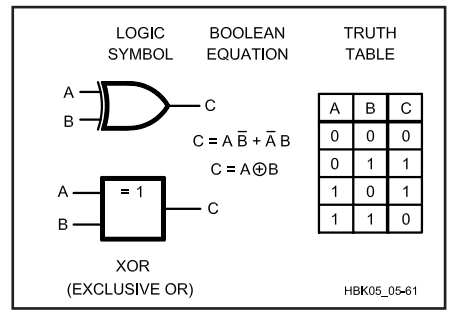


Fig 5.61 — Two-input XOR gate.

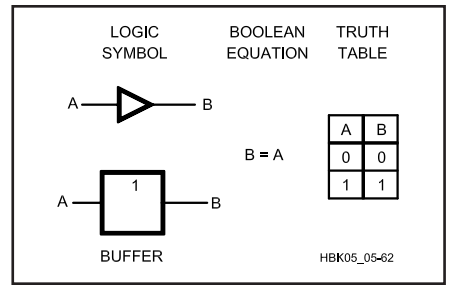


Fig 5.62 — Noninverting buffer.

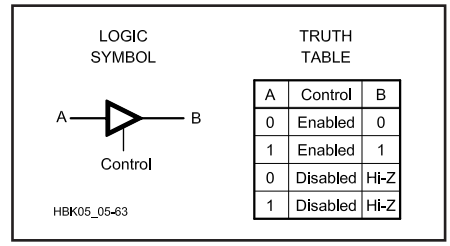


Fig 5.63 — Tri-State Gate.

either a 0 or a 1 to enable it) the gate operates normally, according to the truth table for that type of gate. However, when the gate is not enabled, the output goes to a high impedance, and so far as the output wire is concerned, the gate does not exist.

Each device that has to send data down a bus wire is connected to the bus wire through a tri-state gate. However, as long as only one device, through its tri-state gate, is enabled, it is as though all the other connected tri-state gates do not exist.

BOOLEAN THEOREMS

The analysis of a circuit starts with a logic diagram and then derives a circuit description. In digital circuits, this description is in the form of a truth table or logical equation. The *synthesis*, or design, of a circuit goes in the reverse: starting with an informal description, determining an equation or truth table and then expanding the truth table to components that will implement the desired response. In both of these processes, we need to either simplify or expand a complex logical equation.

To manipulate an equation, we use mathematical *theorems*. Theorems are statements that have been proven to be true. The theorems of Boolean algebra are very similar to those of standard algebra, such as commutivity and associativity. Proofs of the Boolean algebra theorems can be found in an introductory digital design textbook.

BASIC THEOREMS

Table 5.5 lists the theorems for a single variable and **Table 5.6** lists the theorems for two or more variables. These tables illustrate the *principle of duality* exhibited by the Boolean theorems: Each theorem has a dual in which, after swapping all ANDs with ORs and all 1s with 0s, the statement is still true.

The tables also illustrate the *precedence* of the Boolean operations: the order in which operations are performed when not specified by parenthesis. From highest to lowest, the precedence is NOT, AND then OR. For example, the distributive law includes the expression "A + B•C." This is equivalent to "A + (B•C)." The parenthesis around (B•C) can be left out since an AND operation has higher priority than an OR operation. Precedence for Boolean algebra is similar to the convention of standard algebra: raising to a power, then multiplication, then addition.

Table 5.5
Boolean Algebra Single Variable Theorems

Identities:	$A \bullet 1 = A$	$A + 0 = A$
Null elements:	$A \bullet 0 = 0$	$A + 1 = 1$
Idempotence:	$A \bullet A = A$	$A + A = A$
Complements:	$A \bullet \overline{A} = 0$	$A + \overline{A} = 1$
Involution:	$\overline{(\overline{A})} = A$	

Table 5.6
Boolean Algebra Multivariable Theorems

Commutativity:	$A \bullet B = B \bullet A$ $A + B = B + A$
Associativity:	$(A \bullet B) \bullet C = A \bullet (B \bullet C)$ $(A + B) + C = A + (B + C)$
Distributivity:	$(A + B) \bullet (A + C) = A + B \bullet C$ $A \bullet B + A \bullet C = A \bullet (B + C)$
Covering:	$A \bullet (A + B) = A$ $A + A \bullet B = A$
Combining:	$(A + B) \bullet (A + \overline{B}) = A$ $A \bullet B + A \bullet B = A$
Consensus:	$A \bullet B + \overline{A} \bullet C + B \bullet C = A \bullet B + \overline{A} \bullet C$ $(A + B) \bullet (\overline{A} + C) \bullet (B + C) = (A + B) \bullet (\overline{A} + C)$ $A + \overline{A}B = A + B$

DeMorgan's Theorem

One of the most useful theorems in Boolean algebra is DeMorgan's Theorem: $\overline{A \bullet B} = \overline{A} + \overline{B}$ and its dual $\overline{A + B} = \overline{A} \bullet \overline{B}$. The truth table in **Table 5.7** proves these statements. DeMorgan's Theorem provides a way to simplify the complement of a large expression. It also enables a designer to interchange a number of equivalent gates, as shown by **Fig 5.64**.

The equivalent gates show that the duality principle works with symbols the same as it does for Boolean equations: just swap ANDs with ORs and switch the bubbles. For example, the NAND gate — an AND gate followed by an inverter bubble — becomes an OR gate preceded by two inverter bubbles. DeMorgan's Theorem is important because it means any logical function can be implemented using either inverters and AND gates or inverters and OR gates. Also, the ability to change placement of the bubbles using DeMorgan's Theorem is useful in dealing with mixed logic, to be discussed next.

POSITIVE AND NEGATIVE LOGIC

The truth tables shown in the figures in this chapter are drawn for positive logic. In *positive logic*, or *high true*, a higher voltage means true (logic 1) while a lower

Table 5.7
DeMorgan's Theorem

(A) $\overline{A \bullet B} = \overline{A} + \overline{B}$
(B) $\overline{A + B} = \overline{A} \bullet \overline{B}$
(C)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A	B	$\overline{A}$	$\overline{B}$	$A \bullet B$	$\overline{A \bullet B}$	$A + B$	$\overline{A + B}$	$\overline{A} \bullet \overline{B}$	$\overline{A} + \overline{B}$
0	0	1	1	0	1	0	1	1	1
0	1	1	0	0	1	1	0	0	1
1	0	0	1	0	1	1	0	0	1
1	1	0	0	1	0	1	0	0	0

(A) and (B) are statements of DeMorgan's Theorem. The truth table at (C) is proof of these statements: (A) is proven by the equivalence of columns 6 and 10 and (B) by columns 8 and 9.

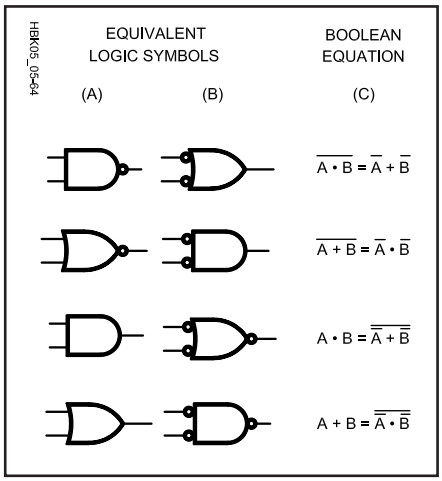


Fig 5.64 — Equivalent gates from DeMorgan's Theorem: Each gate in column A is equivalent to the opposite gate in column B. The Boolean equations in column C formally state the equivalences.

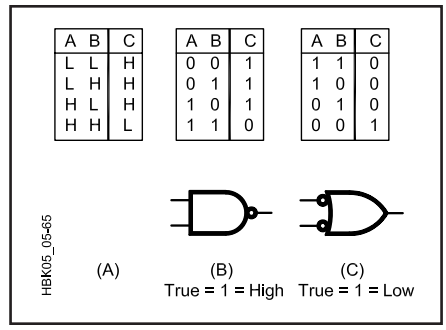


Fig 5.65 — (A) A general truth table, (B) a truth table and NAND symbol for positive logic and (C) a truth table and NOR symbol for negative logic.

voltage means false (logic 0). This is also referred to as *active high*: a signal performs a named action or denotes a condition when it is "high" or 1. In *negative logic*, or *low true*, a lower voltage means true (1) and a higher voltage means false (0). An *active low* signal performs an action or denotes a condition when it is "low" or 0.

In both logic types, true = 1 and false = 0; but whether true means high or low differs. Company databooks are drawn for general truth tables: an "H" for high and an "L" for low. (Some tables also have an "X" for a "don't care" state.) The function of the table can differ depending on whether it is interpreted for positive logic or negative logic.

Device data sheets often show positive logic convention, or same is assumed. However, a signal into an IC is represented with a bar above it, indicating that the "enable" on that wire is active low — it does ***not*** *mean negative logic (0 volts = a logical 1) is used! Similarly a bubble on the input of a logic element also usually means active low. These can be sources of confusion.*

Fig 5.65 shows how a general truth table differs when interpreted for different logic types. The same truth table gives two equivalent gates: positive logic gives the function of a NAND gate while negative logic gives the function of a NOR gate.

Note that these gates correspond to the equivalent gates from DeMorgan's Theorem. A bubble on an input or output terminal indicates an active low device. The absence of bubbles indicates an active high device.

Like the bubbles, signal names can be used to indicate logic states. These names can aid the understanding of a circuit by indicating control of an action (GO, /ENABLE) or detection of a condition (READY, /ERROR). The action or condition occurs when the signal is in its active state. When a signal is in its active state, it is called *asserted*; a signal not in its active state is called *negated* or *deasserted*. A prefix can easily indicate a signal's active state: active low signals are preceded by a "/," like /READY, while active high signals have no prefix. Standard practice is that the signal name and input pin match (have the same active level). For example, an input with a bubble (active low) may be called /READY while an input with no bubble (active high) is called READY. Output signal names should always match the device output pin.

In this chapter, positive logic is used unless indicated otherwise. Although using mixed logic can be confusing, it does have some advantages. Mixed logic combined with DeMorgan's Theorem can promote more effective use of available gates. Also, well-chosen signal names and placement of bubbles can promote more understandable logic diagrams.

Sequential Logic

The previous section discussed combinational logic, whose outputs depend only on the present inputs. In contrast, in *sequential logic* circuits, the new output depends not only on the present inputs but also on the present outputs. The present outputs depended on the previous inputs and outputs and those earlier outputs depended on even earlier inputs and outputs and so on. Thus, the present outputs depend on the previous *sequence of inputs* and the system has *memory*. Having the outputs become part of the new inputs is known as *feedback*.

This section first introduces a number of terms necessary to understand sequential logic: types of synchronicity, types of control signals and ways to illustrate circuit function. Numerous sequential logic circuits are then introduced. These circuits provide an overview of the basic sequential circuits that are commercially available. Depending on your approach to learning, you may choose to either (1) read the material in the order presented, definitions then examples, or (2) start with the example circuits, which begin with the flip-flop, referring back to the definitions as needed.

SYNCHRONICITY AND CONTROL SIGNALS

When a combinational circuit is given a set of inputs, the outputs take on the expected values after a propagation delay during which the inputs travel through the circuit to the output. In a sequential circuit, however, the travel through the circuit is more complicated. After application of the first inputs and one propagation delay, the outputs take on the resulting state; but then the outputs start trickling back through and, after a second propagation delay, new outputs appear. The same happens after a third propagation delay. With propagation delays in the nanosecond range, this cycle around the circuit is rapidly and continually generating new outputs. A user needs to know when the outputs are valid.

There are two types of sequential circuits: synchronous circuits and asynchronous circuits, which are analyzed differently for valid outputs. In *asynchronous* operation, the outputs respond to the inputs immediately after the propagation delay. To work properly, this type of circuit must eventually reach a *stable* state: the inputs and the fed back outputs result in the new outputs staying the same. When the nonfeedback inputs are changed, the feedback cycle needs to eventually reach a new stable state. Generally, the output of this type of logic is not valid until the last input has changed, and enough time has elapsed for all propagation delays to have occurred.

In *synchronous* operation, the outputs change state only at specific times. These times are determined by the presence of a particular input signal: a clock, toggle, latch or enable. Synchronicity is important because it ensures proper timing: all the inputs are present where needed when the control signal causes a change of state.

Some authors vary the meanings slightly for the different control signals. The following is a brief illustration of common uses, as well as showing uses for noun, verb and adjective. *Enabling* a circuit generally means the control signal goes to its asserted level, allowing the circuit to change state. *Latch* implies memory: (noun) a circuit that stores a bit of information or (verb) to hold at the same output state. *Gate* has many meanings, some unrelated to synchronous control: (A) a signal used to trigger the passage of other signals through a circuit (for example, "A gate circuit passes a signal only when a gating pulse is present."), (B) any logic circuit with two or more inputs and one output (used earlier in this chapter) or (C) one of the electrodes of an FET (as described in the analog portion of this chapter). To *toggle* means a signal changes state, from 1 to 0 or vice versa. A *clock* signal is one that toggles at a regular rate.

Clock control is the most common method, so it has some additional terms, illustrated by **Fig 5.66**. The *clock period* is the time between successive transitions in the same direction; the *clock frequency* is the reciprocal of the period. A *pulse* or *clock tick* is the first edge in a clock period, or sometimes the period itself or the first half of the period. The *duty cycle* is the percentage of time that the clock signal is at its asserted level. A common application of the use of clock pulses is to limit the input to a logic circuit such that the circuit is only enabled on one clock phase; that is the inputs occur before the clock changes to a logic 1, The outputs are sampled only after this point; perhaps when the clock next changes back to a logic 0.

The reaction of a synchronous circuit to its control signal is *static* or *dynamic*. Static, *gated* or *level-triggered* control allows the circuit to change state whenever the control signal is at its active or asserted level. Dynamic, or *edge-triggered*, control allows the circuit to change state only when the control signal *changes* from unasserted to asserted. By convention, a control signal is active high if state changes occur when the signal is high or at the rising edge and active low in the opposite case. Thus, for positive logic, the convention is enable = 1 or enable goes from 0 to 1. This transition from 0 to 1 is called

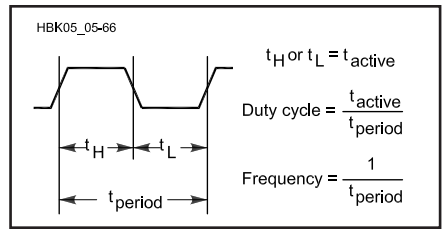


Fig 5.66 — Clock signal terms. The duty cycle would be t_H / t_{PERIOD} for an active high signal and t_L / t_{PERIOD} for an active low signal.

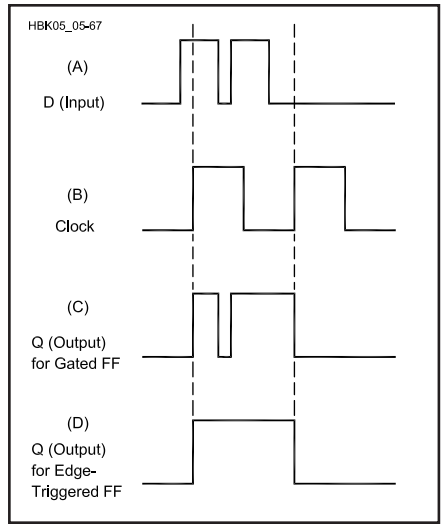


Fig 5.67 — Level-triggered vs edge-triggered for a D flip-flop: (A) input D, (B) clock input, (C) output Q for level-triggered: circuit responds whenever clock is 1. (D) output Q for edge-triggered: circuit responds only at rising edge of clock. Notice that the short negative pulse on the input D is not reproduced by the edge-triggered flip-flop.

positive edge-triggered and is indicated by a small triangle inside the circuit box. A circuit responding to the opposite transition, from 1 to 0, is called *negative edge-triggered*, indicated by a bubble with the triangle. Whether a circuit is level-triggered or edge-triggered can affect its output, as shown by **Fig 5.67**. Input D includes a very brief pulse, called a *glitch*, which may be caused by noise. The differing results at the output illustrate how noise can cause errors.

CIRCUIT DESIGN — FLIP-FLOPS

Flip-flops are the basic building blocks of sequential circuits. A *flip-flop* is a device with two stable states: the *set* state (1) or the *reset* state (0). (The reset state is also called the *cleared* state.) The flip-flop can be placed in one or the other of the two states by applying the appropriate input. (Since a common use of flip-flops is to store one bit of information, some use the term *latch* interchangeably with flip-flop. A set of latches, or flip-flops holding an n-bit number is called a register.) While gates have special symbols, the schematic symbol for most components is a rectangular box with the circuit name or abbreviation, the signal names and assertion bubbles. For flip-flops, the circuit name is usually omitted since the signal names are enough to indicate a flip-flop and its type. The four basic types of flip-flops are the S-R, D, T and J-K. The most common flip-flops available to Amateurs today are the J-K and D- flip-flops; the others can be synthesized by utilizing these two varieties.

Triggering a Flip-Flop

Although the S-R (Set-reset) flip flop is no longer generally available or used, it does provide insight in basic flip-flop operations and triggering. In **Fig 5.68** the symbol for an S-R flip flop and its truth table are accompanied by a logic implementation, using NAND gates. As the truth table shows, this basic implementation requires a positive or logic 1 input on the set input to put the flip-flop in the Q or set state. Remove the input, and the flip flop stays in the Q state, which is what is expected of a flip-flop. Not until the S input receives a logic 1 input does the flip-

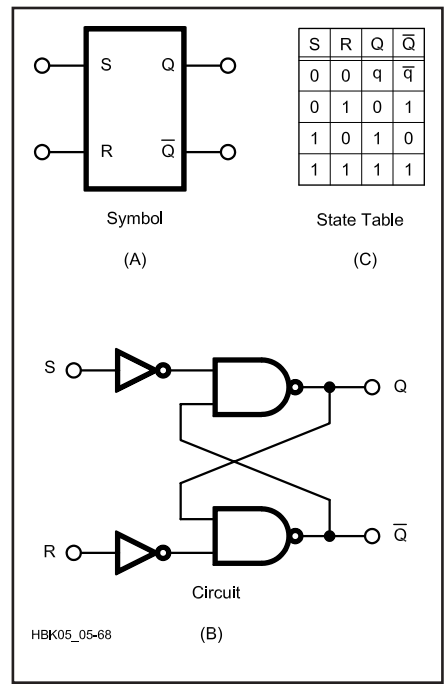


Fig 5.68 — Unclocked S-R Flip-Flop. (A) schematic symbol. (B) circuit diagram. (C) state table or truth table.

flop change state and go to the reset or Q=0 state.

Note that the input can be a short pulse or a level; as long as it is there for some minimum duration, the flip-flop will respond. By contrast the clocked S-R flip-flop in **Fig 5.69** requires both a positive level to be present at either the S or R inputs and a positive clock pulse; the clock pulse is ANDed with the S or R input to trigger the flip–flop. In this case the flip-flop shown is implemented with a set of NOR gates.

A final triggering method is edge triggering. Here, instead of using the clock pulse as shown in the timing diagram of Fig 5.69, just the edge of the clock pulse is used. This will be discussed more in the section on J-K flip-flops. The edge-triggered flip-flop helps solves a problem with noise. Edge-triggering minimizes the time during which a circuit responds to its inputs: the chance of a glitch occurring during the nanosecond transition of a clock pulse is remote. A side benefit of edge-triggering is that only one new output is produced per clock period. Edge-triggering is denoted by a small rising-edge or falling-edge symbol in the clock column of the flip-flop's truth table. It can also appear, instead of the clock triangle, inside the schematic symbol.

Although this description uses positive levels and positive clock pulses, other implementations of these flip-flops can be made with negative levels, negative clock pulses or combinations of positive and negative levels and pulses. This will again be illustrated in the section on J-K flip-flops.

Master/Slave Flip-Flop

One major problem with the simple flip-flop shown up to now is the question of when is there a valid output. Suppose a flip-flop receives input that causes it to change state; at the same time the output of this flip-flop is being sampled to control some other logic element. There is a real risk here that the output will be sampled just as it is changing and thus the validity of the output is questionable.

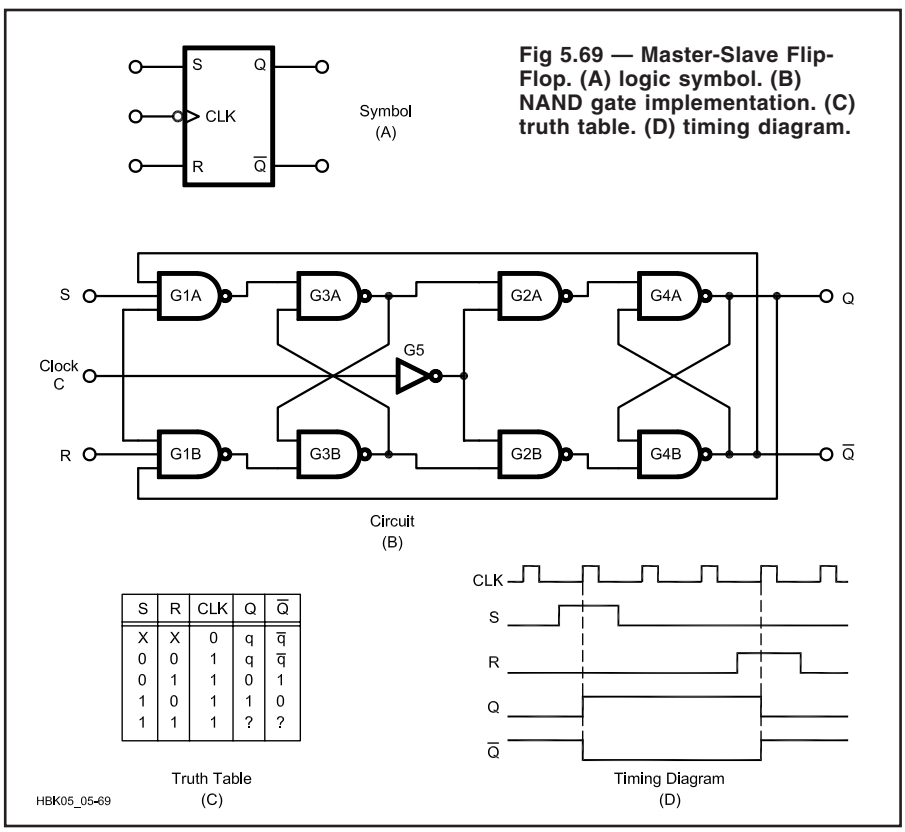
A solution to this problem is a circuit that samples and stores its inputs before changing its outputs. Such a circuit is built by placing two flip-flops in series; both flip-flops are triggered by a common clock but an inverter on the second flip-flop's clock input causes it to be asserted only when the first flip-flop is not asserted. The action for a given clock pulse is as follows: The first, or master, flip-flop can change only when the clock is high, sampling and storing the inputs. The second, or slave, flip-flop gets its input from the master and changes when the clock is low. Hence, when the clock is 1, the input is sampled; then when the clock becomes 0, the output is generated. Note that a bubble may appear on the schematic symbol's clock input, reminding us that the output appears when the clock is asserted low. This is conventional for TTL-style J-K flip-flops, but it can be different for CMOS devices.

The master/slave method isolates output changes from input changes, eliminating the problem of series-fed circuits. It also ensures only one new output per clock period, since the slave flip-flop responds to only the single sampled input. A problem can still occur, however, because the master flip-flop can change more than once while it is asserted; thus, there is the potential for the master to sample at the wrong time. There is also the potential that either flip-flop can be affected by noise.

A master-slave, S-R clocked input flip-flop synthesized from NAND gates, Fig 5.69B, is accompanied by its logic symbol, Fig 5.69A. From the logic symbols you can tell that the output changes on a negative-going clock edge.

G3A and G3B form the master set-rest flip-flop, and G4A and G4B the slave flip flop. The input signals S and R are controlled by the positive going edge of the clock through gates G1A and G1B. G2A and G2B control the inputs into the slave flip-flop; these inputs are the outputs of the master flip-flop. Note G5 inverts the clock; thus while the positive-going edge places new data into the master flip-flop, the other edge of the clock transfers the output of the master into the slave on the following negative clock edge.

Table 5.8 provides a summary of the two basic flip-flops currently readily available, the D flip-flop and the J-K flip-flop. The S-R (Set-Reset) used for illustration previously and the T or toggle flip-flop can be synthesized by using the J-K or D flip-flops. The T flip-flop is primarily used to count. With proper input, it changes state for each input pulse – alternating between Q = 1 and Q = 0 each time a clock pulse appears.



S	R	CLK	Q	$\overline{Q}$
X	X	0	q	$\overline{q}$
0	0	1	q	$\overline{q}$
0	1	1	0	1
1	0	1	1	0
1	1	1	?	?

Truth Table
(C)

Fig 5.69 — Master-Slave Flip-Flop. (A) logic symbol. (B) NAND gate implementation. (C) truth table. (D) timing diagram.

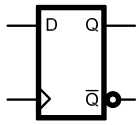
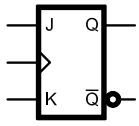
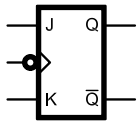
D Flip-Flop

In a D (data) flip-flop, the *data* input is transferred to the outputs when the flip-flop is enabled. The logic level at input D is transferred to Q when the clock is positive; the Q output retains this logic level until the next positive clock pulse (see **Fig 5.70**). The truth table summarizes this operation. If D = 1 the next clock pulse makes Q = 1. If D = 0, the next clock pulse makes Q = 0. A D flip-flop is useful to store one bit of information. A collection of D flip-flops forms a register.

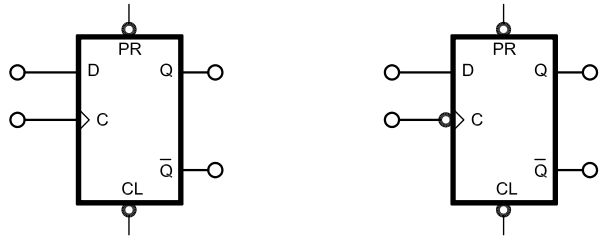
Table 5.8

Summary of Standard Flip-Flops

q = current state Q = next state X = don't care

Flip-Flop Type	Symbol	Truth Table	Characteristic Equation	Excitation Table
D		D CLK | Q X ⌟ | q 0 ⌟ | 0 1 ⌟ | 1	Q = D • CLK	q Q | D CLK 0 0 | X ⌟ 0 1 | 1 ⌟ 1 0 | 0 ⌟ 1 1 | X ⌟
J K		J K CLK | Q X X ⌟ | q 0 0 ⌟ | q 0 1 ⌟ | 0 1 0 ⌟ | 1 1 1 ⌟ | t	$Q = (J \cdot \bar{q} + \bar{K} \cdot q) \cdot CLK$ Positive Edge Clock	q Q | J K CLK 0 0 | 0 X ⌟ 0 1 | 1 X ⌟ 1 0 | X 1 ⌟ 1 1 | X 0 ⌟
J K		J K CLK | Q 0 0 ⌝ | 0 0 1 ⌝ | 0 1 0 ⌝ | 1 1 1 ⌝ | t	$Q = (J \cdot \bar{q} + \bar{K} \cdot q) \cdot CLK$ Negative Edge Clock	q Q | J K CLK 0 0 | 0 X ⌝ 0 1 | 1 X ⌝ 1 0 | X 1 ⌝ 1 1 | X 0 ⌝

HBK05_05-Tab08 t: If J = K, the clock toggles the flip-flop



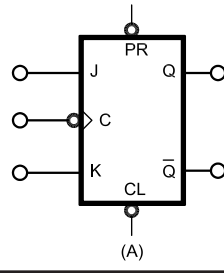
X=1 or 0

Clear	Preset	Clock	D	Q	Q̄	
0	1	X	X	0	1	
1	0	X	X	1	0	
0	0	X	X	1	1	Unused State
1	1	⌐	1	1	0	
1	1	⌐	0	0	1	

Truth Table for a positive edge-triggered D flip-flop

HBK05_05-70 (C)

Fig 5.70 — (A & B) The D flip-flop. (C) A truth table for the positive edge-triggered D flip-flop.



Preset	Clear	J	K	C	Q	Q̄	
0	1	X	X	X	1	0	
1	0	X	X	X	0	1	
0	0	X	X	X	1	1	Unused State
1	1	0	1	⌐	0	1	
1	1	1	0	⌐	1	0	
1	1	0	0	X	Q	Q̄	Unchanged State
1	1	1	1	⌐	Toggle		

JK flip-flop truth table

(B) HBK05_05-71

Fig 5.71 — (A) JK flip-flop. (B) JK flip-flop truth table.

J-K Flip-Flop

The most readily available flip-flop today is the J-K flip-flop, shown schematically in **Fig 5.71A**. It has five inputs, and the unit shown uses both positive active (the J and K inputs) and negative active inputs (notes the "bubbles" on the C or clock, PR or preset and CL or clear inputs). With these inputs almost any other type of flip-flop may be synthesized.

The truth table of Fig 5.71B provides an explanation. Lines 1 and 2 show the preset and clear inputs and their use. These are active low, meaning that when one (and only one) of them goes to a logic 0, the flip-flop responds, just as if it was a S-R or set-reset flip-flop. Make PR a logic 0, and leave CL a logic 1, and the flip-flop goes into the Q = 1 state (line 1). Do the reverse (line 2) – PR = 1, CL = 0 and the flip-flop goes into a Q' = 1 state. When these two inputs are used, J, K and C are marked as X or don't care, because the PR and CL inputs override them. Line 3 corresponds to the unused state of the R-S flip-flop.

Lines 4 and 5 show that if J = 1 and K = 0, the next clock transition from high to low sets Q = 1 and Q' = 0. Alternately, J = 0 and K = 1 sets Q = 0 and Q' = 1. Therefore if a signal is applied to J, and the inverted signal sent to K, the J-K flip-flop will mimic a D flip-flop, echoing its input.

The most unique feature of the J-K flip-flop is line 7. If both J and K are connected to a 1, then each clock 1 to 0 transition will flip or toggle the flop-flop. Thus the J-K flip-flop can be used as a T flip-flop, as in a ripple counter (see the following COUNTERS section.)

Summary

Only the D and J-K flip-flops are generally available as commercial integrated circuit chips. Since memory and temporary storage are so often desirable, the D flip-flop is manufactured as the simplest way to provide memory. When more functionality is needed, the J-K flip-flop is available. The J-K flip-flop (using its PR and CL inputs) can substitute for an S-R flip-flop and a T flip-flop can be created from either the D or J-K flip-flop.

GROUPS OF FLIP-FLOPS

Counters

Groups of flip-flops can be combined to make counters. Intuitively, a counter is a circuit that starts at state 0 and sequences up through states 1, 2, 3, to m, where m is the maximum number of states available. From state m, the next state will return the counter to 0. This describes the most common counter: the *n-bit binary counter*, with n outputs corresponding to 2^n = m states.

Such a counter can be made from n flip-flops, as shown in **Fig 5.72**. This figure shows implementations for each of the types of synchronicity. Both circuits pass the data count from stage to stage. In the asynchronous counter, Fig 5.72A, the clock is also passed from stage to stage and the circuit is called *ripple* or *ripple-carry*.

The J-K flip-flop truth table shows that with PR (Preset) and CL (Clear) both positive, and therefore not effecting the operation, the flip-flop will toggle if J and K are tied to a logic 1. In Fig 5.72A the first stage has its J and K inputs permanently tied to a logic 1, and each succeeding stage has its J and K inputs tied to Q of the proceeding stage. This provides a direct ripple counter implementation.

Design of a synchronous counter is bit more involved. It consists of determining, for a particular count, the conditions that will make the next stage change at the same clock edge when all the stages are changing.

To illustrate this, notice the binary counting table of Fig 5.72. The right-hand column represents the lowest stage of the counter. It alternates between 1 and 0 on every line. Thus, for the first stage the J and K inputs are tied to logic 1. This provides the alternation required by the counting table.

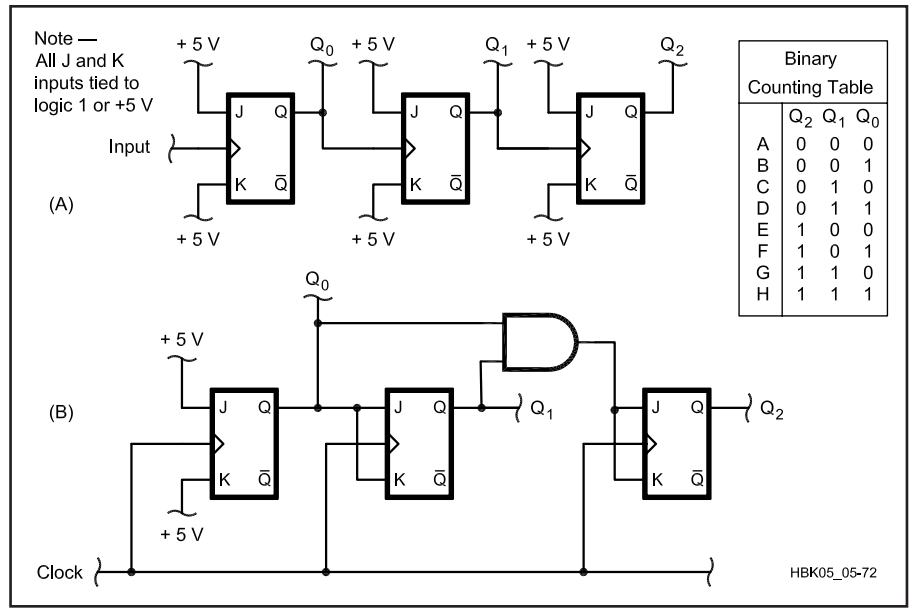
The middle column or second stage of the counter changes state right after the lower stage is a one (lines C, E and G). Thus if the Q output of the lowest stage is tied to the J and K inputs of the second stage, each time the output of the lowest stage is a 1 the second stage toggles on the next clock pulse.

Finally, the third column (third stage) toggles when both the first stage and the second stage are both 1s (line D). Thus by ANDing the Q outputs of the first two stages, and then connecting them to the J and K inputs of the third stage, the third stage will toggle whenever the first two stages are 1s.

There are formal methods for determining the wiring of synchronous counters. The illustration above is one manual method that may be used to design a counter of this type. The advantage of the synchronous counter is that at any instant, except during clock pulse transition, all counter stage outputs *are correct* and delay due to propagation through the flip flops is not a problem.

In the synchronous counter, Fig 5.72B, each stage is controlled by a common clock signal.

There are numerous variations on this first example of a counter. Most counters have the ability to *clear* the count to 0. Some counters can also *preset* to a desired



Binary Counting Table			
	Q_2	Q_1	Q_0
A	0	0	0
B	0	0	1
C	0	1	0
D	0	1	1
E	1	0	0
F	1	0	1
G	1	1	0
H	1	1	1

Fig 5.72 — Three-bit binary counter using J-K flip-flops: (A) asynchronous or ripple, (B) synchronous.

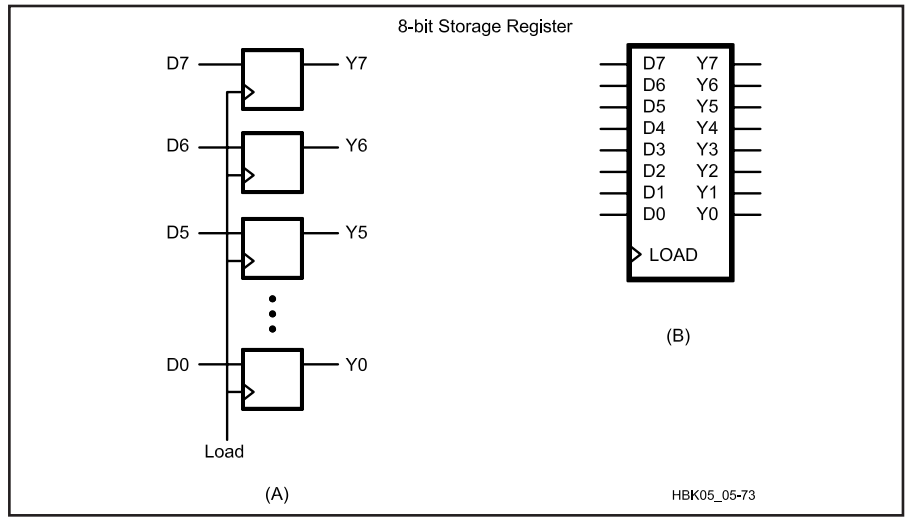


Fig 5.73 — An eight-bit storage register: (A) circuit and (B) schematic symbol.

count. The clear and preset control inputs are often asynchronous — they change the output state without being clocked. Counters may either count up (increment) or down (decrement). *Up/down* counters can be controlled to count in either direction. Counters can have sequences other than the standard numbers, for example a BCD counter.

Counters are also not restricted to changing state on every clock cycle. An n-bit counter that changes state only after m clock pulses is called a *divider* or *divide-by-m* counter. There are still $2^n = m$ states; however, the output after p clock pulses is now p / m. Combining different divide-by-m counters can result in almost any desired count. For example, a base 12 counter can be made from a divide-by-2 and a divide-by-6 counter; a base 10 (decade) counter consists of a divide-by-2 and a BCD divide-by-5 counter.

The outputs of these counters are binary. To produce output in decimal form, the output of a counter would be provided to a binary-to-decimal decoder chip and/or an LED display.

Registers

Groups of flip-flops can be combined to make registers, usually implemented with D flip-flops. A *register* stores n bits of information, delivering that information in response to a clock pulse. Registers usually have asynchronous *set* to 1 and *clear* to 0 capabilities.

Storage Register

A storage register simply stores temporary information, for example, incoming information or intermediate results. The size is related to the basic size of information handled by a computer: 8 flip-flops for an 8-bit or *byte register* or 16 bits for a *word register*. **Fig 5.73** shows a typical circuit and schematic symbols for an 8-bit storage register. In (C), although the bits are passed on 8 separate lines (from 8 flip-flops), a slash and number, "/8," is used to simplify the symbol. Storage registers are important to computer architecture; this topic is discussed in depth later in the chapter.

Shift Register

Shift registers also store information and provide it in response to a clock signal, but they handle their information differently: When a clock pulse occurs, instead of each flip-flop passing its result to the output, the flip-flops pass their data to each other, up and down the row. For example, in up mode, each flip-flop receives the output of the preceding flip-flop. A data bit starting in flip-flop D0 in a left shifter would move to D1, then D2 and so on until it is shifted out of the register. If a 0 was input to the least significant bit, D0, on each clock pulse then, when the last data bit has been shifted out, the register contains all 0s.

Shift registers can be left shifters, right shifters or controlled to shift in either direction. The most general form, a *universal shift register*, has two control inputs for four states: Hold, Shift right, Shift left and Load. Most also have asynchronous inputs for preset, clear and parallel load. The primary use of shift registers is to convert parallel information to serial or vice versa. This is useful in interfacing between devices, and is discussed in detail in the Interfacing section.

Additional uses for a shift register are to (1) delay or synchronize data, (2) multiply or divide a number by a factor 2^n or (3) provide random data. Data can be delayed simply by taking advantage of the Hold feature of the register control inputs. Multiplication and division with shift registers is best explained by example: Suppose a 4-bit shift register currently has the value 1000 = 8. A right shift results in the new parallel output 0100 = 4 = 8 / 2. A second right shift results in 0010 = 2 = (8 / 2) / 2. Together the 2 right shifts performed a division by 2^2. In general, shifting right n times is equivalent to dividing by 2^n. Similarly, shifting left multiplies by 2^n. This can be useful to compiler writers to make a computer program run faster. Random data is provided via a ring counter. A *ring counter* is a shift register with its output fed back to its input. At each clock pulse, the register is shifted up or down and some of the flip-flops feedback to other flip-flops, generating a random binary number. Shift registers with several feedback paths can be used as a *pseudorandom number generator*, where the sequence of bits output by the generator meets one or more mathematical criteria for randomness.

MULTIVIBRATORS

Multivibrators are a general type of circuit with three varieties: bistable, monostable and astable. The only truly digital multivibrator is bistable, having two stable states. The flip-flop is a *bistable multivibrator*: both of its two states are stable; it can be triggered from one stable state to the other by an external signal. The other two varieties of multivibrators are partly analog circuits and partly digital. While their output is one or more pulses, the internal operation is strictly analog.

Monostable Multivibrator

A *monostable* or *one-shot* multivibrator has one energy-storing element in its feedback paths, resulting in one stable and one quasi-stable state. It can be switched, or *triggered*, to its quasi-stable state; then returns to the stable state after a time delay. Thus, when triggered, the one-shot multivibrator puts out a pulse of some duration, T.

A very common integrated circuit used for non-precision generation of a signal pulse is the 555 timer IC. **Fig 5.74** shows a 555 connected as a one-shot multivibrator. The one-shot is activated by a negative-going pulse between the trigger input and ground. The trigger pulse causes the output (Q) to go positive and capacitor C to charge through resistor R. When the voltage across C reaches two-thirds of V_{CC}, the capacitor is quickly discharged to ground and the output returns to 0. The output remains at logic 1 for a time determined by T = 1.1 RC, where R is the resistance in ohms and C is the capacitance in farads.

A very common, but again, non-precision application of this circuit is the generation a delayed pulse. If there is a requirement to generate a 50-microsecond pulse, but delayed from a trigger by 20 ms, two 555s might be used. The first 555, configured as an astable multivibrator, generates the 20-ms pulse, and the trailing edge of the 20-ms pulse is used to trigger a second 555 that in turn generates the 10 microsecond pulse.

Astable Multivibrator

An *astable* or *free-running* multivibrator has two energy-storing elements in its feedback paths, resulting in two quasi-stable states. It continuously switches between these two states without external excitation. Thus, the astable multivibrator puts out a sequence of pulses. By properly selecting circuit components, these pulses can be of a desired frequency and width.

Fig 5.75 shows a 555 timer IC connected as an astable multivibrator. The

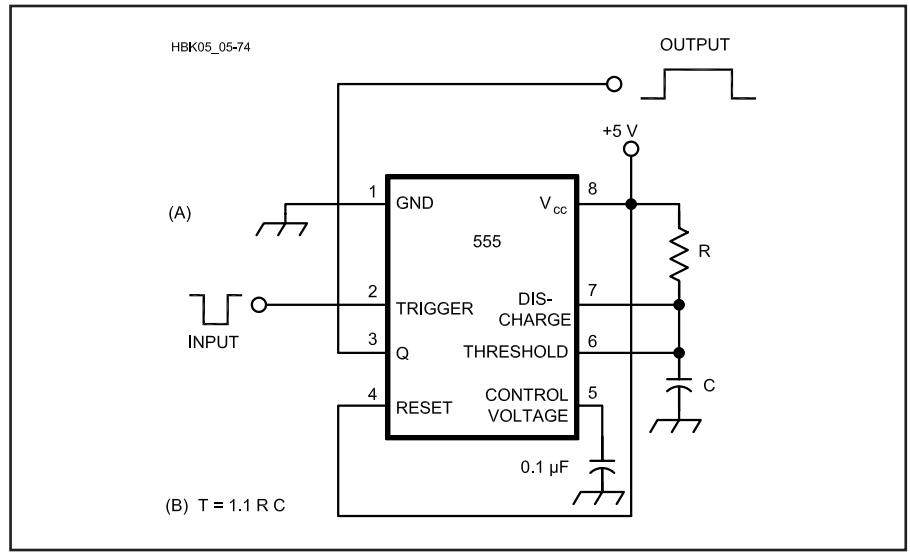


Fig 5.74 — (A) A 555 timer connected as a monostable multivibrator. (B) The equation to calculate values for R in ohms and C in farads, where T is the pulse duration in seconds.

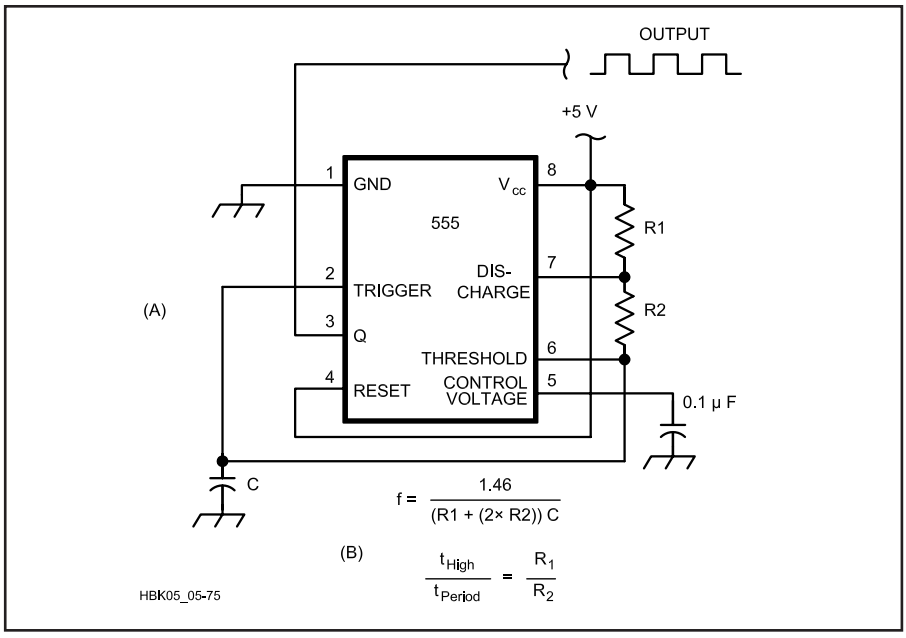


Fig 5.75 — (A) A 555 timer connected as an astable multivibrator. (B) The equations to calculate values for R1, R2 in ohms and C in farads, where f is the clock frequency in Hertz.

capacitor C charges to two-thirds V_{CC} through R1 and R2 and discharges to one-third V_{CC} through R2. The ratio R1 : R2 sets the asserted high duty cycle of the pulse: t_{HIGH} / t_{PERIOD}. The output frequency is determined by:

$$f = \frac{1.46}{(R1 + 2\,R2)C}$$

where:

R1 and R2 are in ohms,
C is in farads and
f is in hertz.

It may be difficult to produce a 50% duty cycle due to manufacturing tolerance for the resistors R1 and R2. One way to ensure a 50% duty cycle is to run the astable multivibrator at 2f and then divide by 2 with a toggle flip-flop.

Astable multivibrators, and the 555 integrated circuit in particular, are very often used to generate clock pulses. Although this is a very inexpensive and minimum hardware approach, the penalty is stability with temperature. Since the frequency and the pulse dimensions are set by resistors and capacitors, drift with temperature and to some extent aging of components will result in changes with time. However, this is no different than the problem faced by designers of L-C controlled VFOs.

SUMMARY

Digital logic plays an increasingly important role in Amateur Radio. Most of this logic is binary and can be described and designed using Boolean algebra. Using the NOT, AND and OR gates of combinational logic, designers can build sequential logic circuits that have memory and feedback. The simplest sequential logic circuit is called a flip-flop. By using control inputs, a flip-flop can latch a data value, retaining one bit of information and acting as memory. Combinations of flip-flops can form useful circuits such as counters, storage registers and shift registers. While this section discussed discrete logic, synthesized with available integrated circuits, most new commercial designs begin with a gate/flip-flop design and then use automated tools to build the entire system or major parts of it with programmable logic devices (PLDs) or equivalent custom integrated circuits.

Digital Integrated Circuits

Integrated circuits (ICs) are the cornerstone of digital logic devices. Modern technology has enabled electronics to become miniature in size and less expensive. Today's complex digital equipment would be impossible with vacuum tubes or even with discrete transistors.

An IC is a miniature electronic module of components and conductors manufactured as a single unit. All you see is a ceramic or black plastic package and the silver-colored pins sticking out. Inside the package is a piece of material, usually silicon, created (fabricated) in such a way that it conducts an electric current to perform logic functions, such as a gate, flip-flop or decoder.

As each generation of ICs surpassed the previous one, they became classified according to the number of gates on a single chip. These classifications are roughly defined as:

Small-scale integration (SSI):
10 or fewer gates on a chip.
Medium-scale integration (MSI):
10-100 gates.
Large-scale integration (LSI):
100-1000 gates.
Very-large-scale integration (VLSI):
1000 or more gates.

This chapter will primarily deal with SSI ICs, the basic digital building blocks. Microprocessors, memory chips and programmable logic devices are discussed later in the Computer Hardware section of this chapter.

The previous section discussed the design of a digital circuit. To build that circuit, the designer must choose between IC chips available in various logic families. Each family and subfamily has its own desirable characteristics. This section reviews the primary IC logic families of interest to radio amateurs. The designer may also be challenged to interface between different logic families or between a logic device and a peripheral device. The former is discussed at the end of this section; the latter with Computer Hardware, later in the chapter.

COMPARING LOGIC FAMILIES

When selecting devices for a circuit, a designer is faced with choosing between many families and subfamilies of logic ICs. The determination of which logic subfamily is right for a specific application is based upon several desirable characteristics: logic speed, power consumption, fan-out, noise immunity and cost. From a practical

view-point, the primary integrated circuit families available from most suppliers today are the TTL and CMOS ICs. Within these families, there are tradeoffs that can be made with respect to individual circuit capabilities, especially in the areas of speed and power consumption. Except under the most demanding circumstances, normal commercial grade temperature rating will do for amateur service. However, at a premium price and with perhaps some problems in availability, military temperature grade equivalent circuits can be selected.

Fan-out

A gate output can supply only a limited amount of current. Therefore, a single output can only drive a limited number of inputs. The measure of driving ability is called fan-out, expressed as the number of inputs (of the same subfamily) that can be driven by a single output. If a logic family that is otherwise desirable does not have sufficient fan-out, consider using noninverting buffers to increase fan-out, as shown by **Fig 5.76**.

Noise Immunity

The noise margin was illustrated in Fig 5.53. The choice of voltage levels for the binary states determines the noise margin. If the gap is too small, a spurious signal can too easily produce the wrong state. Too large a gap, however, produces longer, slower transitions and thus decreased switching speeds.

Circuit impedance also plays a part in noise immunity, particularly if the noise is from external sources such as radio transmitters. At low impedances, more energy is needed to change a given voltage level than at higher impedances.

BIPOLAR LOGIC FAMILIES

Two broad categories of digital logic ICs are *bipolar* and *metal-oxide semiconductor* (MOS). Numerous manufacturing techniques have been developed to fabricate each type. Each surviving, commercially available family has its particular advantages and disadvantages and has found its own special niche in the market.

Bipolar semiconductor ICs usually employ NPN junction transistors. (Bipolar ICs can be manufactured using PNP transistors, but NPN transistors make faster circuits.) While early bipolar logic was faster and had higher power consumption than MOS logic, these distinctions have blurred as manufacturing technology has developed. There are several families of bipolar logic devices, and within some of these families there are subfamilies. The most-used digital logic family is Transistor-Transistor Logic (TTL). Another bipolar logic family, Emitter Coupled Logic (ECL), has exceptionally high speed but high power consumption.

Transistor-Transistor Logic (TTL)

The TTL family has seen widespread acceptance because it is fast and has good noise immunity. It is by far the most commonly used logic family. TTL levels were shown earlier in Fig 5.53: An input voltage between 0.0-0.4 V will represent LOW and an input voltage between 2.4-5.0 V will represent HIGH.

TTL Subfamilies

The original standard TTL is infrequently used today. In the standard TTL circuit, the transistors saturate, reducing the operating speed. TTL variations cure this by clamping the transistors with Schottky diodes to prevent saturation, or by using a dopant in the chip fabrication to reduce transistor recovery time. Schottky-clamped TTL is the faster of these two manufacturing processes.

TTL IC identification numbers begin with either 54 or 74. The 54 prefix denotes a military temperature range of –55 to 125°C, while 74 indicates a commercial temperature range of 0 to 70°C. The next letters, in the middle of the TTL device number, indicate the TTL subfamily. Following the subfamily designation is a 2, 3 or 4-digit device-identification number. For example, a 7400 is a standard TTL NAND gate and a 74LS00 is a low-power Schottky NAND gate. (The NAND gate is the workhorse TTL chip. Recall, from Fig 5.68, the alternative implementation of the S-R flip-flop.) The following TTL subfamilies are available:

	74xx	standard TTL
H	74Hxx	High-speed
L	74Lxx	Low-power
S	74Sxx	Schottky
LS	74LSxx	Low-power Schottky
AS	74ASxx	Advanced Schottky
ALS	74ALSxx	Advanced Low-power Schottky

Each subfamily is a compromise between speed and power consumption. Because the speed-power product is approximately constant, less power consumption results in less speed and vice versa. For the amateur, an additional consideration to the speed-versus-power trade-off is the cost trade-off. The advanced Schottky devices offer both increased speed and reduced power consumption but at a higher cost.

In addition to the above power/speed/cost trade-offs, each TTL subfamily has particular characteristics that can make it suitable or unsuitable for a specific design. **Table 5.9** shows some of these parameters. The actual parameter values may vary slightly from manufacturer to manu-

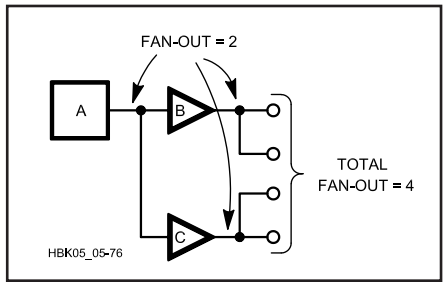


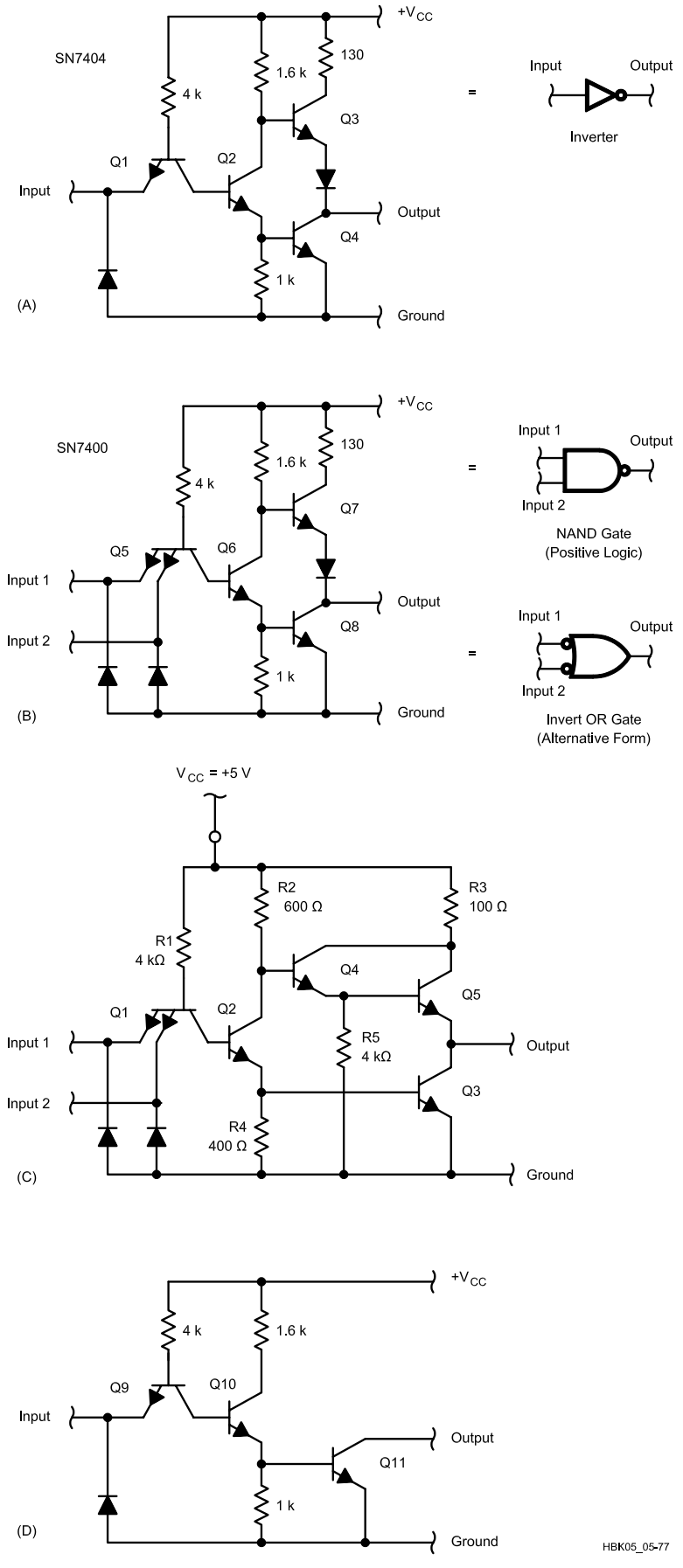
Fig 5.76 — Nonverting buffers used to increase fan-out: Gate A (fan-out = 2) is connected to two buffers, B and C, each with a fan-out of 2. Result is a total fan-out of 4.

Table 5.9

TTL and CMOS Subfamily Performance Characteristics

TTL Family	*Propagation Delay (ns)*	*Per Gate Power Consumption (mW)*	*Speed Power Product (pico-joules)*
Standard	9	10	90
L	33	1	33
H	6	22	132
S	3	20	60
LS	9	2	18
AS	1.6	20	32
ALS	5	1.3	6.5

CMOS Family Operating with $4.5 < V_{CC} < 5.5$ V		*f=100 kHz*	*f=1 MHz*	*f=10 MHz*	*f=100 kHz*	*f=1 MHz*	*f=10 MHz*
HC	18	0.0625	0.6025	6.0025	1.1	10.8	108
HCT	18	0.0625	0.6025	6.0025	1.1	10.8	108
AC	5.25	0.080	0.755	7.505	0.4	3.9	39
ACT	4.75	0.080	0.755	7.505	0.4	3.6	36



facturer, so always consult the manufacturers' data books for complete information.

TTL Circuits

Fig 5.77A shows the schematic representation of a TTL hex inverter. A 7404 chip contains four of these inverters. When the input is low, Q1 is ON, conducting current from base to emitter through the input lead and into ground. Thus, a low TTL input device must be prepared to sink current from the input. Since Q1 is saturated, Q2 is OFF because there is not enough voltage at its base. Similarly, Q4 is also OFF. With Q2 and Q4 OFF, Q3 will be ON and pull the output high, about one volt below V_{CC}. When the input is high, an unusual situation occurs: Q1 is operating in the inverse mode, with current flowing from base to collector. This current causes Q2 to be ON, which causes Q4 to be ON. With Q2 and Q4 ON, there is not enough current left for Q3, so Q3 is OFF. Q4 is pulling the output low.

By replacing Q1 with a multiple-emitter transistor, as is done with the two-input Q5 in Fig 5.77B, the inverter circuit becomes a NAND gate. Commercially available TTL NAND gates have as many as 13 inputs, the limiting factor being the number of input pins on the standard 16-pin chip. The operation of this multiple-input NAND circuit is the same as described for the inverter, the difference being that any one of the emitter inputs being low will conduct current through the emitter, leading to the conditions described above to produce a high at the output. Similarly, all inputs must be high to produce the low output.

In the TTL circuit of Fig 5.77A, transistors Q3 and Q4 are arranged in a *totem-pole* configuration. This configuration gives the output circuit a low source impedance, allowing the gate to source (supply) or sink substantial output current. The 130-Ω resistor between the collector of Q3 and $+V_{CC}$ limits the current through Q3.

When a TTL gate changes state, the amount of current that it draws changes rapidly. These changes in current, called switching transients, appear on the power supply line and can cause false triggering

Fig 5.77 — Example TTL circuits and their equivalent logic symbols: (A) an inverter and (B) a NAND gate, both with totem-pole outputs. (C) A NAND gate with a Darlington output. (D) A NAND gate with an open-collector output. (Indicated resistor values are typical. Identification of transistors is for text reference only. These are not discrete components but parts of the silicon die.)

of other devices. For this reason, the power bus should be adequately decoupled. For proper decoupling, connect a 0.01 to 0.1 μF capacitor from V_{CC} to ground near each device to minimize the transient currents caused by device switching and magnetic coupling. These capacitors must be low-inductance, high-frequency RF capacitors (disk-ceramic capacitors are preferred). In addition, a large-value (50 to 100 μF) capacitor should be connected from V_{CC} to ground somewhere on the board to accommodate the continually changing I_{CC} requirements of the total V_{CC} bus line. These are generally low-inductance tantalum capacitors rather than rolled-foil Mylar or aluminum-electrolytic capacitors.

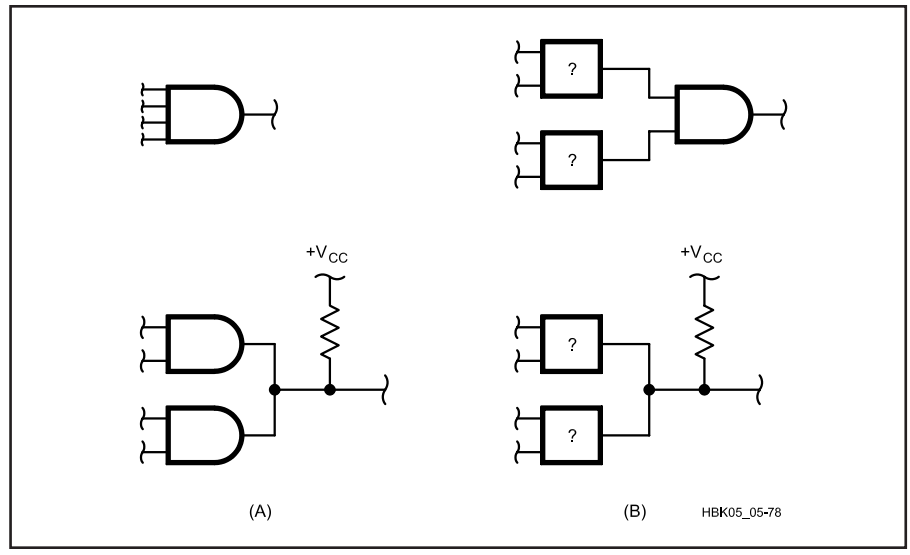


Fig 5.78 — The outputs of two open-collector-output AND gates are shorted together (wire ANDed) to produce an output the same as would be obtained from a 4-input AND gate.

Darlington and Open-Collector Outputs

Fig 5.77C and D show variations from the totem-pole configuration. They are the Darlington transistor pair and the open-collector configuration respectively.

The Darlington pair configuration replaces the single transistor Q4 with two transistors, Q4 and Q5. The effect is to provide more current-sourcing capability in the high state. This has two benefits: (1) the rise time is decreased and (2) the fan-out is increased.

Transistor(s) on the output in both the totem-pole and Darlington configurations provide active pull-up. Omitting the transistor(s) and providing an external resistor for passive pull-up gives the open-collector configuration. This configuration, unfortunately, results in slower rise time, since a relatively large external resistor must be used. The technique has some very useful applications, however: driving other devices, performing wired logic, busing and interfacing between logic devices.

Devices that need other than a 5-V supply can be driven with the open-collector output by substituting the device for the external resistor. Example devices include light-emitting diodes (LEDs), relays and solenoids. Inductive devices like relay coils and solenoids need a "flyback" protection diode across the coil. You must pay attention to the current ratings of open-collector outputs in such applications. You may need a switching transistor to drive some relays or other high-current loads.

Open-collector outputs can perform wired logic, rather than gated IC logic, by wire-ANDing the outputs. This can save the designer an AND gate, potentially simplifying the design. Wire-ANDed outputs are several open-collector outputs connected to a single external pull-up resistor. The overall output, then, will only be high when all pull-down transistors are OFF (all connected outputs are high), effectively performing an AND of the connected outputs. If any of the connected outputs are low, the output after the external resistor will be low. **Fig 5.78** illustrates the wire-ANDing of open-collector outputs.

The wire-ANDed concept can be applied to several devices sharing a common bus. At any time, all but one device has a high-impedance (off) output. The remaining device, enabled with control circuitry, drives the bus output.

Open-collector outputs are also useful for interfacing TTL gates to gates from other logic families. TTL outputs have a minimum high level of 2.4 V and a maximum low level of 0.4 V. When driving non-TTL circuits, a pull-up resistor (typically 2.2 kΩ) connected to the positive supply can raise the high level to 5 V. If a higher output voltage is needed, a pull-up resistor on an open-collector output can be connected to a positive supply greater than 5 V, so long as the chip output voltage and current maximums are not exceeded.

Three-State Outputs

While open-collector outputs can perform bus sharing, a more popular method is three-state output, or tristate, devices. The three states are low, high and high impedance, also called Hi-Z or *floating*. An output in the high-impedance state behaves as if it is disconnected from the circuit, except for possibly a small leakage current. Three-state devices have an additional disable input. When enable is low, the device provides high and low outputs just as it would normally; when enable is high, the device goes into its high-impedance state.

A bus is a common set of wires, usually used for data transfer. A three-state bus has several three-state outputs wired together. With control circuitry, all devices on the bus but one have outputs in the high-impedance state. The remaining device is enabled, driving the bus with high and low outputs. Care should be taken to ensure only one of the output devices can be enabled at any time, since simultaneously connected high and low outputs may result in an incorrect logic voltage. (The condition when more than one driver is enabled at the same time is called *bus contention*.) Also, the large current drain from V_{CC} to ground through the high driver to the low driver can potentially damage the circuit or produce noise pulses that can affect overall system behavior.

Unused TTL Inputs

A design may result in the need for an n-input gate when only an n + m input gate is available. In this case, the recommended solution for extraneous inputs is to give the extra inputs a constant value that won't affect the output. A low input is easily provided by connecting the input to ground. A high input can be provided with either an inverter whose input is ground or with a pull-up resistor. The pull-up resistor is preferred rather than a direct connection to power because the resistor limits the current, thus protecting the circuit from transient voltages. Usually, a 1-kΩ to 5-kΩ resistor is used; a single

1-kΩ resistor can handle up to 10 inputs.

It's important to properly handle all inputs. Design analysis would show that an unconnected, or floating TTL input is usually high but can easily be changed low by only a small amount of capacitively-coupled noise.

METAL-OXIDE SEMICONDUCTOR (MOS) LOGIC FAMILIES

While bipolar devices use junction transistors, MOS devices use field effect transistors (FETs). MOS is characterized by simple device structure, small size (high density) and ease of fabrication. MOS circuits use the NOR gate as the workhorse chip rather than the NAND. MOS families are used extensively in digital watches, calculators and VLSI circuits such as microprocessors and memories.

P-Channel MOS (PMOS)

The first MOS devices to be fabricated were PMOS, conducting electrical current by the flow of positive charges (holes). PMOS power consumption is much lower than that of bipolar logic, but its operating speed is also lower. The only extensive use of PMOS is in calculators and watches, where low speed is acceptable and low power consumption and low cost are desirable.

N-Channel MOS (NMOS)

With improved fabrication technology, NMOS became feasible and provided improved performance and TTL compatibility. The speed of NMOS is at least twice that of PMOS, since electrons rather than holes carry the current. NMOS also has greater gain than PMOS and supports greater packaging density through the use of smaller transistors.

Complementary MOS (CMOS)

CMOS combines both P-channel and N-channel devices on the same substrate to achieve high noise immunity and low power consumption: less than 1 mW per gate and negligible power during standby. This accounts for the widespread use of CMOS in battery-operated equipment. The high impedance of CMOS gates makes them susceptible to electromagnetic interference, however, particularly if long traces are involved. Consider a trace ¼-wavelength long between input and output. The output is a low-impedance point, hence the trace is effectively grounded at this point. You can get high RF potentials ¼-wavelength away, which disturbs circuit operation.

A notable feature of CMOS devices is that the logic levels swing to within a few millivolts of the supply voltages. The input-switching threshold is approximately one half the supply voltage ($V_{DD} - V_{SS}$). This characteristic contributes to high noise immunity on the input signal or power supply lines. CMOS input-current drive requirements are minuscule, so the fan-out is great, at least in low-speed systems. For high-speed systems, the input capacitance increases the dynamic power dissipation and limits the fan-out.

CMOS Subfamilies

There are a number of CMOS subfamilies available. Like TTL, the original CMOS has largely been replaced by later subfamilies using improved technologies. The original family, called the 4000-series, has numbers beginning with 40 or 45 followed by two or three numbers to indicate the specific device. 4000B is second generation CMOS. When introduced, this family offered low power consumption but was fairly slow and not easy to interface with TTL.

Later CMOS subfamilies provided improved performance and TTL compatibility. For simplicity, the later subfamilies were given numbers similar to the TTL numbering system, with the same leading numbers, 54 or 74, followed by 1 to 3 letters indicating the subfamily and as many as 5 numbers indicating the specific device. The subfamily letters usually include a "C" to distinguish them as CMOS.

The following CMOS device families are available:

4000	4071B	standard CMOS
C	74Cxx	CMOS versions of TTL

Devices in the 74C subfamily are pin and functional equivalents of many of the most popular parts in the 7400 TTL family. It may be possible to replace all TTL ICs in a particular circuit with 74C-series CMOS, but this family should not be mixed with TTL in a circuit without careful design considerations. Devices in the C series are typically 50% faster than the 4000 series.

HC	74HCxx	High-speed CMOS

Devices in the 74HC subfamily have speed and drive capabilities similar to Low-power Schottky (LS) TTL but with better noise immunity and greatly reduced power consumption. High-speed refers to faster than the previous CMOS family, the 4000-series.

HCT	74HCTxx	High-Speed CMOS, TTL compatible

Devices in this subfamily were designed to interface TTL to CMOS systems. The HCT inputs recognize TTL levels, while the outputs are CMOS compatible.

AC	74ACxxxxx	Advanced CMOS

Devices in this family have reduced propagation delays, increased drive capabilities and can operate at higher speeds than standard CMOS. They are comparable to Advanced Low-power Schottky (ALS) TTL devices.

ACT	74ACTxxxxx	Advanced CMOS, TTL compatible

This subfamily combines the improved performance of the AC series with TTL-compatible inputs.

As with TTL, each CMOS subfamily has characteristics that make it suitable or unsuitable for a particular design. You should consult the manufacturer's data books for complete information on each subfamily being considered.

CMOS Circuits

A simplified diagram of a CMOS logic inverter is shown in **Fig 5.79**. When the

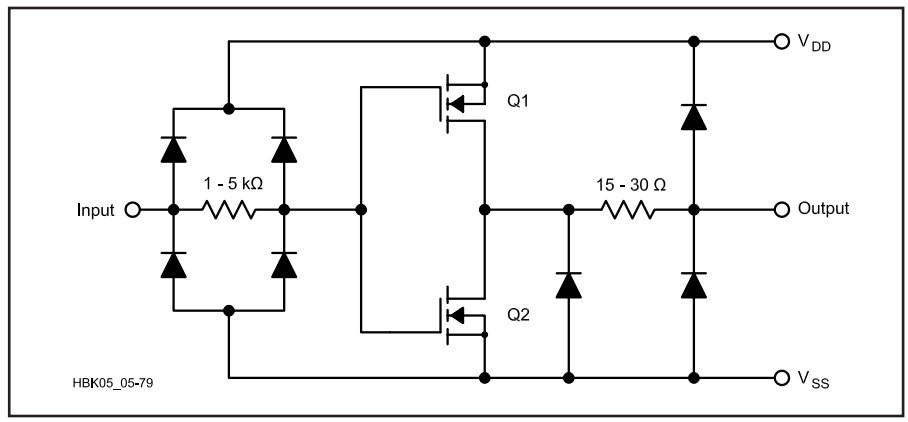


Fig 5.79 — Internal structure of a CMOS inverter.

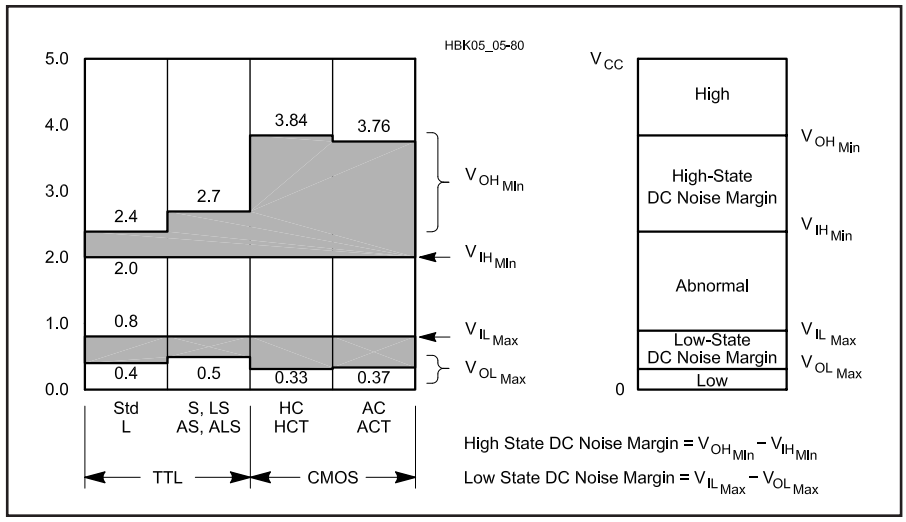


Fig 5.80 — Differences in logic levels for some TTL and CMOS families.

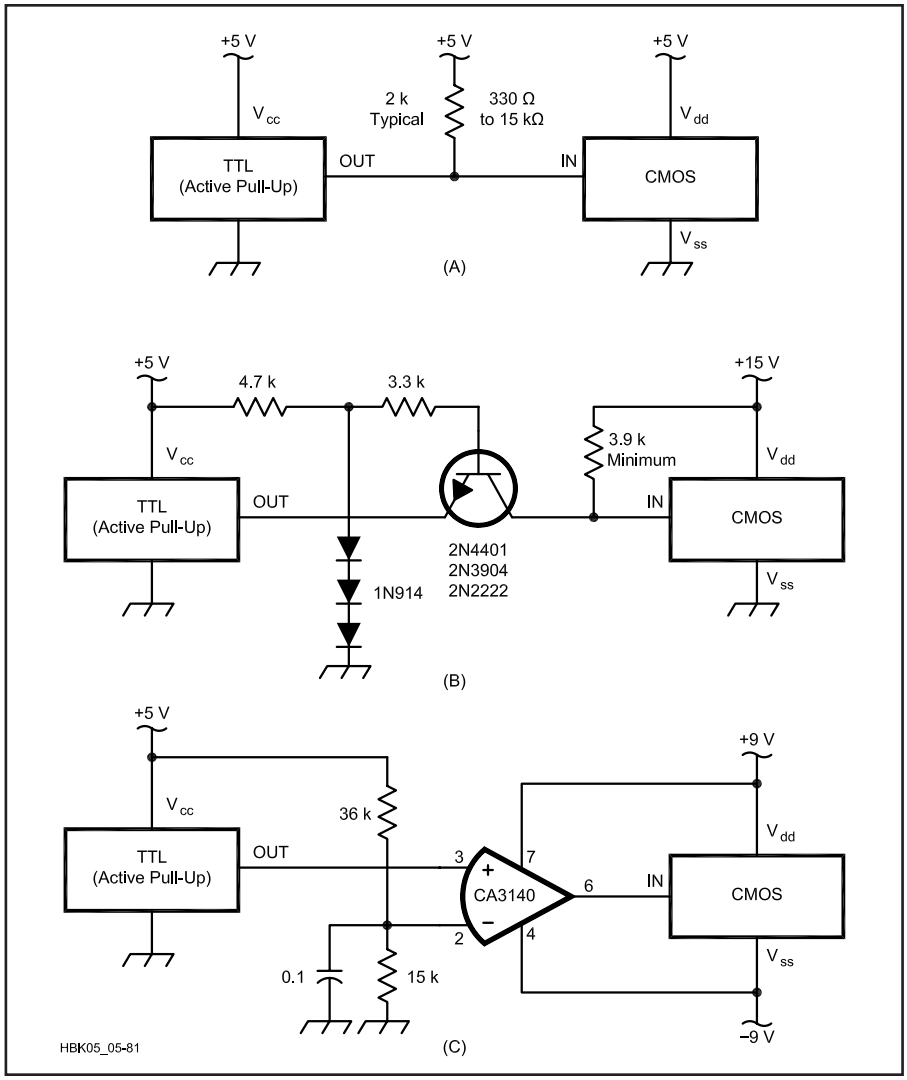


Fig 5.81 — TTL to CMOS interface circuits: (A) pull-up resistor, (B) common-base level shifter and (C) op amp configured as a comparator.

input is low, the resistance of Q2 is low so a high current flows from V_{CC}. Since Q1's resistance is high, the high current flows to the output. When the input is high, the opposite occurs: Q2's resistance is low, Q1's is high and the output is low. The diodes are to protect the circuit against static charges.

Special Considerations

Some of the diodes in the input- and output-protection circuits are an inherent part of the manufacturing process. Even with the protection circuits, however, CMOS ICs are susceptible to damage from static charges. To protect against damage from static, the pins should not be inserted in Styrofoam as is sometimes done with other components. Instead, a spongy conductive material is available for this purpose. Before removing a CMOS IC from its protective material, make certain that your body is grounded. Touching nearly any large metal object before handling the ICs is probably adequate to drain any static charge off your body. Some people prefer to touch a grounded metal object or to use a conductive bracelet connected to the ground terminal of a three-wire ac outlet through a 10-MΩ resistor. Since wall outlets aren't always wired properly, you should measure the voltage between the ground terminal and any metal objects you might touch. Connecting yourself to ground through a 1-MΩ to 10-MΩ resistor will limit any current that might flow through your body.

All CMOS inputs should be tied to an input signal. A positive supply voltage or ground is suitable if a constant input is desired. Undetermined CMOS inputs, even on unused gates, may cause gate outputs to oscillate. Oscillating gates draw high current, overheat and self destruct.

The low power consumption of CMOS ICs made them attractive for satellite applications, but standard CMOS devices proved to be sensitive to low levels of radiation — cosmic rays, gamma rays and X rays. Later, radiation-hardened CMOS ICs, able to tolerate 10^6 rads, made them suitable for space applications. (A rad is a unit of measurement for absorbed doses of ionizing radiation, equivalent to 10^{-2} joules per kilogram.)

SUMMARY

There are many types of logic ICs, each with its own advantages and disadvantages. Regardless of the application, consult up-to-date literature when designing logic circuits. IC databooks and application notes are usually available from IC manufacturers and distributors. Also, just about all of this information is available

on the Internet. By using a search engine and entering a few key word specifications, you will locate application notes, tutorials and a host of other information.

INTERFACING LOGIC FAMILIES

Each semiconductor logic family has its own advantages in particular applications. When a design mixes ICs from different logic families, the designer must account for the differing voltage and current requirements each logic family recognizes. The designer must ensure the appropriate interface exists between the point at which one logic family ends and another begins. A knowledge of the specific input/output (I/O) characteristics of each device is necessary, and a knowledge of the general internal structure is desirable to ensure reliable digital interfaces. Typical internal structures have been illustrated for each common logic family. **Fig 5.80** illustrates the logic level changes for different TTL and CMOS families. Databooks should be consulted for manufacturer's specifications.

Often more than one conversion scheme is possible, depending on whether the designer wishes to optimize power consumption or speed. Usually one quality must be traded off for the other. The following section discusses some specific logic conversions. Where an electrical connection between two logic systems isn't possible, an optoisolator can sometimes be used.

TTL Driving CMOS

TTL and low-power TTL can drive 74C series CMOS directly over the commercial temperature range without an external pull-up resistor. However, they cannot drive 4000-series CMOS directly, and for HC-series devices, a pull-up resistor is recommended. The pull-up resistor, connected between the output of the TTL gate and V_{CC} as shown in **Fig 5.81A**, ensures proper operation and enough noise margin by making the high output equal to V_{DD}. Since the low output voltage will also be affected, the resistor value must be chosen with both desired high and low voltage ranges in mind. Resistor values in the range 1.5 kΩ to 4.7 kΩ should be suitable for all TTL families under worst conditions. A larger resistance reduces the maximum possible speed of the CMOS gate; a lower resistance generates a more favorable RC product but at the expense of increased power dissipation.

HCT-series and ACT-series CMOS devices were specifically designed to interface non-CMOS devices to a CMOS system. An HCT device acts as a simple buffer between the non-CMOS (usually TTL) and CMOS device and may be com-

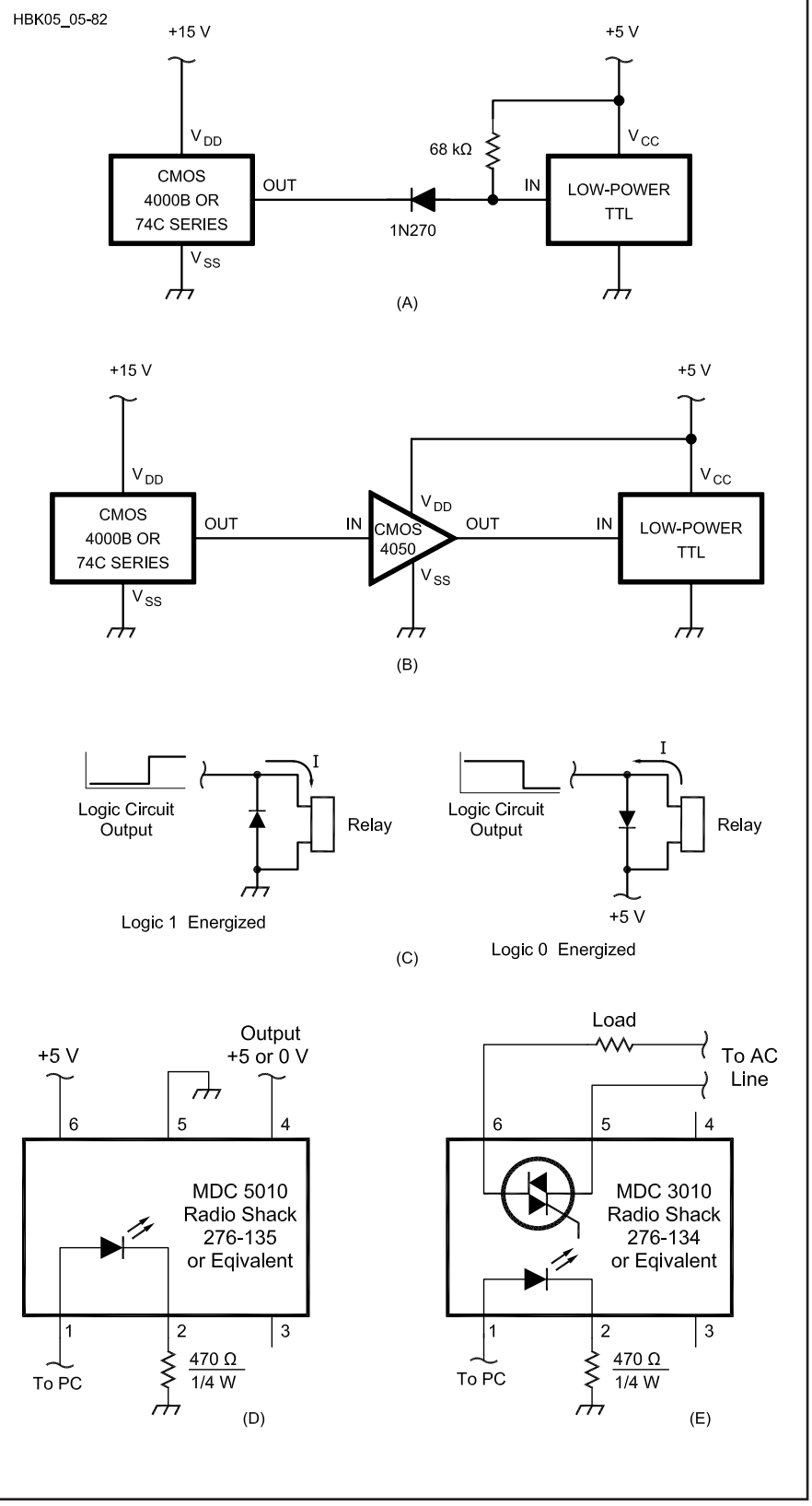


Fig 5.82 — CMOS to TTL interface circuits: (A) blocking diode chosen when different supply voltages are used. The diode is not necessary if both devices operate with a +5 V supply. (B) CMOS noninverting buffer IC. (C) Relay interface for total isolation. (D) An electro-optical coupler can be used in place of a relay. (E) An electro-optical coupler can also drive a triac-based unit for switching ac.

bined with a logic function if a suitable HCT device is available.

When the CMOS device is operating from a power supply other than +5 V, the TTL interface is more complex. One fairly simple technique uses a TTL open-collector output connected to the CMOS input, with a pull-up resistor from the CMOS input to the CMOS power supply. Another method, shown in Fig 5.81B, is a common-base level shifter. The level shifter translates a TTL output signal to a +15 V CMOS signal while preserving the full noise immunity of both gates. An excellent converter from TTL to CMOS using dual power supplies is to configure an operational amplifier as a comparator, as shown in Fig 5.81C. An FET op amp is shown because its output voltage can usually swing closer to the rails (+ and – supply voltages) than a bipolar device.

CMOS Driving TTL

Certain CMOS devices can drive TTL loads directly. The output voltages of CMOS are compatible with the input requirements of TTL, but the input-current requirement of TTL limits the number of TTL loads that a CMOS device can drive from a single output (the fan-out).

Interfacing CMOS to TTL is a bit more complicated when the CMOS is operating at a voltage other than +5 V. One technique is shown in **Fig 5.82A**. The diode blocks the high voltage from the CMOS gate when it is in the high output state. A germanium diode is used because its lower forward-voltage drop provides higher noise immunity for the TTL device in the low state. The 68-kΩ resistor pulls the input high when the diode is back biased.

There are two CMOS devices specifically designed to interface CMOS to TTL when TTL is using a lower supply voltage. The CD4050 is a noninverting buffer that allows its input high voltage to exceed the supply voltage. This capability allows the CD4050 to be connected directly between the CMOS and TTL devices, as shown in Fig 5.82B. The CD4049 is an inverting buffer that has the same capabilities as the CD4050.

Real-World Interfacing

Quite often logic circuits must either drive or be driven from non-logic sources. A very common requirement is sensing the presence or absence of a high (as compared to +5 volts) voltage or perhaps turning on or off a 120-VAC motor, such as an antenna rotor. A similar problem occurs when two different units in the shack must be interfaced since induced AC voltages or ground loops can cause problems with the desired signals.

A slow speed but safe way to interface such circuits is to use a relay. This provides absolute isolation between the logic circuits and the load. Fig 5.82C shows the correct way to provide this connection. The relay coil is selected to draw less than the available current from the driving logic circuit. The diode, most often a 1N914 or equivalent switching diode, prevents the inductive load from back-biasing the logic circuit and possibly destroying it.

Electro-optical couplers can also be used for this circuit interfacing. Fig 5.82D uses one to interface two sets of logic circuits, and Fig 5.82E interfaces with the AC line.

MSI, LSI VLSI Circuits and Controllers

In addition to using the basic logic elements discussed in the previous sections, there are many integrated circuits available for Amateur Radio applications that include the equivalent of dozens, hundreds or perhaps even thousands of gates. Often, it is no harder to use one of these units than it is to use a few gates and flip-flops.

The analog to digital (A/D) and digital-to-analog ((D/A) converters discussed in the section following (on the parallel port) are examples. Hybrid (containing both analog and digital functions) integrated circuits provide other opportunities for builders. As additional examples, the LM3914 is a LED driver that takes an analog signal in and turns on one or more LEDs, depending on the analog voltage. A self-contained signal generator, the MAX038, can accept either switch closures or digital control signals and generates selectable sin, square or triangle waves with variable frequency.

At the other end of the spectrum are microcontrollers, which can be considered *PCs on a chip.* These include a reduced version of an arithmetic-logic unit, as described in the next section as well as interfacing and some data and program storage. However, these units require a special purpose programmer to load and store the programs, and programming them can become quite involved.

A larger, but perhaps friendlier microcontroller is the various versions (and clones) of the *BASIC Stamp* (*BASIC Stamp* is a registered trademark of Parallax, Inc.) Several varieties are available, with varying capabilities. They are programmed in a version of the *BASIC* language using a PC, and, then, the program is downloaded from the PC into the BASIC Stamp.

Computer Hardware

So far, this chapter segment has discussed digital logic, the implementation of that logic with integrated circuits, interfacing IC logic families and the use of memory to store information used by the ICs. The synthesis of all this technology is the microcomputer — combining a microprocessor IC, memory, peripheral devices, and user interface into the modern personal computer. A computer has both physical components (hardware) and a collection of programs (software) to tell it what to do. This section (by Bob Wolbert, K6XX with additional material by Paul Danzer, N1II) will focus on the physical components of the computer: its internal physical components, their interaction, and peripheral I/O devices that communicate with other systems and the operator.

Material relating to computers in general, and PCs in particular, tends to become obsolete very quickly. For that reason, a quick look at the ever-growing PC-related section at your local bookstore will show very few books on the shelf older than perhaps 12 to 18 months. While the basic technology of electronics as applied to PCs does not change, the standard, performance and configurations change monthly, weekly and perhaps even daily. Thus, this section will provide only an overview of PC hardware and technology.

A WORD ABOUT AVAILABILITY

While many of an application's programs such as RTTY, SSTV, PSK AMTOR (and several dozen others) work best with fast, new PC hardware, many Amateur applications using the hardware as a controller do not require very much capability. Since many used computers are available at prices ranging from nothing (just haul it away) to perhaps $100, these units are worth considering for single, special applications. Want a modulation

monitor or a recording voltmeter? How about basing the design on an old PC? Application of the material in this chapter to interfacing with an old PC can provide a host of ham-shack capabilities.

WHAT IS A COMPUTER?

The strictest definition of the term "computer" includes special purpose digital systems optimized for a particular task. For example, a modern synthesized transceiver, with its memory, I/O, serial control, DSP, etc. meets the definition of a computing device. Many of the concepts discussed in this section apply equally well to your HF rig as well as to your PC; however, our definition of a computer will be restricted to a general purpose machine whose task is quickly and easily changed by loading or changing software. If its task cannot be readily modified — to compute a spreadsheet or compose e-mail, for example — we will exclude that system from our discussion. The personal computer (PC) will be emphasized due to its ubiquitous nature.

The three major divisions of a PC are its hardware, its software, and its firmware. See **Fig 5.83**. The hardware includes the *central processing unit* (CPU) and input/output (I/O) devices. Software refers to the programs that are loaded into the computer to configure it for the task at hand. Firmware, also called microcode or BIOS (Basic Input/Output System), is a hybrid of both hardware and software that is used to perform specific tasks. The microcode is the basis of the microprocessor's command set that tells it how to fetch data and add numbers. For example, the BIOS is firmware generally used to start-up (boot) the system.

COMPUTER ARCHITECTURE

Unlike many present textbooks, where computer architecture is narrowly defined as only including those attributes of the system that interest programmers, our discussion deals with the structural organization and hardware design of the digital computer system. All modern computer systems consist of three basic sections: the CPU, memory, and peripherals for interfacing with the operator and the real world. The architecture of a computer is the arrangement of these two specific internal subsystems, the CPU and the bus. The CPU (central processing unit), called a *microprocessor* in personal computers, is an IC consisting of three major parts: a control unit, an arithmetic logic unit (ALU), and temporary storage registers. The *bus* is a set of wires carrying address, data and control information, which interconnects all of the subsystems. Virtually all computers are designed based on the basic "Von Neumann" architecture shown in **Fig 5.84**.

The microprocessor, memory chips and other circuitry are all part of the system's hardware, the physical components of a system. The computer case, the nuts and bolts and physical parts are other parts of the hardware. A computer also includes software, a collection of programs or sequence of instructions to perform a specified task. The design of computers is so complex, however, that it is nearly impossible to design an original architecture without any bugs. Thus many designers use microprocessors that include *microcode* or microinstructions, which are instructions in the control unit of a microprocessor. This hybrid between hardware and software is called firmware. Firmware also includes software stored in ROM or EPROM rather than being stored on magnetic disk or tape.

Computer designers make decisions on hardware, software and firmware based on cost versus performance. Today's computer market includes a wide range of systems, from high-performance super-computers costing millions of dollars, to the personal microcomputer, with prices in the high hundreds to a few thousands new and ranging from free on up for older used models.

THE CENTRAL PROCESSING UNIT

The central processing unit is usually a single microprocessor chip, although its subsystems can be on more than one chip. The CPU at least includes a control unit, timing circuitry, an arithmetic logic unit (ALU) and registers for temporary storage. Modern microprocessors have tens of millions of transistors and are designed in modules.

Control Unit

The control unit directs the operation of the computer, managing the interaction between subunits. It takes instructions from the memory and executes them, performing tasks such as accessing data in memory, calling on the ALU or performing I/O. Control is one of the most difficult parts to design; thus it is the most likely source of bugs in designing an original architecture.

Microprocessors consist of both hardwired control and micro-programmed control. In both cases, the designer determines a sequence of states through which the computer cycles, each with inputs to examine and outputs to activate other CPU sub-

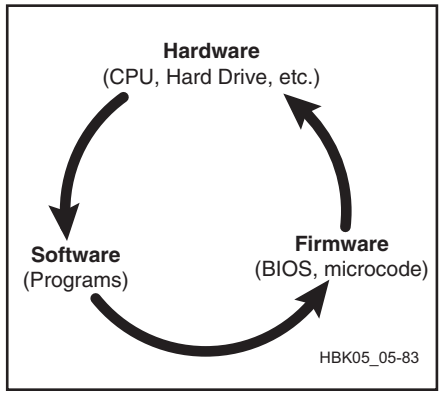


Fig 5.83 — Hardware, software and firmware comprise a computer.

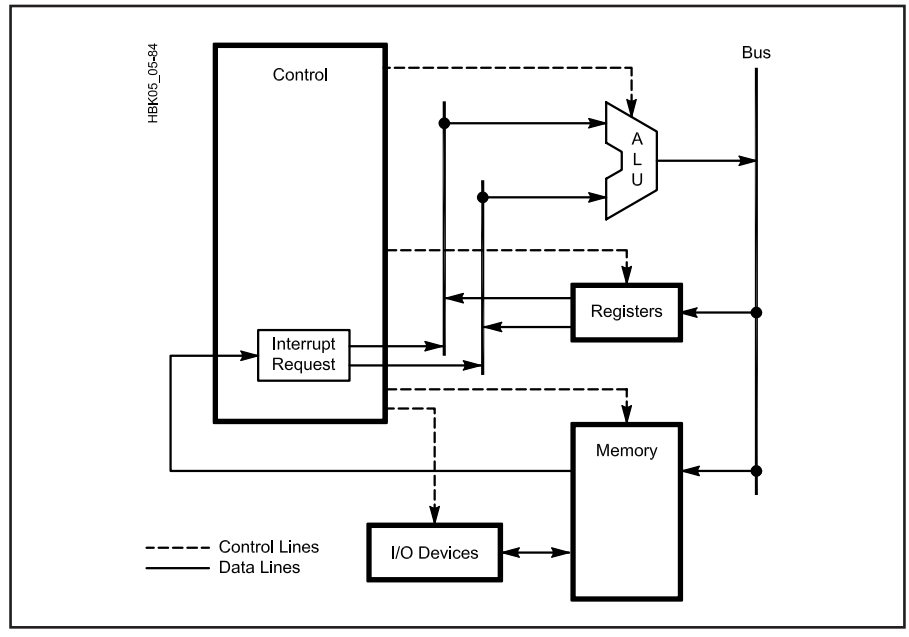


Fig 5.84 — Example of a basic computer architecture.

systems (including activating itself, indicating which state to do next). For example, the sequence usually starts with "Fetch the next instruction from memory," with control outputs to activate memory for a read, a program counter to send the address to be fetched and an instruction register to receive the memory contents. Hardwired control is completely via circuitry, usually with a programmed logic array. Microprogrammed control uses a microprocessor with a modifiable control memory, containing microcode or micro-instructions. An advantage of microprogrammed control is flexibility; the code can be changed without changing the hardware, making it easier to correct design errors.

Timing

Usually, an oscillator controlled by a quartz crystal generates the microcomputer's clock signal. The output of this clock goes to the microprocessor and to other ICs. The clock synchronizes the microcomputer subunits. For example, each of the microinstructions is designed to take only one clock cycle to execute, so any components triggered by a microinstruction's control outputs should finish their actions by the end of the clock cycle. The exception to this is memory, which may take multiple clock cycles to finish, so the control unit repeats in its same state until memory says it's done. Since the clock rate effectively controls the rate at which instructions are executed, the clock frequency is one way to measure the speed of a computer. Clock frequency, however, cannot be the only criteria considered because the actions performed during a clock cycle vary for different designs, particularly processors with superscalar or pipelined designs capable of computing multiple items simultaneously.

Arithmetic Logic Unit

The *arithmetic logic unit* (ALU) performs logical operations such as AND, OR and SHIFT and two number arithmetic operations such as addition, subtraction, multiplication and division. The ALU depends on the control unit to tell it which operation to perform and also to trigger other devices (memory, registers and I/O) to supply its input data and to send out its results to the appropriate place.

The ALU often only performs simple operations. Complex operations, such as multiplication, division and operations involving decimal numbers, are performed by dedicated hardware, called floating-point processors, or *coprocessors*.

Registers

Microprocessor chips have some internal memory locations that are used by the control unit and ALU. Because they are inside the microprocessor IC, these registers can be accessed more quickly than main memory locations. Special purpose registers or *dedicated registers* are purely internal, have predefined uses and cannot be directly accessed by programs. *General-purpose registers* hold data and addresses in use by programs and can be directly accessed, although usually only by assembly level programs.

The dedicated registers include the instruction register, program counter, effective address register and status register. The first step to execute an instruction is to fetch it from memory and put it in the *instruction register* (IR). The *program counter* (PC) is then incremented to contain the address of the next instruction to be fetched. An instruction may change the program counter as a result of a conditional branch (if-then), loop, subroutine call or other nonlinear execution. If data from memory is needed by an instruction, the address of the data is calculated and fetched with the *effective address register* (EAR). The *status register* (SR) keeps track of various conditions in the computer. For example, it tells the control unit when the keyboard has been typed on so the control unit knows to get input. It also notices if something goes wrong during an instruction execution, for example an attempted divide by 0, and tells the control unit to halt the program or fix the error. Certain bits in the status register are known as the *condition codes*, flags set by each instruction. These flags tell information about the result of the latest instruction, such as if the result was negative, positive or zero and if an arithmetic overflow or a carry error occurred. The flags can then be used by a conditional branch to decide if that branch should be chosen.

MEMORY

Computers and other digital circuits rely on stored information, either data to be acted upon or instructions to direct circuit actions. This information is stored in memory devices, in binary form. Computers use four main types of memory, as shown in **Fig 5.85**.

Accessing a Memory Item

Memory devices consist of a large number of memory cells each capable of remembering one bit of binary information. The information in memory is stored in digital form with collections of bits, called words, representing numbers and symbols. The most common symbol set is the American National Standard Code for Information Interchange (ASCII). Words in memory, just like the letters in this sentence, are stored one after the other. They are accessed by their location or address. The number of bits in each word, equal to the number of memory cells per memory location, is constant within a memory device but can vary for different devices. Common memory devices have word sizes of 8, 16 and 32 bits.

Addresses and Chip Size

An *address* is the identifier, or name, given to a particular location in memory. Since this address is expressed as a binary number, the number of unique addresses available in a particular memory chip is determined by the number of bits to express the address. For example, a memory chip with 8 bit addresses has $2^8 = 256$ memory locations. These locations are accessed as the addresses 00000000 through 11111111, 0 through 255 decimal or 00 through FF hex. (For ease of notation, programmers and circuit designers use hexadecimal (base 16) notation to avoid long strings of 1s and 0s.) The memory chip size can be expressed as M × N, where M is the number of unique addresses, or memory locations and N is the word size, or number of bits per memory location. Memory chips come in a variety of sizes and can be arranged, together with control circuitry and decoders, to meet a designer's needs.

Memory chips, no matter how large or small, have several things in common. Each chip has address, data and control lines. A memory chip must have enough

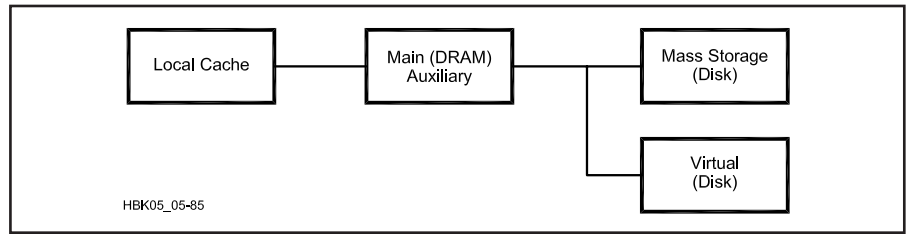


Fig 5.85 — Computer memory types.

address lines to uniquely address each of its words and as many data lines as there are bits per word. Memory size is usually specified in kilobytes (kbyte). This is usually abbreviated as K. Notice this is not quite the same as the metric prefix kilo, because it represents 1024, rather than 1000. When checking memory size on a PC, using various programs and tools, it is very common to see slightly different numbers as a result of these checks and tools. These differences are usually due to some processes giving results in increments of 1024 and some is actual increments of 1000.

Memory Types

The concepts described above are applied to several types of random-access, semiconductor memory. Semiconductor memories are categorized by the ease and speed with which they can be accessed and their ability to "remember" in the absence of power.

SAM versus RAM

One way to categorize memory is by which memory cells can be accessed at a given instant. *Sequential-access memory* (SAM) must be accessed by stepping past each memory location until the desired location is reached. Magnetic tapes implement SAM. To reach information in the middle of the tape, the tape head must pass over all of the information on the beginning of the tape. Two special types of SAM are the queue and the push-down stack. In a *queue*, also called a first-in, first-out (FIFO) memory, locations must be read in the order that they were written. The queue is a first-come, first-served device, like a line at a ticket window. The push-down stack is also called last-in, first-out (LIFO) memory. In LIFO memory, the location written most recently is the next location read. LIFO can be visualized as a stack, always adding to and removing from the top of the stack. *Random-access memory* (RAM) allows any memory cell to be accessed at any instant, with no time wasted stepping past the beginning parts of the data. Random-access memory is like a bookcase; any book can be pulled out at any time.

It is usually faster to access a desired word in RAM than in SAM. Also, all words in RAM have the same access time, while each word in a SAM has a different access time based on its position. Generally, the semiconductor memory devices internal to computers are random-access memories. Magnetic devices, such as tapes and disks, have at least some sequential access characteristics. We will leave tapes and disks for a later section and concentrate here on random-access, solid-state memories.

Random Access Memory

Most RAM chips are *volatile*, meaning that stored information is lost if power is removed. RAM is either static or dynamic. *Dynamic RAM* (DRAM) stores a bit of information as the presence or absence of charge. This charge, since it is stored in a capacitor, slowly leaks away and must be refreshed periodically. Memory refresh typically occurs every few milliseconds and is usually performed by a dynamic RAM controller chip. *Static RAM* (SRAM) stores a bit of information in a flip-flop. Since the bit will retain its value until either power is removed or another bit replaces it, refresh is not necessary.

Both types of RAM have their advantages and disadvantages. The advantage of DRAM is increased density and ease of manufacture, making them significantly less expensive. SRAMs, however, have much faster access times. Most general purpose computers use DRAMs, since large memory size and low cost are the major objectives. Where the amount of memory required doesn't justify the use of DRAM, and the faster access time is important, SRAMs are common, for example, in embedded systems (telephones, toasters), battery powered devices, and for cache memories. Cost, power consumption and access time, provided in manufacturers' data sheets, are factors to consider in selecting the best RAM for a given application.

New acronyms for various new memory types appear to be invented daily. The only solution to this problem is to do an internet-based search on an unfamiliar acronym; any result brought up by the search engine over 6-months old is almost certain to be obsolete!

Read-Only Memory

Read-only memory (ROM) is nonvolatile; its contents are not lost when power is removed from the memory. Despite its name, all ROMs can be written or programmed at least once. The earliest ROM designs were "written" by clipping a diode between the memory bit and power supply wherever a 0 was desired. Modern MOS ROMs use a transistor instead of a diode. Mask ROMs are programmed by having ones and zeros etched into their semiconductors at manufacturing time, according to a pattern of connections and non-connections provided in a mask. Since the programming of a mask ROM must be done by the manufacturer, adding expense and time delays, this type of ROM is primarily used only in high volume applications.

For low-volume applications, the programmable ROM (PROM) is the most effective choice since the data can be written after manufacture. A PROM is manufactured with all its diodes or transistors connected. A PROM programmer device then burns away undesired connections. This type of PROM can be written only once.

Two types of PROMs that can be erased and reprogrammed are EPROMs and EEPROMs. The transistors in UV erasable PROMs (EPROMs) have a floating gate surrounded by an insulating material. When programming with a bit value, a high voltage creates a negative charge on the floating gate. Exposure to ultraviolet light erases the negative charge. Similarly, electrically erasable PROMs (EEPROMs) erase their floating-gate values by applying a voltage of the opposite polarity.

Besides being nonvolatile, PROMs are also distinguished from RAMs by their read and write times. RAM read and write times are nearly equal, in the nanosecond range. Naturally, since PROMs are only written to infrequently, they can have slow write times (in the millisecond range). Their read times, however, are near those of RAM. Two factors make it hard to write to PROMs: (1) PROMs must be erased before they can be reprogrammed and (2) PROMs often require a programming voltage higher than their operating voltage.

ROMs are practical only for storing data or programs that do not change frequently and must survive when power is removed from the memory. The BIOS program that starts a computer when it is first switched on or the memory that holds the call sign in a repeater IDer are prime candidates for ROM.

Nonvolatile RAM

For some situations, the ideal memory would be as nonvolatile as ROM but as easy to write to as RAM. The primary example is data that must not be allowed to perish despite a power failure. Low-power RAMs can be used in such applications if they are supplied with NiCd or lithium cells for backup power. A more elegant and durable solution is nonvolatile RAM (NVRAM), which includes both RAM and ROM. The standard volatile RAM, called shadow RAM, is backed up by nonvolatile EEPROM. When the RECALL control is asserted, such as when power is first applied, the contents of the ROM are copied into the RAM. During normal operation, the system reads and writes to the RAM. When the STORE control is triggered, such as by a power failure or before turning off the system, the entire contents of the RAM are copied into the ROM for nonvolatile storage. In

the event of primary power failure, to successfully save the RAM data, some power must be maintained until the memory store is complete, generally about 20 ms.

Cache versus Main Memory

Memory is in high demand for many applications. To balance the trade-off of speed versus cost, most computers use a larger, slower, but cheaper main memory in conjunction with a smaller, faster, but more expensive cache memory. As you run a computer program, it accesses memory frequently. When it needs an item, a piece of data or the next part of the program to execute, it first looks in the cache. If the item is not found in the cache, it is copied to the cache from the main memory. As you run a computer program, it often repeats certain parts of the program and repeatedly uses pieces of data. Since this information has been copied to the high-speed cache, your computer game or other application can run faster. Information used less often or not being used at all (programs not currently being run) can stay in the slower main memory.

A "cache" is a place to store treasure; the treasure, the information you are using frequently, can be accessed quickly because it is in the high-speed cache. The use of cache versus main memory is managed by a computer's CPU so it is transparent to the user. The improvement in program execution time is similar to accessing a floppy disk versus the computer's internal memory.

I/O TRANSFERS

No computer will perform useful work without some means of communicating with the real world. Its input and output system allow the computer to react to and affect the outside world. The ability to interact with their environment is a primary reason why computers are so useful and cost-effective. Often, I/O is provided by a user, and a great deal of effort goes towards making computers user-friendly. Alongside the drive for user-friendly computers is the drive for automation. Data are acquired and operations are performed automatically, such as the packet bulletin board automatically forwarding a message.

BUS STRUCTURE: LOCAL BUSSES

Tying the blocks together is the data bus, the main information corridor inside the computer. The bus carries signals to and from various components, such as the CPU, the keyboard, mass storage, and communications ports. The first IBM PC and compatibles used the 62-pin, 8-bit, 8-MHz ISA bus, which was revised to the AT-bus, a 98-pin, 16-bit wide 8-MHz bus. Other pre-PCI bus architectures include Apple's NuBus, the Extended Industry Standard Architecture (EISA), and the VESA Local Bus.

The slow ISA bus was a bottleneck to system performance, so a separate bus was implemented between the microprocessor and main memory. This bus was called a *local bus*, as it was local to the CPU and memory only. Eventually, graphics and hard disk drive speeds increased to the point where they could ride on the local bus as well, without impacting memory access performance. Present computer systems are built around the Peripheral Components Interface (PCI-X) bus, a 64-bit wide system running at 133 MHz or higher.

The need to remain compatible with older PC architectures and conventions has hampered bus development and changes. As a result, instead of the invention of a host of new bus structures, most new PCs now include external bus attachments, such as USB. This permits attachment of more and faster devices than the older bus structures would allow.

PERIPHERALS

Peripherals work with the CPU and memory to provide additional capabilities. One of the most common examples is communication with a user via input devices and output devices. Peripherals may be divided into three groups: bidirectional (input/output) devices, input devices, and output devices. Bidirectional devices allow data storage and communications with the outside world. Input devices provide the computer both data to work on and programs to tell it what to do. Output devices present the results of computer operations to the user or another system and may even control an external system. Both input and output combine to provide user-friendly interaction. Most of these devices have adapted to certain standards and use readily available connectors and cables, enabling easy incorporation into a system. Knowledge of how external memory devices work is more useful and will be discussed in more detail.

Bidirectional Input/Output Devices

Mass storage and communications devices provide data input as well as output. Perhaps the most important peripheral in the computer system is the mass storage unit, such as the *disk drive*. Another important I/O unit is the *modem* (a contraction of MODulator/DEModulator), which allows easy communication between computers across standard telephone lines. A third is the *local area network* (LAN) card, which provides high speed communications between nearby computers.

Hard Disk Drives

An electromechanical hybrid, the hard disk drive provides the largest capacity at the lowest cost per byte of any random-access storage media. Hard drives are an essential part of present computer systems. Their key features include:

- Low cost per byte of storage.
- Large capacity available.
- Random access to data.
- Non-volatile magnetic storage.

Hard drives consist of three main units, the *head/disk assembly* (HDA), the *read/write channel*, and the *controller*. The HDA comprises the mechanical portion of the assembly, with one or more aluminum disks mounted on a spindle, which is rotated by a brushless dc motor, generally between 3600 and 7200 rpm. Read/write heads are mounted on an actuator arm that sweeps across the disk surfaces. The read/write amplifier is affixed to the actuator by a flexible cable to provide the lowest possible noise pickup. Read/write heads do not touch the disk surface; instead, air flowing over the rapidly spinning disk causes them to fly slightly above the disk. "Slightly" is no exaggeration — typical flying heights of drives made in the year 2000 are approximately two millionths of an inch (about 500 nanometers). Due to these extremely tight tolerances, the entire HDA is enclosed in a sealed aluminum casting that prevents contamination. As shown in **Fig 5.86**, debris such as a hair or a dust particle tower over the flying heads.

The drive controller performs data caching and communication with the host bus. Common hard drive busses include the EIDE (Extended Integrated Device Electronics), the (similar) ATA (AT-Attachment), and SCSI (small computer systems interface). External hard drives, interfacing with USB ports, are now both readily available and provide portability from PC to PC. In fact, some newer PCs now will boot from an external drive. Instead of carrying a PC from location to location, just unplug these thumb-sized drives and plug them back in at the new location.

Data is stored on both surfaces of each disk in concentric arcs called sectors, as shown in **Fig 5.87**. All sectors on one surface a given distance from the disk edge constitute a track. A cylinder is the collection of all similar tracks on all surfaces of the drive. When a new drive is installed, most newer PCs and their BIOS will recognize the new drive and set the BIOS

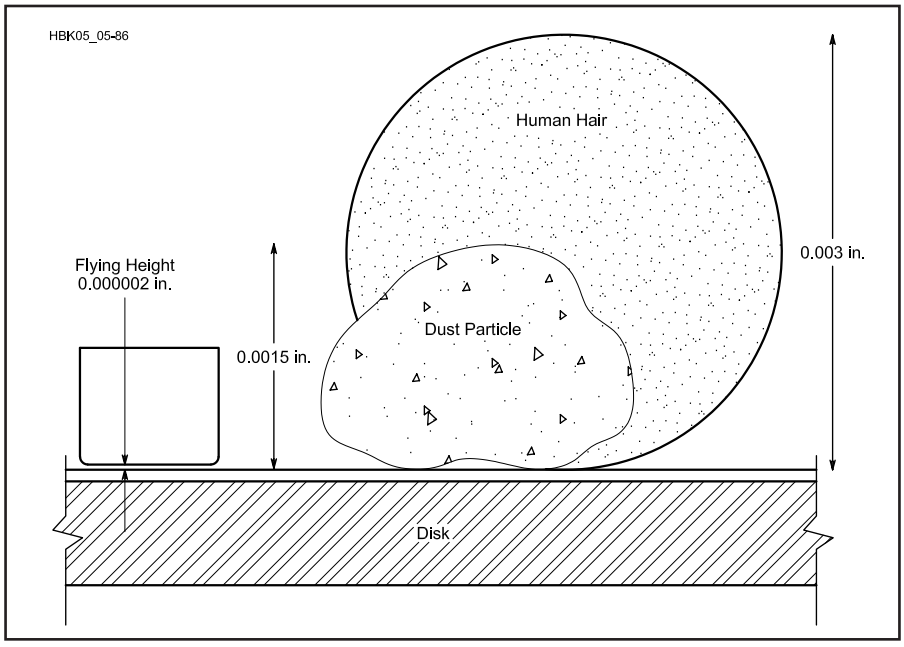


Fig 5.86 — Hard disk drive head flying height compared to common debris size.

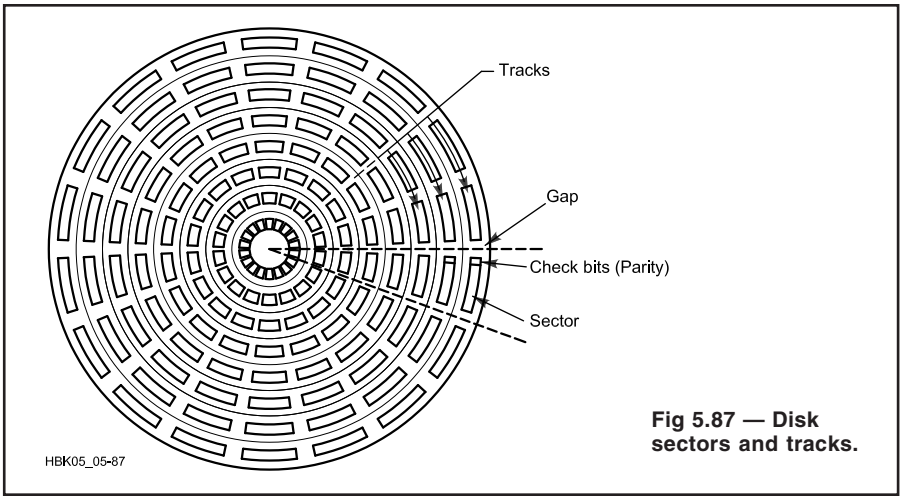


Fig 5.87 — Disk sectors and tracks.

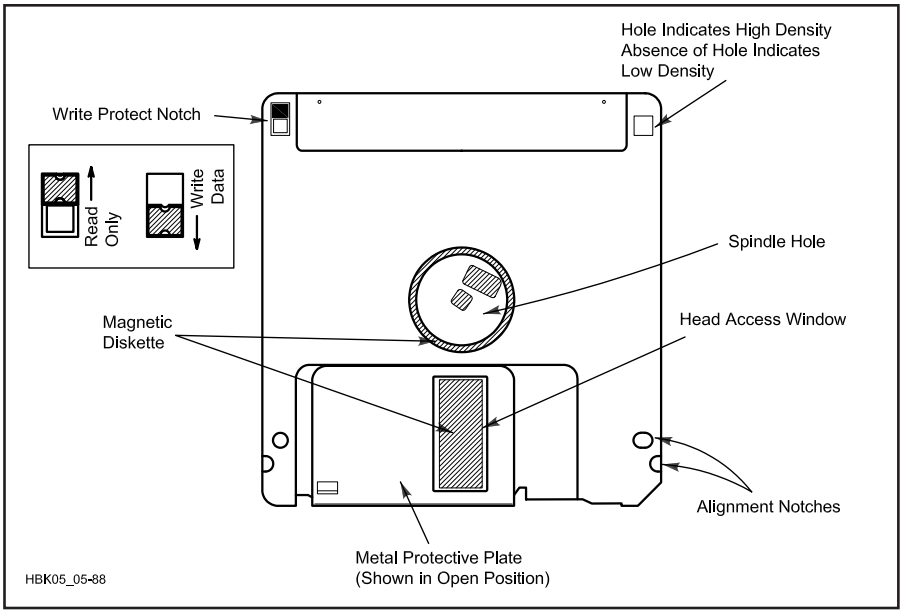


Fig 5.88 — Standard 3.5-in. floppy disk.

directly for the new drive.

Occasionally, when installing a hard drive in an older PC, the motherboard must be informed of the number of disk heads and sectors available. Often, the number entered is not the actual number of physical heads and sectors, as intelligent controllers map data requests to the proper physical location without burdening the CPU with the fine details.

Important hard disk drive parameters include bus type (for compatibility with your system), storage capacity, average and worst-case seek time, and size of onboard cache. Seek times are dominated by mechanical considerations, such as the time the read head arms require to settle from track to track or from one edge to the other, and by the rotational speed of the disk. Better overall performance is expected with larger on-board cache memory. The best drives have high capacity, fast seek times, and large caches.

Floppy Disk Drive

The most common storage peripheral in the PC is the 3.5-in., 1.44 MB floppy disk drive. The floppy drive operates similarly to its higher speed/higher capacity brother, but its media is removable and it does not offer much on-board intelligence. Floppy disks enclose the magnetic-media platter in a protective casing, as shown in **Fig 5.88**, so the disk can be carried around. The floppy disk can be inserted into a disk drive and the read/write head automatically extended. When the recording is complete, the read/write head is automatically retracted before the disk is ejected from the drive.

Most standard PCs contain a cable to connect a single floppy. This cable may be replaced with one that allows connection of two floppies. Some PCs have a cable that permits the connection of two floppies. The first drive, drive A, attaches to the *end* of the standard "twisted" control cable; if only one drive is used, leave the middle connector free. The twist in the cable causes the floppy controller to recognize the end drive as **drive A:** while the second drive — if you connect one — registers as **drive B:**, and is before the twist.

Other Magnetic Disk Drives

Disk drives featuring moderate storage capacity and performance with removable media are available, such as the Iomega Zip and Jazz, and the Imation SuperDrive. These drives use the same basic techniques as the hard drives, but with somewhat looser mechanical tolerances that allow an unsealed disk chamber and more flexible media. These drives are available with EIDE, SCSI, parallel port, USBs, and PC-Card interfaces for internal or external use. Previously very widely used for both back-up and transferring data from PC to PC, these proprietary drives are now competing with very inexpensive hard drives and USB connected hard drives. Adding a second hard drive and cloning the first hard drive (Drive C:) to the second is a popular back-up technique. Using a USB-connected hard drive for data transfer also is very convenient and popular.

Optical Storage

The audio compact disc evolved into the CD-ROM offering over 600 MB of storage on a very low-cost, single-sided platter. Initially a read-only media, writeable and re-writable discs are used as removable mass storage. While media cost is low, the low performance and restricted number of write sessions, even on the re-writable discs, prevent these technologies from competing with the magnetic hard drive except in archival or other applications where removability is important. Another transplant from the consumer entertainment industry, the DVD is another important optical storage format. Similar in appearance to CD-ROM discs, DVDs offer up to 8.5GB of storage capability per disc.

Both the CD-ROM and the DVD use a laser and an optical system to detect the surface deformities that represent data. Unlike the hard disk and floppy drives, which employ concentric tracks, the optical drives use a spiral pattern that works out from the center. Also, more elaborate data handling is necessary since the raw data error rate is significantly higher than that from magnetic drives.

All drives and disks eventually fail, and the data on the disk can be lost. Therefore, it is prudent to make backup copies of your disks, stored in a clean, dry, cool place.

Tape

Tape is one of the more inexpensive options for auxiliary memory. Tape access time is slow, since the data must be accessed sequentially, so tape is primarily used for backup copies of a system's hard drive. Tape is available in cassette form (common sizes are comparable to the cassettes for a portable tape player and VCR tapes) and on digital-audio tape (DAT). A single 4-mm-wide DAT cartridge, which fits in the palm of your hand, can hold over 2 gigabytes (GB) of data (1 GB = 1000 MB). Tape units are rarely used now in home PCs, but they are often seen at hamfests as obsolescent equipment.

Modems

Nearly all computers assembled today include a modem for connecting to the Internet via standard telephone lines. Besides connecting to an ISP or online service, this peripheral may call another modem-equipped PC or send standard facsimiles. The so-called 56-k modem uses V.90 protocol with its sophisticated DSP techniques providing echo cancellation and dynamic line equalization. This protocol uses the telephone line to gain every possible bit per second of data transfer rate. While the data rate never reaches 56 kbps, if the modem is less than four miles from the telephone central office, expect speeds of 40 kbps to 53 kbps. Achieving the 56 kbps rate would necessitate increasing transmission power above the –9 dBm limit and could cause excessive crosstalk between lines. The V.90 specification has an unusual characteristic in that the bit rate is non-symmetrical. The modem originating a call is limited to approximately 33 kbps while the answering modem runs up to 53 kbps. This compromise was made because most modem traffic is between a user and an ISP, where the user downloads much more data from the World Wide Web than he uploads.

Modems are available in three configurations, internal, external, and PC-Card. Internal modems plug directly into the computer motherboard, external devices attach to a serial port, and PC-Card (often called PCMCIA) modems plug into a PC-Card slot. The internal version is the least expensive and most common. PC-Card modems are popular with notebook computers due to their small size. External modems offer the advantage of providing immediate visual feedback of all data transfer activity. Additionally, external modems provide another level of surge protection to the computer system; if a destructive surge travels through the phone line it *might* be stopped outside the PC cabinet by the external modem.

Input Devices

The keyboard is probably the most familiar input device. A keyboard simply makes and breaks electrical contacts. The open or closed contacts are usually sensed by a microprocessor built into the circuit board under the keys. This microprocessor decodes the key closures and sends the appropriate ASCII code to the main computer unit. Keyboards will generate the entire 128-character ASCII set and often, with CONTROL and ALT (Alternate) keys, the 256-character extended ASCII set.

The mouse is a close second in familiarity to the keyboard. This pointing device controls the position of a cursor on the screen, and switches on the mouse make and break connections (clicking) to select and activate items (icons) on the screen. Touchpads, trackballs, and pen input on a sensitive screen are variations of the mouse and may offer a more natural, human-friendly interface to the computer. Voice recognition systems promise even easier data entry.

Digital cameras and streaming video cameras allow quick transfer of images into the computer realm.

Image scanners digitize photographs and older printed pages, allowing reuse of material without laboriously recreating the work. Scanners are available in four major configurations: handheld, sheet feed, flatbed, and drum. The handheld scanner is low cost and very portable. The sheet feed scanner is also small and is easier to use, since pages are better aligned. Flatbed scanners are an economical and a moderately high resolution means of entering data from book or other bound sources. Drum scanners provide the best resolution and color reproduction, but their high cost relegates them to professional graphics shops.

Output Devices

The most familiar output device is the computer screen, or monitor. The next most common output device is the printer, to produce paper hardcopy. Sound cards provide high-fidelity-stereo audio.

Monitors and most printers share a common display technique: images, such as characters and graphics, are formed by tiny dots, called *pixels* (picture elements). On screens, these are dots of light turned on and off. In printers, they are dots of ink or electrostatic toner imposed onto the paper. For color displays, pixels in red, green and blue (RGB) are spaced closely together and appear as colors to the human eye.

Video Displays

Video monitors are usually specialized high-resolution cathode-ray tube (CRT) displays, except in notebook computers, which use screens fabricated with liquid-crystal displays (LCDs). Most monitors employ *raster scanning* techniques to turn on the screen pixels, similar to that used by standard broadcast television receiv-

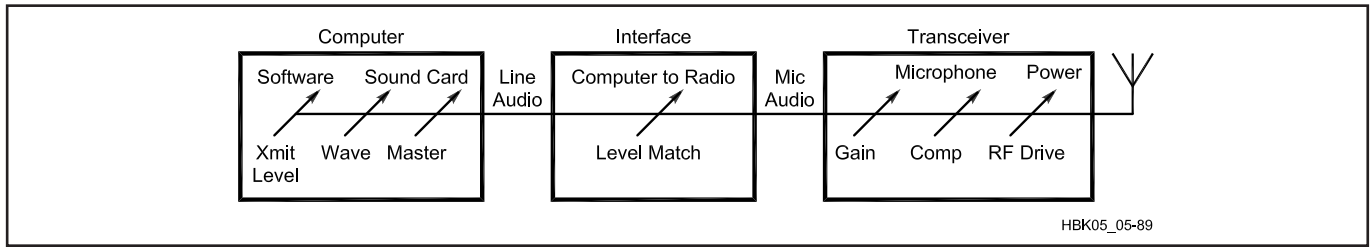


Fig 5.89 — An example of daisy-chained transmit level controls. Note that the controls are in series, and the resultant output is affected by each of the prior control settings.

ers. The electron beam paints the screen one row of pixels at a time, from left to right and top to bottom. Then, a vertical retrace brings the beam back to the top of the screen to begin again. Raster scanning signals every pixel on or off for each screen pass.

Printers

Printers suitable for hamshack use generally fall into two categories, inkjet and laser. The inkjet printer uses a controlled spray of liquid ink to produce images. Photographic-quality full-color prints are possible when the proper paper and ink is employed.

Laser printers produce exceptionally crisp text and graphics in black or a few colors at relatively high speed and low cost per page. While color laser printers do not (yet) produce the lifelike quality images of inkjets, they are not as fussy about the quality of the paper used, and its powdered *toner*, or "ink," does not dry up when stored in the printer over time as does the inkjet pigment. The use of dry paper is important, especially with many inkjet printers.

Today's state of the art in manufacturing and construction has made possible a host of very inexpensive inkjet printers, both black and white and color. There are several disadvantages to many of these units. First, and most noticeable, is the cost of replacement cartridges, especially color cartridges. It is not uncommon to pay under $100 for one of these printers and discover that two or three sets of cartridges, easily used up in a single year, cost the same as the printer. Some manufacturers have configured their printers so that when the ink goes below a certain level, the printer stops and no further operation is possible. Thus even if you wish to print in black and white, an exhausted color cartridge might prevent you from doing so. Finally, some printers are configured with a *smart chip* that prevents you from manually refilling a cartridge. The refilled cartridge may be full; the chip tells the printer it is empty.

Sound Cards

Your PC will produce and record full CD-quality audio when a suitable sound card and speaker system is deployed. Microphone and auxiliary inputs and line level outputs on the sound card let the PC serve as a contest voice keyer. When the proper software is used, it can also serve as a RTTY, CW, Pactor, PSK31, etc, terminal.

PCs running various versions of the Windows operating system do have a major annoyance in their use of soundcards. **Figure 5.89** shows a typical installation, where a PC with a sound card is the source of transmitted data. There may be as many as three level or gain controls on the PC software, perhaps one in the interfacing unit and effectively one to three in the transmitter. While all controls in the interface units and the transmitter tend to remain set, unless manually changed, that is not always true for the controls in the PC software. Some of these level set controls may revert back to a nominal value after re-booting the PC, and often several tries will be needed to set a system so that the constantly reverting levels do not have to be touched. For further information on this problem, see *QST*, Oct 2003, page 33, *The Ins and Outs of a Sound Card*.

COMMUNICATIONS: INTERNAL AND EXTERNAL INTERFACING

Designing an interface, or simply using an existing interface, to connect two devices involves a number of issues. For example, digital interfacing can be categorized as parallel or serial, internal or external and asynchronous or synchronous. Additional issues are the data rate, error detection methods and the signaling format or standards. The format can be especially important since many standards and conventions have developed that should be taken into consideration. This section focuses on some basic concepts of digital communications for interfacing between devices.

Parallel Versus Serial Signaling

To communicate a word to you across the room, you could hold up flash cards displaying the letters of the word. If you hold up four flash cards, each with a letter on it, all at once, then you are transmitting in parallel. If instead, you hold up each of the flashcards only one at a time, then you are transmitting in serial. *Parallel* means all the bits in a group are handled exactly at the same time. *Serial* means each of the bits is sent in turn over a single channel or wire, according to an agreed sequence. **Fig 5.90** gives a graphic illustration of parallel and

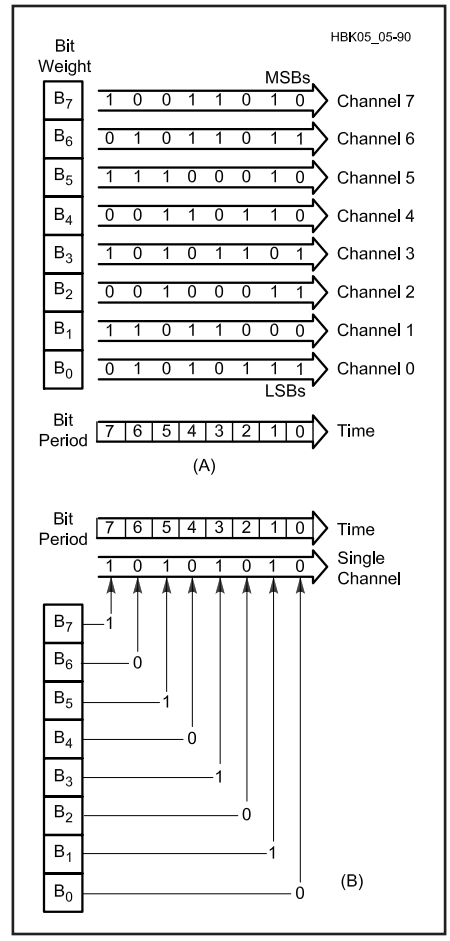


Fig 5.90 — Parallel (A) and serial (B) signaling. Parallel signaling in this example uses 8 channels and is capable of transferring 8 bits per bit period. Serial transfer only uses 1 channel and can send only 1 bit per bit period.

Interfacing To The Parallel Port

While there is a choice of ports on the PC for Amateur use and direct interfacing, the parallel port is probably the simplest. With eight data wires, several control wires and bidirectional capability, it offers a convenient way to get information in and out of a PC. The examples in the next two sections use an older software language, BASIC or GWBASIC to get information in and out. Newer languages can be used, however several varieties of BASIC are available on the internet at no cost, and the learning curve for someone who has never used a programming language is very short — usually a matter of a few minutes. The two examples that follow interface single chip analog to digital and digital to analog converters to the parallel port of a PC.

Single-Chip Dual-Channel A/D

In this analog world, often there is need to measure an analog voltage and convert it to a digital value for further processing in a PC. This single chip converter and accompanying software performs this task for two analog voltages.

Circuit Description

The circuit consists of a single-chip A/D converter, U2, and a DB-25 male plug (**Fig A**). Pins 2 and 3 are identical voltage inputs, with a range from 0 to slightly less than the supply voltage V_{CC} (+5 V). RI, R2, C3 and C4 provide some input isolation and RF bypass. There are four signal leads on U2. DO is the converted data from the AID out to the computer; DI and CS are control signals from the computer, and CLK is a computer-generated clock signal sent to pin 7 of U2.

The +5-V supply is required. It may be obtained from a +12-V source and regulator UI. Current drain is usually less than 20 mA, so any 5-V regulator may be used for UI. The power supply ground, the circuit ground and the computer ground are all tied together. If you already have a source of regulated 5 V, UI is not needed.

In this form the circuit will give you two identical dc voltmeters.

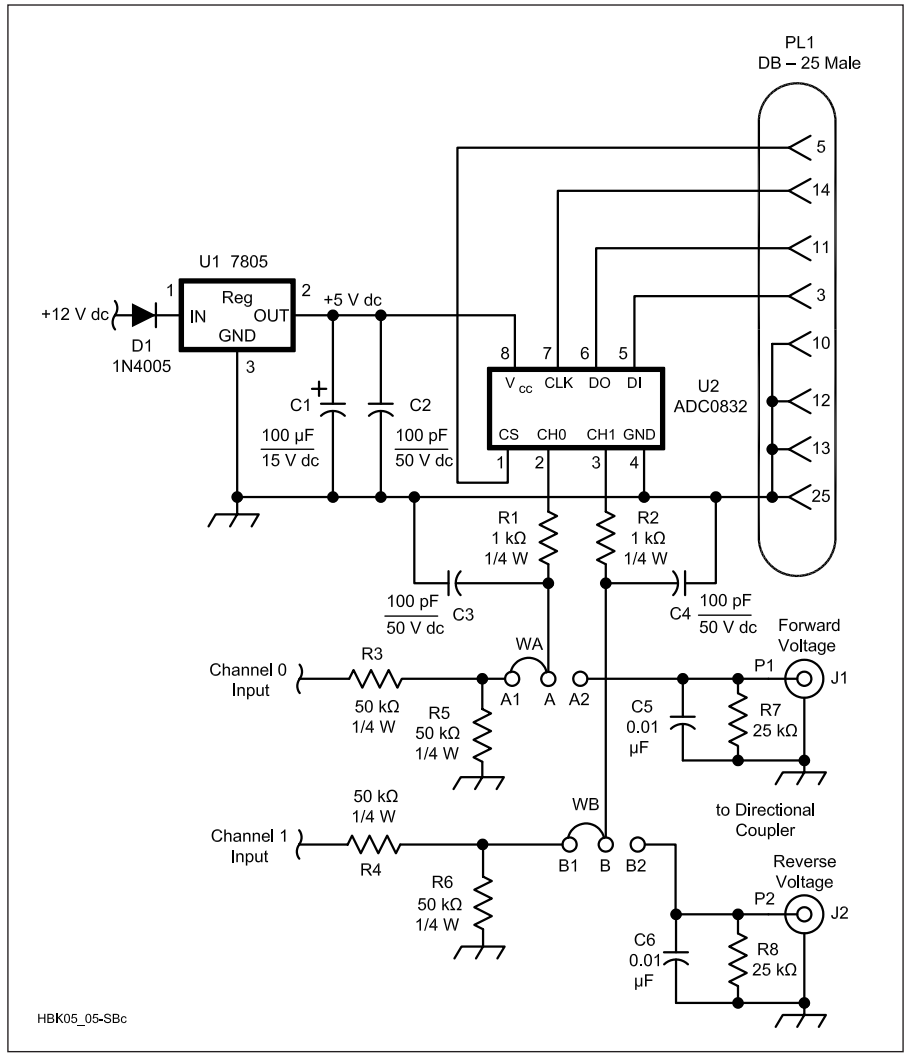


Fig A — Only two chips are used to provide a dual-channel voltmeter. PL1 is connected through a standard 25-pin cable to your computer printer port. U2 requires an 8-pin IC socket. All resistors are ¼ W. You can use the A/D as an SWR display by connecting it to a sensor such as the one shown in Chapter 19 of this *Handbook* (Tandem Match Wattmeter project). A few more resistors are all that are needed to change the voltmeter scale. The 50-kΩ resistors from 2:1 voltage dividers, extending the voltmeter scale on both channels to almost 10 V dc.

To extend their range, connect voltage dividers to the input points A and B. A typical 2:1 divider, using 50-kΩ resistors, is shown in the figure. Resistor accuracy is not important, since the circuit is calibrated in the accompanying software.

Software

The software, *A2D.BAS*, includes a voltmeter function and an SWR function. It is written in *GWBASIC* and saved as an ASCII file. Therefore, you can read it on any word processor, but if you modify it, make sure you re-save it as an ASCII file. It can be imported into *QBasic* and most other Basic dialects.

The program was written to be understandable rather than to be highly efficient. Each line of basic code has a comment or explanation. It can be modified for most PCs. The printer port used is LPT1, which is at a hex address of 378h. If you wish to use LPT2

(printer port 2), try changing the address to 278h. To find the addresses of your printer ports, run *FINDLPT.BAS*.

A2D.BAS was written to run on computers as slow as 4.7-MHz PC/XTs. If you get erratic results with a much faster computer, set line 1020(CD=1) to a higher value to increase the width of the computer-generated clock pulses.

The software is set up to act as an SWR meter. Connecting points A and B to the forward and reverse voltage points on any conventional SWR bridge will result in the program calculating the value of SWR.

Initially the software reads the value of voltage at point A into the computer, followed by the voltage at point B. It then prints these two values on the screen, and computes their sum and difference to derive the SWR. If you use the project as a voltmeter, simply ignore the SWR reading on the screen or suppress it by deleting lines 2150, 2160 and 2170. If the two voltages are very close to each other (within 1 mV), the program declares a bad reading for SWR.

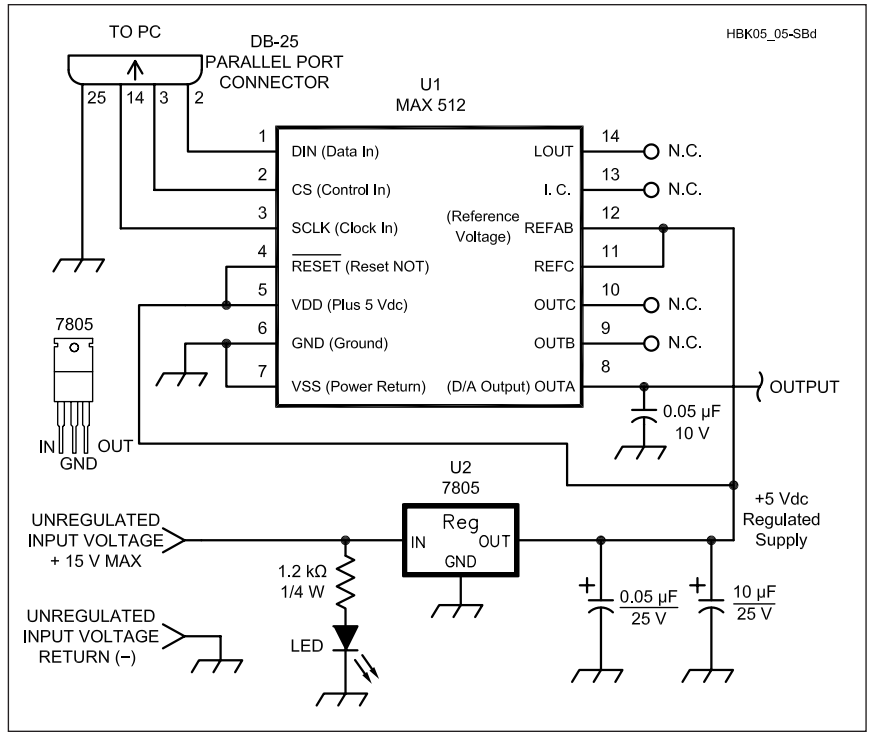


Fig B — Only three wires and a ground lead are needed to connect the converter to your PC.

Calibration

Lines 120 and 130 in the program independently set the calibration for the two voltage inputs. To calibrate a channel, apply a known voltage to input point A. Read the value on the PC screen. Now multiply the constant in line 120 by the correct value and divide the result by the value you previously saw on the screen. Enter this constant on line 120. Repeat the procedure for input point B and line 130.

D to A CONVERTERS — CONTROLLING ANALOG DEVICES

The complement to A/D converters is D/A (digital-to-analog) converters. Once there is a digital value in your PC, a D/A will provide an analog voltage proportional to the digital value. Normally the actual value is scaled. As an example, an 8-bit converter allows a maximum count of 255. If the converter is set up with a +5 Vdc reference voltage, a maximum value digital value of 255 would result in a D/A output value of 5-V. Lower digital inputs would give proportionally lower voltages.

Circuit Description

This project is the complement of the parallel port A/D converter described earlier. It takes a digital number from the computer, and converts it to a voltage from 0 to 5 V dc. Only one chip, the MAX 512, is required. It operates from a 5-V supply and is connected to the computer by a standard DB-25 parallel port connector. The chip may be ordered from Digi-Key, Allied Electronics and other ham suppliers as MAX512CPD-ND. The voltage regulator in **Fig B** provides the 5-volt source required to power the chip.

Software

The software needed to run the chip, *D2A.BAS*, can be downloaded from Internet sites. It is about 60-lines long, fully commented and written in *GWBASIC*, so it may be readily modified. The parallel port address is defined on line 105 as PORTO=&H378. Your computer may use a different address. To find the correct address, run *FINDLPT.BAS*.

The program takes the value AIN from the keyboard (line 230), converts it to a number between 0 and 255, and then sends it out as a serial word to the DIA chip. If you would like to use the project with another program, use your other program to set AIN to the value you want to generate, and then run this program as a subroutine.

At the end of the program is the clock pulse subroutine. In the event your computer is too fast for the converter chip, you can stretch the clock pulses by changing CD in line 5010 to a value greater than the default value of 1.

Applications

This circuit provides the capability of setting a voltage under computer control. It can be calibrated to match the power supply and the actual chip used. Tests with several chips showed an error of 25 mV or less over the range of 0 to 5 V dc output.

— *Paul Danzer, N1II*

serial signaling.

Both parallel and serial signaling are appropriate for certain circumstances. Parallel signaling is faster, since all bits are transmitted simultaneously, but each bit needs its own conductor, which can be expensive. Parallel signaling is more likely to be used for internal communications. For spanning longer distances, such as to an external device, serial signaling is more appropriate. Each bit is sent in turn, so communication is slower, but it is also less expensive, since fewer channels are needed between the devices.

Most amateur digital communications use serial transmission to minimize cost and complexity. The number of channels needed for signaling also depends on the operational mode. One channel is required per bit for simplex (one-way, from sender to receiver only) and for half-duplex (two-way communication, but only one person can talk at a time), but two channels per bit are needed for full-duplex (simultaneous communications in both directions).

Parallel I/O Interfacing

Fig 5.91 shows an example of a parallel input/output chip. Typically, they have eight data lines and one or more handshaking lines. *Handshaking* involves a number of functions to coordinate the data transfer. For example, the READY line indicates that data are available on all 8 data lines. If only the READY line is used, however, the receiver may not be able to keep up with the data. Thus, the STROBE line is added so the receiver can determine when the transmitter is ready for the next character.

On standard PCs, a parallel port is available using a 25-pin DB-25 connector. This port was originally intended for use with a printer. Several versions of the parallel port exist, and late-model PCs feature a high speed, bidirectional parallel port that, in addition to high-speed printing, may also interface with mass storage devices, scanners, and other I/O devices. The most common parallel ports are:

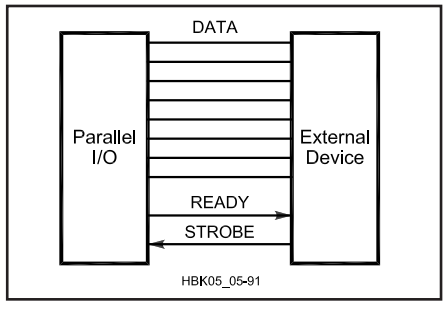


Fig 5.91 — Parallel interface with READY and STROBE handshaking lines.

- Printer Mode — The most basic, output mode only port.
- Standard & Bidirectional (SPP) — The low-speed bidirectional port
- Enhanced Parallel Port (EPP) — Uses local hardware handshaking and strobing to accomplish 500 kB/s to 2 MB/s transfer rates.
- Extended Capabilities Port (ECP) — Similar to EPP, except negotiates a reverse channel with the external peripheral and requires that peripheral controls handshaking. It is optimized for the Windows operating system and uses DMA channels, a FIFO buffer, and real time data compression of up to 64:1.

Most PCs offer a choice of port protocol in the BIOS setup. Unless you have a reason to do otherwise, select the "ECP and EPP 1.9 Mode" for maximum flexibility and performance.

Serial I/O Interfacing

Serial input/output interfacing is more complex than parallel, since the data must be transmitted based on an agreed sequence. For example, transmitting the 8 bits (b7, b6, . . . b0) of a word includes specifying whether the least significant bit, b0, or the most significant bit, b7, is sent first. Fortunately, a number of standards have been developed to define the agreed sequence, or encoding scheme. Use only IEEE specified cable to ensure availability of the correct number of conductors.

Data Rate

There are a number of limitations on how fast data can be transferred: (1) The sending equipment has an upper limit on how fast it can produce a continuous stream of data. (2) The receiving equipment has an upper limit on how fast it can accept and process data. (3) The signaling channel itself has a speed limit, often based on how fast data can be sent without errors. (4) Finally, standards and the need for compatibility with other equipment may have a strong influence on the data rate.

Two ways to express data transmission rates are *baud* and *bits per second (bps)*. These two terms are not interchangeable: Baud describes the signaling, or symbol, rate — a measure of how fast individual signal elements *could* be transmitted through a communications system. Specifically, the baud is defined as the reciprocal of the shortest element (in seconds) in the data- encoding scheme. For example, in a system where the shortest element is 1-ms long, the maximum signaling rate would be 1000 elements per second. (Note that, since baud is measured in elements per second, the term "baud rate" is incorrect since baud is already a measure of speed, or rate.) Continuous transmission is not required, because signaling speed is based only on the shortest signaling element.

Signaling rate in baud says nothing about actual information transfer rate. The maximum information transfer rate is defined as the number of equivalent binary digits transferred per second; this is measured in bits per second.

When binary data encoding is employed, each signaling element represents one bit. Complications arise when more sophisticated data encoding schemes are used. In a quadrature- phase-shift keying (QPSK) system, a phase transition of 90° represents a level shift. There are four possible states in a QPSK system; thus, two binary digits are required to represent the four possible states. If 1000 elements per second are transmitted in a quadriphase system where each element is represented by two bits, then the actual information rate is 2000 bps.

This scheme can be extended. It is possible to transmit three bits at a time using eight different phase angles (bps = 3 × baud). In addition, each angle can have more than one amplitude. A 9600 bps modem uses 12 phase angles, 4 of which have two amplitude values. This yields 16 distinct states, each represented by four binary digits. Using this technique, the information transfer rate is four times the signaling speed. This is what makes it possible to transfer data over a phone line at a rate that produces an unacceptable bandwidth using simpler binary encoding. This also makes it possible to transfer data at 2400 bps on 10 m, where FCC regulations allow only 1200-baud signals.

When are transmission speed in bauds and information rate in bps equal? Three conditions must be met: (1) binary encoding must be used, (2) all elements used to encode characters must be equal in width and (3) synchronous transmission at a constant rate must be employed. In all other cases, the two terms are not equivalent.

Within a given piece of equipment, it is desirable to use the highest possible data rate. When external devices are interfaced, it is normal practice to select the highest standard signaling rate at which both the sending and receiving equipment can operate.

Error Detection

Since data transfers are subject to errors, data transmission should include some method of detecting and correcting errors. Numerous techniques are avail-

able, each used depending on the specific circumstances, such as what types of errors are likely to be encountered. Some error detection techniques are discussed in the Modes and Modulation Sources chapter. One of the simplest and most common techniques, parity check, is discussed here.

Parity Check

Parity check provides adequate error detection for some data transfers. This method transmits a parity bit along with the data bits. In systems using odd parity, the parity bit is selected such that the number of 1 bits in the transmitted character (data bits plus parity bit) is odd. In even parity systems, the parity bit is chosen to give the character an even number of ones. For example, if the data 1101001 is to be transmitted, there are 4 (an even number) ones in the data. Thus, the parity bit should be set to 1 for odd parity (to give a total of 5 ones) or should be 0 for even parity (to maintain the even number, 4). When a character is received, the receiver checks parity by counting the ones in the character. If the parity is correct, the data is assumed to be correct. If the parity is wrong, an error has been detected.

Parity checking only detects a small fraction of possible errors. This can be intuitively understood by noting that a randomly chosen word has a 50% chance of having even parity and a 50% chance of having odd parity. Fortunately, on relatively error-free channels, single-bit errors are the most common and parity checking will always detect a single bit in error. However, an even number of errors will go undetected, whereas an odd number of errors will be detected. Parity checking is a simple error detection strategy. Because it is easy to implement, it is frequently used. Other more complex techniques are used in commercial data transmission services.

Standard Interface Busses

Signaling Levels

Inside equipment and for short runs of wire between equipment, the normal practice is to use neutral keying; that is, simply to key a voltage such as + 5 V on and off. In neutral keying, the off condition is considered to be 0 V. Over longer runs of wire, the line is viewed as a transmission line, with distributed inductance and capacitance. It takes longer to make the transition from 0 to 1 or vice versa because of the additional inductance and capacitance. This decreases the maximum speed at which data can be transferred on the wire and may also cause the 1s and 0s to be different lengths, called bias distortion. Also, longer lines are more likely to pick up noise, which can make it difficult for the receiver to decide exactly when the transition takes place. Because of these problems, bipolar keying is used on longer lines. Bipolar keying uses one polarity (for example +) for a logical 1 and the other (– in this example) for a 0. This means that the decision threshold at the receiver is 0 V. Any positive voltage is taken as a 1 and any negative voltage as a 0.

EIA-RS-232

The most common serial bus protocol, EIA-RS-232, addresses this issue (however, a Mark "1" is a negative voltage and a Space "0" is positive). Generally called RS-232, this protocol defines connectors and voltages between data terminal equipment (DTE) such as a PC, and data communications equipment (DCE), such as a modem or TNC. The connector is the DB-25, or the presently more popular DB-9 version. Signaling voltages are defined between + 3V and + 25V for logic "0" and between – 3V and – 25V for logic "1." Although the top data rate addressed in the specification is only 20 kbps, speeds of up to 115 kbps are commonly used. Communications distances of hundreds of meters are possible at reasonable data rates.

Since neutral keying is usually used inside equipment and bipolar keying for lines leaving equipment, signals must be converted between bipolar and neutral. Discrete level shifters or op amp circuits may perform this task, or low cost specialized IC line drivers and receivers are available.

RS-422

RS-422 is a serial protocol similar to RS-232, but employing fully differential data lines. Differential data offers the important advantage that common grounds between remote units are not necessary, and an important cause of ground loops (and their associated problems) is eliminated. Available on many Apple Macintosh computers, RS-422 systems may connect to standard RS-232 modems and TNCs by building a cable that makes the following translations:

RS-422 DTE	*RS-232 DCE*
RXD–	RXD
TXD–	TXD
RXD+	GND
TXD+	No Connection
GPi	CD

IrDA (Infrared Data Access)

Another high-speed serial protocol is the IrDA, which is a simple, short range wireless system using infrared LEDs and detectors. Data rates up to 3 MB/sec are possible between compatible units.

Universal Serial Bus (USB)

USB is a computer standard for an intelligent serial data transfer protocol. In addition to its higher speed than RS-232, USB offers reasonable power availability to its loads, or *functions*. Under certain circumstances, up to 127 hubs and functions may connect to a single computer. USBs requires that each function have onboard intelligence and that it negotiate with the host for power and bandwidth allocation, and has the major advantage of *hot-pluggablity* — the PC need not reboot when new functions are added. The USB connectors use four-conductor cable, with two bidirectional, differential data lines, power, and ground. Approximately 5 V at 100 mA is allowed per function, with up to 500 mA available if the host system has the capability. This means that relatively sophisticated devices, such as modems, small video cameras, or hand-held scanners may operate from the bus without additional power supplies. Preventing power back-flowing up from function to host is accomplished by configuring the connector shapes such that the host has a rectangular connector while that of functions are nearly square.

There are currently two USB standards. USB 1.1, somewhat obsolete but common in PCs just a few years old, is capable of 12 megabits per second (12 Mbps). USB 2.0 is the later standard and is rated up to 480 Mbps. Most USB 2.0 ports will allow the use of older USB 1.1 devices—that is they are backwards compatible. However maximum cable lengths and available power to devices may be affected.

IEEE-1394 (FireWire)

A very high speed serial protocol, IEEE-1394 (christened "FireWire" by its creator, Apple Computer), is capable of up to 400 Mbps of sustained transfer. It is ideal for high bandwidth systems, such as live video, external hard drives, or high-speed DVD player/recorders. Up to 63 devices may daisy-chain together at once via a standard six-wire cable. Unlike USB, 1394 is peer-to-peer, meaning any device may initiate a data transfer — the PC does not have to initiate a data transfer. Similar to USB, IEEE-1394 is hot-pluggable and provides power on the cable, but the voltage may vary from 7 V to almost 40 V, and may be sourced by any device. Allowable current drain per device may reach 1 A.

PC-Card (PCMCIA)

The PC-Card Standard is a collection of specifications for miniature plug-in peripherals. The most common is colloqui-

ally referred to as PCMCIA cards — 68-pin devices the size of thick credit cards that contain a modem, LAN, GPS receiver, USB port, FireWire port, high resolution video, an extra serial or parallel port, or memory storage expansion. The standard PC-Card allows up to 5 V at 1 A peak current, or 3.3 V at 1 A, depending upon configuration. Other voltages may be used if available from the host. Other portions of the specification define memory storage-only cards in even smaller footprints.

Small Computer System Interface (SCSI)

SCSI interfaces provide for up to 320 MB/s transmission rates with up to 15 devices. Used mostly with disk drives, the "skuzzy" bus also supports a very wide variety of high-speed peripherals. A wide variety of bus widths, speeds, and connectors exist. With the widespread use of USB, SCSI is no longer generally used except where large numbers of storage disks are needed — primarily in commercial installations.

10Base2, 10BaseT, 10Base4, 100BaseT

Common office/home networks use 10BaseN protocol. 10Base2 is generally recommended for Amateur Radio installations, since it uses shielded cable (RG-58, renamed "thin coax" in this application). Additionally, no separate hub is needed as the connected computers work on a peer-to-peer basis. A drawback of 10Base2 is its maximum data rate is limited to 10 Mbps. Also, shorter runs may occasionally require a simple extension in length of approximately 30-50 feet for proper operation. The other protocols use one or more hubs, RJ-45 connectors, and unshielded Category 5 cable. 100BaseT systems are rated to 100 Mbps.

Explanation of the *MBaseN* and *Category N* terminology can be found in many available networking books. Newer home networks are wireless; often, they are more trouble-free from high power Amateur Radio installations because the unshielded cable used for wired networks tends to pick up RF readily. Unfortunately, high power VHF and UHF installations may pose problems with wireless networks.

POWER SYSTEMS AND ATX

When the initial personal computers were designed, the most common logic family was TTL and CMOS that interfaced with TTL levels at 5 V. Disk drive and fan motors preferred +12 V. RS-232 demanded a higher voltage bipolar supply, so –12 V was added to complement the 12 V already used. The analog portion of the early modems required –5 V. Thus, the initial "silver box" PC supply provided +5 V at high current, +12 V at moderate current, and –5 V and –12 V at low current. Advances in semiconductor technology allowed shrinking transistor geometries. The smaller transistors were faster, but they had lower breakdown voltages, thus a new logic voltage of 3.3 V was introduced. Initially, computer manufacturers responded by placing IC regulators on the motherboard to power the 3.3-V circuits, but eventually the current demanded by these circuits exceeded that of the traditional +5V components and a new physical standard, called ATX, was introduced.

The ATX standard defines a layout physically different from older "AT-type" computers. The computer case, motherboard mounting holes, expansion slot location, and the power supply and its connector are all changed. ATX computers are recommended for the hamshack due to their better RFI control, resulting from careful mechanical design of connectors and consideration of card slot case penetration. ATX power supplies produce +3.3 V, +5 V, +12 V, –5 V and –12 V. They also provide an output voltage, even when the power supply is otherwise off, allowing *sleep mode* operation. Sleep mode retains the computer RAM contents and configuration so a reboot is not necessary each time the computer is used, and is especially critical to extending battery life in notebook PCs.

As semiconductor technology continued improving, the 3.3 V source became too high for the fine-geometry microprocessors, and an even lower voltage was required. At present, there is no standard for the next lower voltage, but devices are available that need anywhere between 1 V to 2.8 V for the microprocessor core (densest portion). Motherboard manufacturers have addressed this issue by again using on-board voltage regulators to drop an existing ATX standard voltage to the value needed by the CPU. Since there is no standard — in fact, the exact voltage preferred by a given processor family decreases as the manufacturing process evolves — motherboards provide means of selecting the matching voltage. Sometimes this process is automatic, as the on-board power supply communicates with the microprocessor before initializing and rises to the proper voltage, but other times, jumpers must be manually positioned *before* initial power is applied. Using a higher than recommended supply voltage causes excessive operating temperature and stresses the gates of the CMOS transistors, leading to reduced reliability and early (sometimes immediate) circuit death.

Power supplies removed from older computers are readily available, and there is a great temptation to reuse them for other amateur applications. Unfortunately, there is very little control on the manufacturers of these units, and quite often the expected *5-VDC, 10-amp* output is poorly regulated except at currents such as one or two amperes. Many of the newer units do not have an on-off switch, but depend on connecting one lead to ground for some period to turn the supply on and off. Even when off, a portion of the supply is on to provide control voltages. Many computer handbooks and texts contain tables of the connections to the various PC power supplies, complete with wire colors. However, there is no way to tell if compliance with the color codes has been followed. Hence, use of these supplies should be limited to those voltage sources that can be measured on a particular supply.

Powering Small Circuits

Small amounts of power can be 'stolen' from the serial port to power devices interfacing with a PC. The signals on the serial port, in accordance with the various versions of RS-232, generally range from +15 VDC to –15 VDC. By selecting several control signals and knowing that one of these signals will be at the positive voltage level, the power source in **Fig 5.92** can be used. The amount of current available will vary depending on the specific interface chip and chip supplier for the particular PC.

By reversing the diodes and filter capacitor a negative supply can be built. The regulator should be chosen for the negative or positive supply, as needed. Most PCs will not have any difficulty supplying 10 ma, or slightly more.

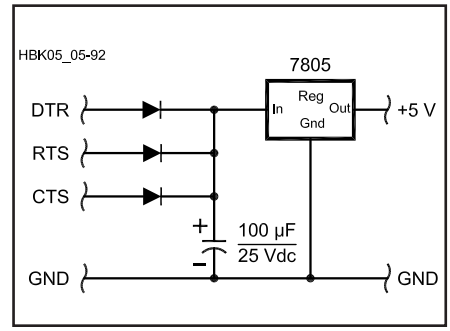


Fig 5.92 — Small amounts of power can be 'stolen' from the serial port with this circuit. See the Component Data and References chapter for the 7805 pin connections.

Power Quality

As operating voltages drop, power quality — the measure of voltage accuracy and transient response to changing load currents — becomes simultaneously more critical and more difficult. A 500-mV spike represents a 10% error in a 5-V supply, but the same 500-mV spike applied to a 2-V processor represents an overvoltage of 25%, grossly exceeding the maximum rated supply voltage for that controller. Further, if the spike is of the opposite polarity, it seriously reduces the noise margin of the logic-high levels, possibly corrupting data.

Providing clean power becomes more difficult when the effects of sleep mode are considered. Microprocessors reduce their power consumption when idle by slowing down internal clocks and other techniques, but when called back to duty, their response occurs in nanoseconds. The result is a huge change in supply current, from nearly zero to maximum current flow in those few nanoseconds. Large voltage spikes may result from this fast rise-time current step working against the inductance of the PC board power supply traces. Careful power supply design, especially during layout, and judicious use of low ESR, and low-inductance bypass capacitors mitigate the transients and keep the system reliable.

STANDARD COMPUTER CONNECTIONS

See the **Component Data and References** chapter for details on computer connector pinouts. You'll also find details on cables, such as a null modem cable.