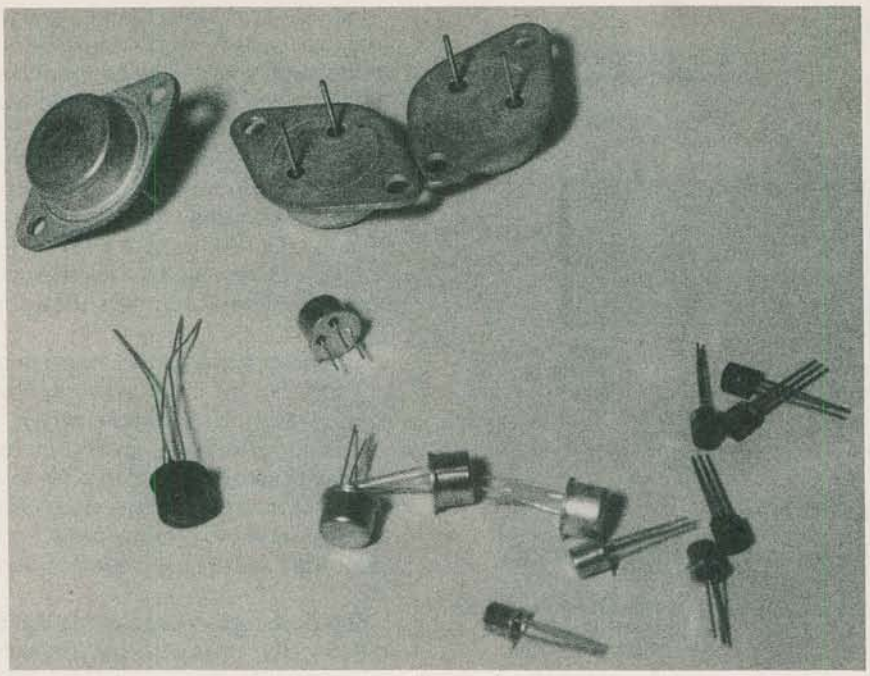

All About Transistors: Bipolar Basics

BY ROBERT A. YOUNG

We look at the tiny devices that have reshaped the world of electronics.

Along with the solid-state diode, the point-contact transistor—invented in 1947 at Bell Labs—started the semiconductor revolution and has gone on to become one of the rudimentary devices in today's electronic equipment. The transistor, whether in discrete or IC form, is at the heart of most modern circuitry. Therefore, understanding how transistors function will help you properly design circuits containing them, and in case of a failure, enable you to find and correct the problem.

Bipolar-Transistor Composition. A bipolar transistor is basically two PN junctions connected back-to-back within the same piece of semiconductor material and sharing a common P- or N-doped semiconductor region. There are two types of bipolar transistor, the NPN and the PNP. Figure 1A is a simplified illustration of the composition of the NPN type of transistor. In our illustration, the NPN type unit is shown as P-doped semiconductor material sandwiched between two layers of N-doped material. The composition of a PNP transistor is just the opposite of that, (*i.e.*, the N- and P-doped materials in the transistor are interchanged). It follows then that biasing considerations for NPN units are also opposite from those for the PNP unit.

Note from Fig. 1A that a bipolar transistor is comprised of a center region called the base surrounded by two other regions known as the collector and the emitter. The difference between them will be discussed shortly. The two junctions are arranged so that they are very close together; that's done by making the shared base region very thin and lightly doped. That causes the two junctions to interact with one another. Conduction in the collector-base junction depends largely on what happens in the emitter-base junction.

Because the base region is lightly doped, it has a relatively small number of free carriers (holes in a P-type base and electrons in an N-type base) to conduct current. On the other hand, the emitter region is quite heavily doped, containing a much larger amount of donor impurity (for the NPN type) or acceptor impurity (for the PNP type), so there are many more free carriers available in the emitter region to conduct current than in the adjacent base region. Because of that, the emitter-base junction, when forward biased, conducts much the same as a common PN-junction diode.

The current that flows (composed of electrons for NPN units and holes, in the case of PNP transistors) is mainly from the emitter to the base rather than vice versa. That is where the emitter derives its name—it emits or injects current carriers into the other regions of the device.

The third region of a transistor, the collector, is lightly doped, much the same as the base, except with the opposite type of doping impurity, so it (like the base region) has relatively few free carriers available to conduct current in the normal way. The collector-base junction is normally reverse biased, so a depletion layer forms, spreading out on either side of the junction. The depletion layer effectively removes the carriers that would otherwise balance out the charges on the fixed impurity atoms of the crystals, setting up a potential barrier to match the applied reverse voltage.

To the normal majority carriers in the base and emitter, that potential barrier is a big wall that must be overcome before they can pass to the other side. So just as in the case of a normal diode, virtually no current flows across the collector-base junction when left to its own devices. However, the junction is not left to its own devices.

Remember that the base region is deliberately made very thin and lightly doped, while the emitter is made much more heavily doped. Because of that, applying a forward bias to the emitter-base junction causes majority carriers to be injected into the base, and straight into the reverse-biased collector-base junction.

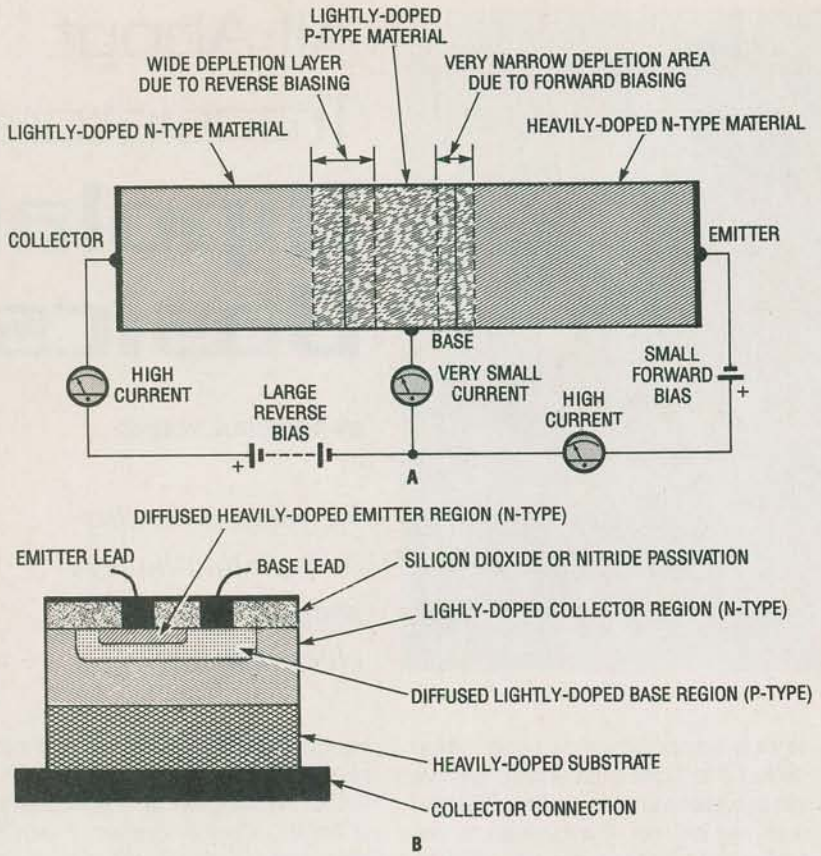

Fig. 1. Shown in A is a simplified illustration of the composition of the NPN type of transistor, consisting of a layer of P-doped semiconductor material sandwiched between two layers of N-doped material. The actual form of the modern planar double-defused epitaxial junction transistor is shown in B.

Those carriers are actually minority carriers in the base region, because that region is of opposite semiconductor type to the emitter. To those majority-turned-minority carriers, the collector-base junction depletion region is not a barrier at all but an inviting, accelerating field; so as soon as they reach the depletion layer, they are immediately swept into the collector region.

Forward biasing the emitter-base junction causes two things to happen that might seem surprising at first: Only a relatively small current actually flows between the emitter and the base, much smaller than would flow in a normal PN diode despite the forward bias applied to the junction between them. A much larger current instead flows directly between the emitter and collector regions, in this case, despite the fact that the collector-base junction is reversed biased.

That effect is illustrated in Fig. 1A, which (hopefully) will help you to understand what is going on. The diagram shows a NPN transistor, but the action in a PNP unit is similar except for the opposite region polarity and conduction mainly by holes rather than electrons.

From a practical point of view, the behavior of bipolar transistors means that, unlike the simple PN-junction diode, it is capable of amplification. In effect, a small input current made to flow between the base and emitter results in a much larger output current flowing between the emitter and collector. Only a small voltage—around 0.6 volts for a typical silicon transistor—is needed to produce the small input current required.

In contrast, the reverse-bias voltage applied across the collector-base junction can be much larger; typically anywhere from 6 to 90 volts or more. So in producing and being able to control a larger current in this much-higher-output circuit, the transistor's small input current and voltage can achieve considerable voltage, power, and current, gains.

Bipolar transistors, therefore, work very well as both amplifiers and electronic switches. That is why they have become the workhorses of modern electronics, virtually replacing the vacuum tube. The diagram in Fig. 1A is designed to show how a bipolar transistor works, rather than its physical construction. The actual form of the modern, planar, double-defused epitaxial-junction transistor is shown in Fig. 1B.

The collector region is formed from a lightly doped layer grown epitaxially on the main substrate, which is made from the same type (but more-heavily doped) material to provide a low resistance connection. Here, both are N-doped material; for a PNP transistor, they would be P-doped material.

The base region is formed by lightly diffusing the opposite type impurity into a medium-sized area of the chip surface to reverse that type of area and create the base-collector junction. The emitter region is formed by making a second and heavier diffusion over the smaller area inside the first, but this time with the same kind of impurity as used for the epitaxial collector region.

The second diffusion is very carefully controlled so that the emitter region that results extends almost—but not quite—to the bottom of the base. That leaves the area of the base right below the emitter quite thin to ensure that as many as possible of the carriers injected from the emitter region will be swept through to the collector. The thinner that active base region, the higher (in general) the gain of the transistor.

Note that although the collector and emitter regions are made of the same type of semiconductor material, the two are physically quite different. The emitter is heavily doped (for good carrier injection) and can be relatively small since the emitter-base junction does not need to dissipate much power (heat). In contrast, the collector is lightly doped (for a wide depletion area) and its junction is much larger since, being reversed biased, it must dissipate much more power.

Connections to the emitter and base regions are made by way of aluminum electrodes deposited on the surface. Tiny wires are bonded to the

electrodes for connection to the main device leads. The low-resistance substrate itself is used to connect to the collector region.

That is the basic construction used for most modern bipolar transistors, whether they are discrete units or part of an IC containing thousands of transistors. The main difference is size, although, in an IC, the collector region of the transistor will generally be in an epitaxial layer grown on the opposite kind of substrate, and separated by diffused walls (of the opposite type material) to separate the transistors from each other.

Inside an IC, the active part of an individual transistor might only be a couple of micrometers square, while a very large transistor (one used to switch hundreds of amperes) might be on a single wafer of 10 mm or more in diameter. Typical small-to-medium power, discrete transistors used in consumer and hobby electronics are grown on chips measuring from 1- to about 3-mm square—the rest of the component is protective packaging.

Transistor Operation. Refer to Fig. 2, a PNP version of the illustration shown in Fig. 1A. Note that both are essentially the same, except that in this instance, the collector is more negative than the base or the emitter. That is an important characteristic to remember when it comes to the operation of bipolar transistors.

If a positive voltage is applied to the P-doped emitter (to the left), current will be swept through the base-emitter junction—with the holes from the P-doped material moving to the right and the electrons from the N-doped material moving to the left. Some of the holes moving into the N-doped base region will combine with electrons and become neutralized, while others will migrate to the base-collector junction.

Normally, if the base-collector junction is negatively biased, there would be no current flow in the circuit. However, there would be additional holes in the junction to travel to the base-collector junction, and electrons can then travel toward the base-emitter junction, so a current flows even though that section of the sandwich is biased (at cutoff) to prevent conduction. Most of the current travels between the emitter and collector

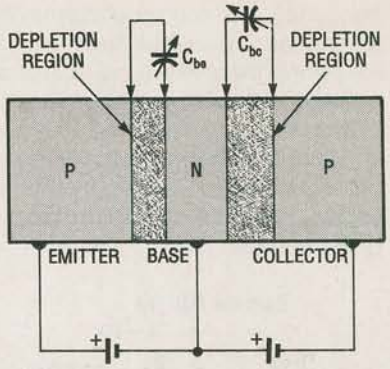

Fig. 2. Shown here is a PNP version of the illustration shown in Fig. 1A. Note that both are essentially the same, except that in this instance, the collector is now more negative than the base or the emitter.

and does not flow out through the base.

The amplitude of the collector current depends principally on the magnitude of emitter current (*e.g.*, the collector current is controlled by emitter current). Note that between each PN junction, there is an area known as the depletion or transition region that is similar in some characteristics to a dielectric layer. That layer varies in accordance with the operating voltage. The semiconductor materials on either side of the depletion regions constitute the plates of a capacitor. The base-collector capacitance is indicated in Fig. 2 as C_{bc}, and the base-emitter capacitance is designated C_{be}. A change in signal and operating voltages causes a non-linear change in those junction capacitances.

There is also a base-emitter resistance (R_{be}) that must be considered. In practical transistors, emitter resistance is on the order of a few ohms, while the collector resistance is many hundreds or even thousands of times larger. The junction capacitance in combination with the base-emitter resistance determines the useful upper-frequency limit of a transistor by establishing an RC time constant.

Because the collector is reversed biased, the collector-to-base resistance is high. On the other hand, the emitter and collector currents are substantially equal, so the power in the collector circuit is larger than the power in the emitter circuit. ($P = I^2R$, so the powers are proportional to the respective resistances, if the currents are the same.) In practical transistors, emitter resistance is on the order of a few ohms, while the collector resistance is many hundreds or thousands of times larger, so power gains of 20 to 40 dB, or even more, are possible.

Figure 3 shows the schematic symbols for both the NPN and PNP versions of the bipolar transistor. The first two letters of the designations (NPN or PNP) indicate the polarities of the voltages applied to the collector and emitter in normal operation. For example, in a PNP unit, the emitter is made more positive with respect to the collector and the base, and the collector is made more negative with respective to the base. Another way of saying that is: the collector is more negative than the base and the base is more negative than the emitter.

Transistor Amplifiers. Transistors are among the most commonly used building blocks in electronics. While they can be used as electronically controlled switches, they are widely configured for amplifier use. In fact, the vast majority of electronic circuits contain one or more amplifiers of some type or another.

However, what exactly do we mean by the term amplifier? By definition an amplifier is a circuit that draws power from a source other than the input signal and produces an output that is usually an enlarged reproduction of the input signal.

We say usually because not all amplifiers are used to magnify the input signal—buffer amplifiers (often called unity-gain amplifiers) are not designed to magnify the input signal. When operated as a buffer, the transistor is used to isolate one stage from the effects of one that follows. Since buffer amplifiers provide no increase in signal level, a 10-millivolt (mV) signal

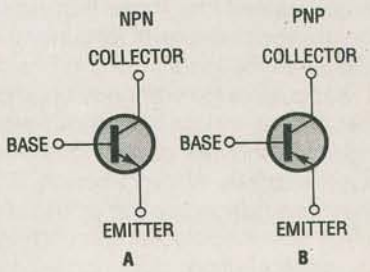

Fig. 3. The schematic symbol for the NPN bipolar transistor is shown in A, while its PNP counterpart is shown in B.

TABLE 1—AMPLIFIER CONDUCTION ANGLES & EFFICIENCY

Class	Angle (Degrees)	Effiency %
A	360	20 – 25
B	180	60 – 78.5
AD	180 – 360	25 – 78.5
C	.<180	>78.5

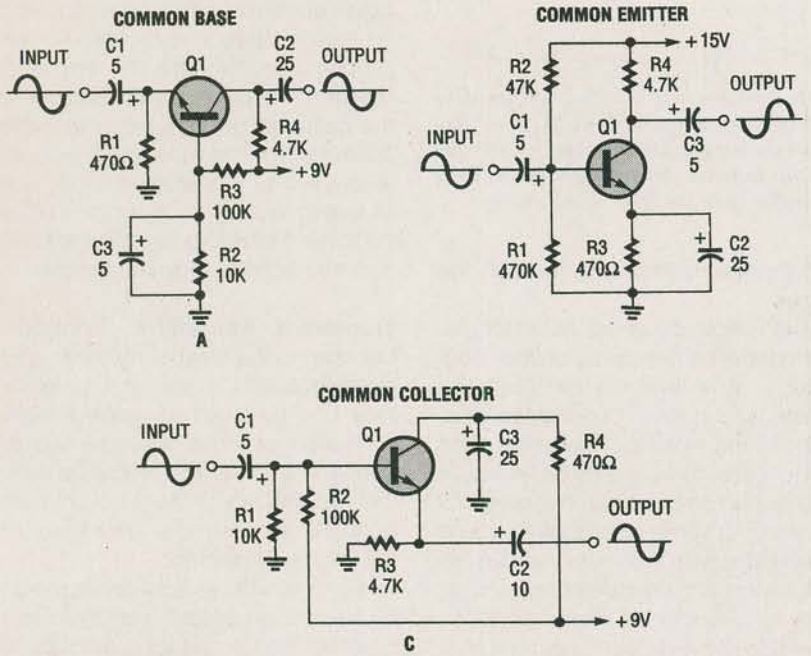

Fig. 4. Examples of the common-base, common-emitter, and common-collector amplifiers are shown in A, B, and C, respectively.

applied to the input of a unity-gain amplifier produces an output signal at the same 10-mV level (a carbon copy of the input signal).

There are many types of amplifiers, however, and all fall into one of two broad categories: voltage amplifiers or current (often referred to as a power) amplifiers. The term voltage amplifier applies to a circuit in which a low voltage is applied to the input to produce a higher voltage at the output. The term power amplifier is generally reserved for those that supply an appreciable power (or current) increase to the load.

Because of the vast array of amplifier circuits in use in modern electronics, amplifier circuits are often subdivided by application—AF, IF, RF, instrumentation, op-amp, etc. Another way of categorizing amplifiers is by configuration: common-emitter, common-collector, and common-base, for example. The important parameters in such circuits are the cutoff frequency and the input/output impedances. The cut-off frequency is the frequency at which the gain of an amplifier falls below 0.707 times the maximum gain of the circuit. The input impedance is the impedance the signal source would see, and the output impedance is the output impedance of the transistor.

Amplifier Configurations. An example of a common-base amplifier is shown in Fig. 4A. The optimum load impedance can range from a few thousand ohms to 100,000 ohms, depending on the circuit's requirements. In this type of circuit, the output signal (at the collector) is in phase with the input signal (applied to the emitter). The current that flows through the base resistance of the transistor is therefore in phase as well, so the circuit tends to be regenerative and will oscillate if the current-amplification factor is greater than one.

A common-emitter (also called a "grounded-emitter") amplifier is shown in Fig. 4B. Base current in this amplifier configuration is small and the input impedance is therefore fairly high (several thousand ohms on the average). Collector resistance on the other hand, can be tens of thousands of ohms, depending on the signal's source impedance. The common-emitter amplifier has a lower cutoff frequency than does the common-base type, but gives the highest power gain of the three configurations. Note that the output signal is 180° out-of-phase with (or the opposite of) the input (base-current) signal, so the feedback that flows through the small emitter resistance is negative (degenerative), keeping the circuit stable. The common-emitter amplifier is one of the most often seen configurations for the bipolar transistor.

The common-collector amplifier (also referred to as an emitter follower), see Fig. 4C, has a high input impedance and a low output impedance. The impedance is approximately:

$$R_s \times (1 - \alpha)$$

The fact that the input resistance is directly related to the load resistance is a disadvantage of this type of amplifier if the load is one whose resistance or impedance varies with frequency. The current transfer ratio of this type of circuit is:

$$1 / (1 - \alpha)$$

and the cutoff frequency is the same as in the common-emitter amplifier circuit. The output and input currents of this type of circuit are in phase.

Amplifier Classifications. Amplifiers may be otherwise classified by their specific operational characteristics, in particular, the bias voltages between the emitter-base and base-collector junctions. The relationship between the bias voltage and the cutoff voltage of an amplifier is what classifies an amplifier as being class A, B, C, or AB. Each class has a specific characteristic that makes it most suitable for a particular application.

In a class-A amplifier—which is the least efficient, but offers the least distortion—the transistor is biased so that its quiescent operating point is in the midddle of the power-supply extremes, *i.e.*, the transistor is always turned on, and the resulting output varies around the bias voltage; see the output waveform in Fig. 5A. Because of that, the input signal must be small enough so that its positive and negative swings do not drive the amplifier near the non-linear cutoff and saturation regions.

(Continued on page 88)

BIPOLAR TRANSISTORS

(Continued from page 48)

Since a high-value resistor is used to change the output voltage to a current (I = V/R) in a class-A configuration, the output current is small. That is important since current flows at all times in such amplifiers, with or without an input signal. Power is wasted and efficiency (the ratio of output to total power consumed) is low—only about 20 to 25%—in class-A amplifiers. Class-A amplifiers can be configured for single-ended or push-pull operation and are used in AF (audio-frequency), IF (intermediate-frequency), and RF (radio-frequency) applications.

transistor is needed. That configuration allows a small input current to control current drawn from a power source, and thus produce a stronger replica of a weaker original signal.

In class-B operation, the transistor is biased at cutoff (see Fig. 5B), so that output current flows during only half of the input cycle. It is used where high efficiency and low distortion are required—for instance, in audio power-output configurations. When the class-B amplifier is used for audio applications, two such amplifiers connected in the push-pull configuration are required, so that current can flow alternately through the two amplifiers. In other words, one amplifier is turned on while the other is turned off.

On the other hand, when the class-

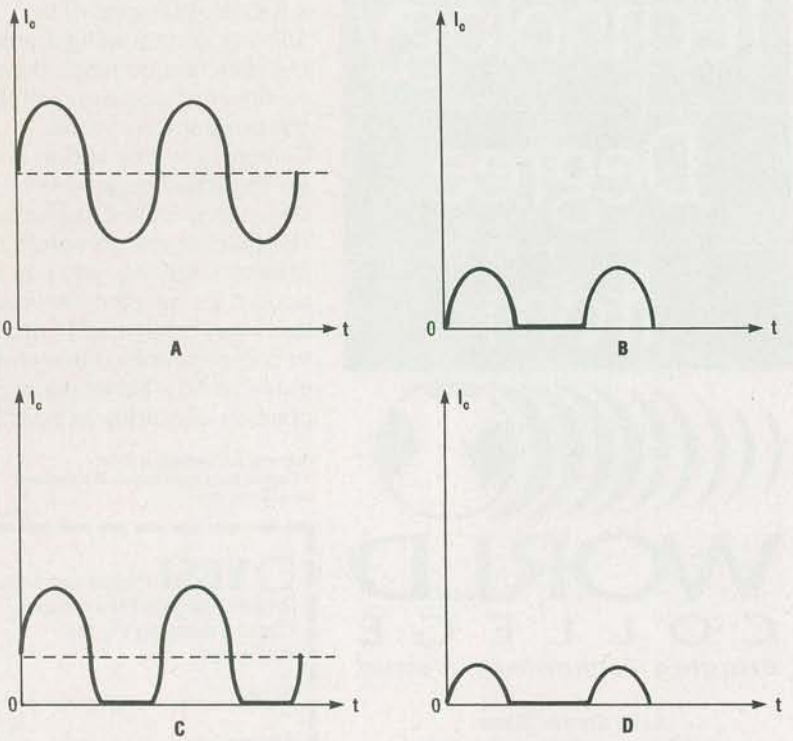

Fig. 5. In a class-A amplifier, the transistor is biased so that it is always turned on (see waveform in A). In class-B operation, the transistor is biased at cutoff (B). Class-AB amplifiers (C) are biased somewhere between class-A and class-B units (D).

Class-A operation is suitable for voltage amplifiers. In a voltage amplifier, the emphasis is on the magnitude of the output voltage. Figure 6 shows a single-ended class-A audio-frequency voltage amplifier. Such an amplifier might be used in a preamplifier stage, where input signals are typically small, and a faithful reproduction of the input using a single B amplifier is used in RF applications, it can be configured for single-ended operation. Since, in the absence of an input signal its current output is negligible, it is used where high efficiency (60–70%) and low distortion are required, which is very important in high-power amplifiers.

Class-AB amplifiers (see Fig. 5C) are biased somewhere between class-A

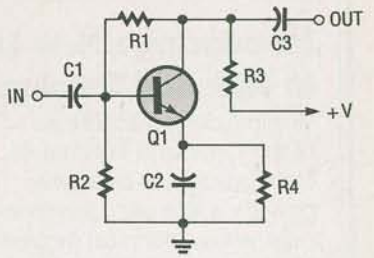

Fig. 6. A single-ended class-A audio-frequency voltage amplifier, like this one, might be used in a preamplifier stage, where input signals are typically small and a faithful reproduction of the input using a single transistor is needed.

and class-B operation, and have efficiencies (25–35%) and distortion characteristics that lie between those of class-A and B amplifiers. Class-AB amplifiers require a somewhat larger input signal than do class-A amplifiers. The class-AB amplifier is used in push-pull configurations for both audio- and radio-frequency applications.

In class-C operation—which has the highest efficiency (perhaps more than 90%), but offers the greatest distortion—the transistor is biased beyond the cutoff region (see Fig. 5D). Because of that, output current flows during less than half (about a third) of the input cycle, making it unsuitable for amplifying signals of varying amplitude, such as audio. That type of amplifier is normally used to amplify a signal of fixed amplitude; for instance, it is often used in the RF power output stages of a transmitter. Current in a class-C amplifier flows in a series of power pulses that excite an LC-tank circuit into oscillation. Because of that the output waveform is a sinewave, that varies in amplitude if modulated. Class C amplifiers can be configured for push-pull or single-ended operation. Table 1 summarizes the conduction angles and efficiency ratings of the various classes of transistor amplifier. ■