4 Amplification

The ability to amplify weak signals was an important milestone in the development of radio technology. This was first achieved around 1912 using a triode thermionic valve and such devices were steadily refined over the next four decades. Around 1947, however, a revolution happened with the development of the transistor and this led to an explosion of semiconductor innovations that eventually resulted in the integrated circuit. This chapter discusses the development of both semiconductor and valve technology. In particular, the development of RF amplification and its impact upon the development of radio. Practical amplifiers suffer from a host of problems that include noise, nonlinearity and parasitic capacitance. The chapter discusses the impact of these factors upon radio performance and, in particular, the ways in which their detrimental effects can be mitigated.

4.1 Thermionic Valves

The first experiments on thermionic valves were essentially carried out by Thomas Edison around 1883 in an attempt to improve the performance of incandescent electric light bulbs. He placed an electrode inside the bulb and found that a current flowed when this was made positive, but no current flowed when it was negative. Edison had created what was essentially a diode. Unfortunately, Edison was not interested in carrying this work forward and it was left to a British physicist, John Ambrose Flemming, to develop Edison's work into something that was much needed in radio, a reliable diode. In the first half of the 1890s, Flemming published several papers on what had become known as the *Edison effect*. We now know that the heating of the filament (the cathode) in the bulb causes electrons to be ejected and these will then be attracted to, or be repelled by, the inserted electrode (the *anode* or plate). In 1904, Flemming patented the first diode valve and the history of radio was changed forever. Figure 4.1a shows the configuration of a typical early diode in conduction. The use of a separate electrode, the *cathode*, to supply the electrons soon became common practice (see Figure 4.1b). The filament now simply heated the cathode to a point where it could release electrons. Such an arrangement meant that the filament supply did not interfere with the electron flow and could be in common with other valves in the circuit. In conduction, the relationship between the current I_P flowing through the valve and the cathode to anode voltage V_P is given by the *Child–Langmuir* law $I_P = \kappa V_P^{3/2}$ where κ is a constant known as the *perveance* of the valve.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

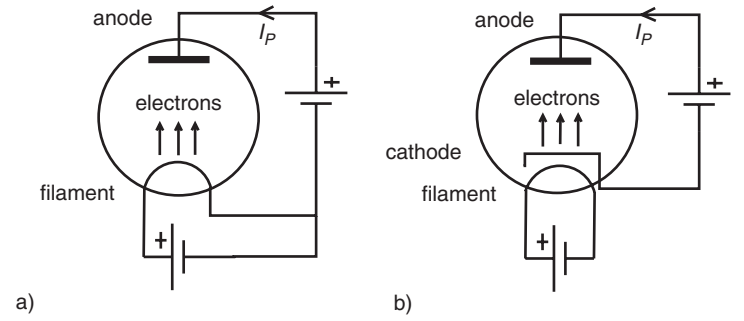


Fig. 4.1 Diode thermionic valve (vacuum tube).

The next major development in the story of thermionic valves (or *vacuum tubes* as they are sometimes known due to the vacuum inside the bulb) was to place an extra electrode (called a *grid*) between the cathode and anode. The grid is a coarse mesh so that it does not impede the flow of electrons from cathode to anode, but is sufficiently close to the cathode for its field to have considerable control over the emission of electrons (a negative grid voltage decreases the emission of electrons, whilst a positive voltage increases the emission). Known as a triode valve, such a device is capable of amplification. A device of this nature was developed by Lee De Forest in 1906 (he called his device an audion), but it took many years before such devices were refined to a point where they could be of much use to radio. By about 1912, however, effective triode valves were available and such valves drove the next stage in the development of radio. Figure 4.2a illustrates the triode (notice we have omitted the filament since it does nothing more than heat the cathode) and 4.2b shows a typical set of characteristic curves. These curves show how the anode current I_P varies with grid voltage V_G and anode voltage V_P. If we insert a load R_L into the anode circuit (see Figure 4.3a), the anode voltage V_P will vary with grid voltage V_G according to the points on the *load line* of Figure 4.3a. For a given grid voltage, the intersection of the corresponding characteristic curve with the load line will provide the anode current I_P and anode voltage V_P. If the grid is negative (the normal mode of operation), the current flow is well approximated by

$$I_P = \kappa (V_P + \mu V_G)^{\frac{3}{2}}, \tag{4.1}$$

where κ is the perveance mentioned above and μ is the *amplification factor*. As a consequence, the voltage drop across the load R_L will be

$$V_L = R_L \kappa (V_P + \mu V_G)^{\frac{3}{2}}. \tag{4.2}$$

The valve will perform voltage amplification, but it will be noted that the amplification can be highly nonlinear. We will return to this issue when we study transistors.

Figure 4.4 is a simple radio that is based on a triode valve and employs what is commonly known as a grid leak detector. The tuned circuit is coupled to the grid through the capacitor C_2 and the grid is biased to the same voltage as the cathode through a very high impedance resistor R. Consequently, the grid will only conduct on positive voltage swings at the output of the tuned circuit. During such periods, the *grid leak capacitor* C_2

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

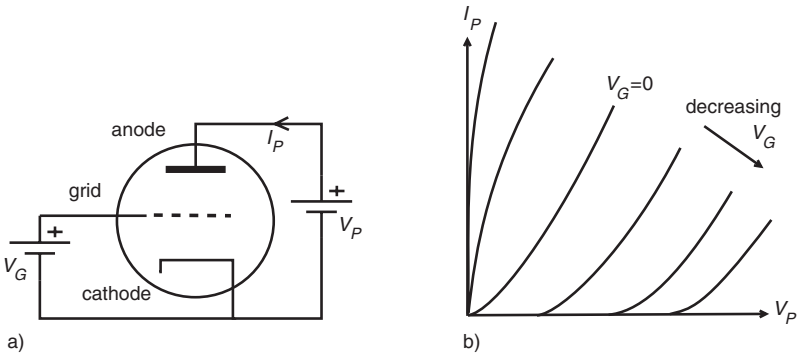


Fig. 4.2 Triode thermionic valve (vacuum tube).

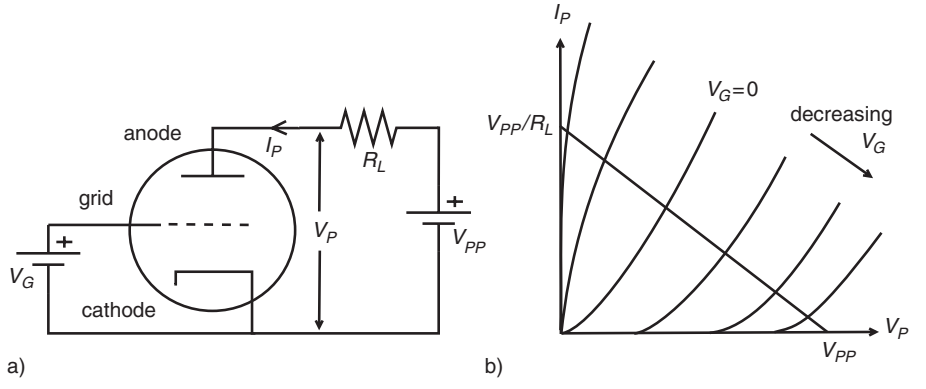


Fig. 4.3 Triode thermionic valve with load.

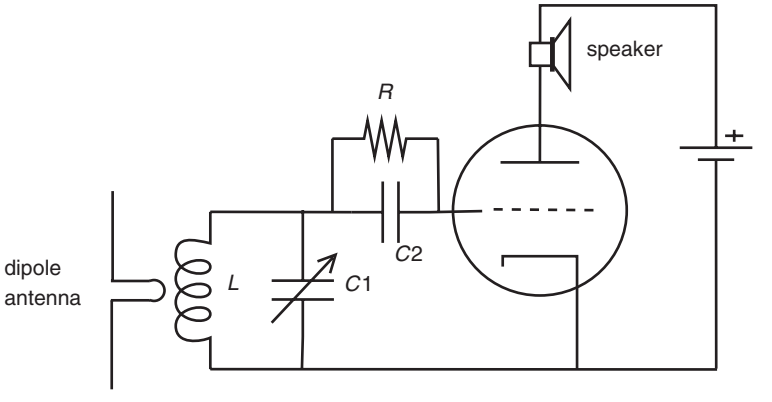


Fig. 4.4 Radio based on a grid leak detector.

will charge up and then maintain a negative bias on the grid during negative swings. The capacitor $C2$ needs to be chosen large enough to smooth out the RF component at the grid, but not so large that it smooths out the modulation. The voltage at the grid will therefore follow the amplitude modulation of the signal and this will be amplified through the valve.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

4.2 Semiconductor Diodes

The valve diode was a reliable replacement for the point contact diodes developed by Bose and others, but its heavy power requirement was a severe drawback. However, in 1940, Russell Ohl discovered the semiconductor pn junction, and this led to the development of highly reliable and effective diodes, without the heavy power consumption of valves. To understand the pn junction we first need to understand what is meant by a semiconductor. A crystal form of many materials results when the atoms are bound together by sharing the electrons in their outermost electron shell (this is often known as a covalent bond). Figure 4.5 depicts a crystal solid consisting of atoms with four electrons in their outermost shell. If the solid consisted of carbon atoms, the outermost shell would be an L shell with four electrons (this solid is in fact diamond). Furthermore, once it has shared its electrons with its neighbours, its L shell would be full. As a consequence, there is no option for electrons to move to higher energy levels (they would need to bridge the large energy gap between the L and M shells) and hence gain the kinetic energy of motion, i.e. the material is an insulator. If the crystal consists of silicon atoms, the outermost shell is an M shell containing four electrons. In this case the outer shell is not full and so electrons can more easily enter higher energy levels for which the additional energy is now the kinetic energy of motion. At room temperature, several of the electrons of silicon will have achieved these higher states and will be available for conduction when an electric field is applied. Such a material is known as a *semiconductor*. Another example of a semiconductor material is germanium, an atom for which the outermost shell is an N shell containing four electrons.

The conductivity of semiconductors can be improved by the addition of some impurity into the solid. Consider adding an impurity consisting of phosphorous into a crystal of silicon. The outermost shell of a phosphorous atom is an M shell containing five electrons. When some phosphorous atoms are added to the silicon crystal (i.e. it is *doped*), four of the five phosphorous electrons will be shared with neighbouring silicon atoms (see Figure 4.6). The fifth electron, however, is not bound by the sharing mechanism and can easily move into the available energy levels of the M shell, empty in the case of silicon. Consequently, very little electric field will be required to move an electron through the

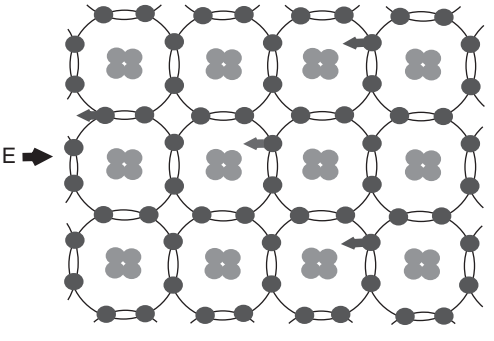


Fig. 4.5 In a semiconductor crystal, some electrons can attain energy levels that allow them to move.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

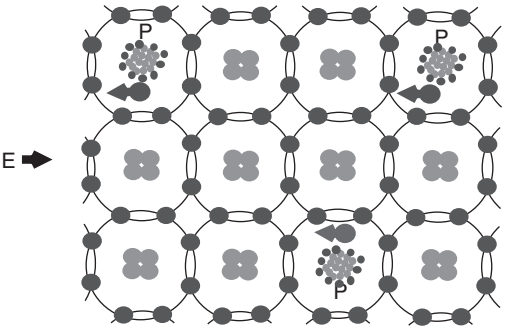


Fig. 4.6 Flow of charge through an n-type semiconductor by movement of electrons.

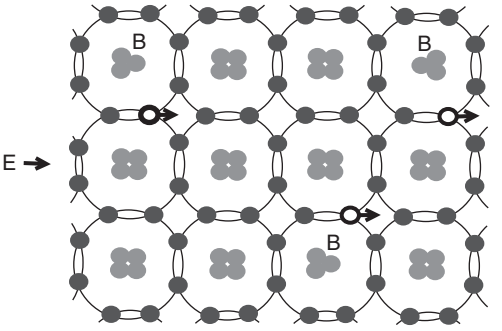


Fig. 4.7 The flow of charge through a p-type semiconductor by movement of holes.

crystal. Now consider adding an impurity consisting of boron to a crystal of silicon. The outermost shell of a boron atom is an L shell containing three electrons. When some boron atoms are added to the silicon crystal, the three electrons will be shared with neighbouring silicon atoms (see Figure 4.7). There will, however, be an electron hole in the L shell of one of the neighbouring silicon atoms. Very little electric field will be required to move an electron of a neighbouring atom into the hole, hence creating a hole in the neighbouring atom. This hole will then be filled by an electron from another neighbouring atom and so on through the crystal. In this way, there will be an effective flow of positive charge through the crystal. In deference to the type of charge carrier, the phosphorous-doped material is known as an n-type semiconductor and the boron-doped material as a p-type semiconductor.

It is possible to grow a semiconductor crystal in which one half is p-type and the other half is n-type. When there is no field across the junction, there is still some charge flow by the process of *diffusion*. Under this process, thermal agitation will cause carriers to flow from a region of strong concentration to a region of low concentration. Electrons will flow from the n-type semiconductor to the p-type where they fill holes and holes will flow from the p-type semiconductor to the n-type. This produces an accumulation of positive charge in the n-type region and negative charge in the p-type region. As a result there is a *depletion region* either side of the junction that is devoid of carriers. The accumulation

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

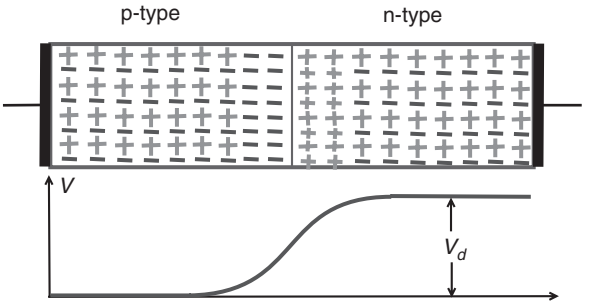


Fig. 4.8 A p-n junction at equilibrium.

of charge on either side of the junction will lead to an electric field that will increasingly oppose the migration of further carriers across the junction until the diffusion is reduced to a trickle. It should be noted, however, that thermal agitation will always produce some minority carriers (holes in the n-type semiconductor and electrons in the p-type) and these will be swept across the junction by this field and form what is known as the *drift current*. Eventually, the drift current and residual diffusion current will balance each other and we will have a state of equilibrium. In this state, the potential difference across the junction will vary as shown in Figure 4.8. The equilibrium potential difference V_d across the junction is known as the diffusion potential and has a value around 0.3 volts for germanium and around 0.8 volts for silicon.

If a voltage is now applied across the diode with the n-type end positive, known as *reverse bias*, it will further impede the diffusion and enhance the drift current. As a consequence, only a small amount of current will flow from n-type to p-type semiconductor (see Figure 4.9b). However, if a voltage is now applied with the p-type end positive, known as *forward bias*, the voltage drop across the depletion layer will be reduced and the diffusion current significantly increased (see Figure 4.9c). The current I through the diode is related to the potential drop across the device through the relation

$$I = I_s \left(\exp\left(\frac{eV}{\eta kT} \right) - 1 \right), \tag{4.3}$$

where I_s is the reverse saturation current, e is the charge on an electron, V is the voltage across the device, k is Boltzmann's constant, T is the absolute temperature and η is a constant that can have a value between 1 and 2.

The original semiconductor diodes were of the point-contact variety. In such devices, an n-type semiconductor is fused to a metal contact (often aluminium) to form the cathode and then the anode is formed by contact with a metal wire (usually tungsten) having a very fine point. Some of the metal ions from the point then migrate into the semiconductor and form a small p-type region. As mentioned above, an alternative process is to grow a semiconductor crystal with the desired regions of semiconductor type. The crystal is steadily pulled out of a semiconductor melt to which is added, at appropriate stages, suitable amounts of impurity in order to attain the desired pn structure.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

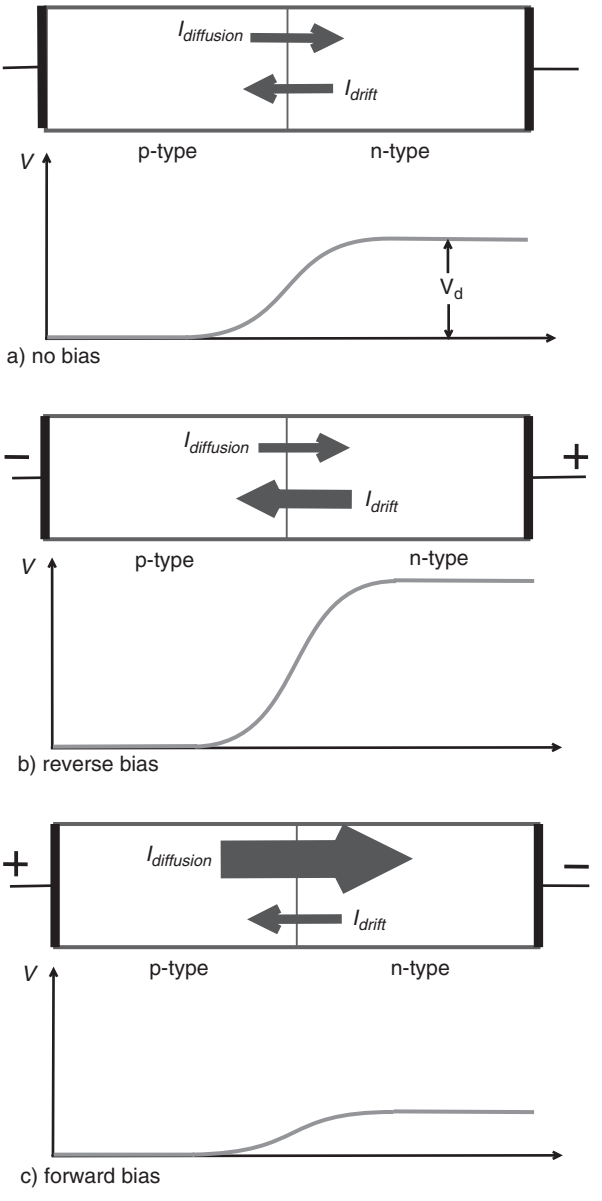


Fig. 4.9 Voltage drop and currents across a pn junction.

4.3 The Bipolar Junction Transistor

The next big step forward was made in 1947 with the invention of the *bipolar junction transistor* (or BJT for short) by the American physicists John Bardeen, Walter Brattain and William Shockley of AT&T's Bell Laboratories. They took the crucial step of adding an additional p-type layer resulting in a germanium device with two junctions. In fact, in the first device invented by Bardeen and Brattain, the p-type regions were created using

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

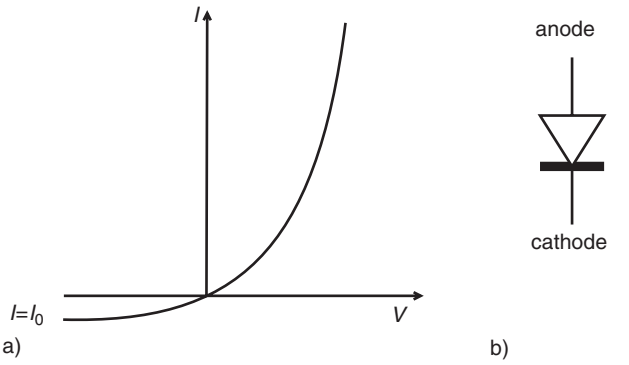


Fig. 4.10 Semiconductor diode characteristic and diode symbol.

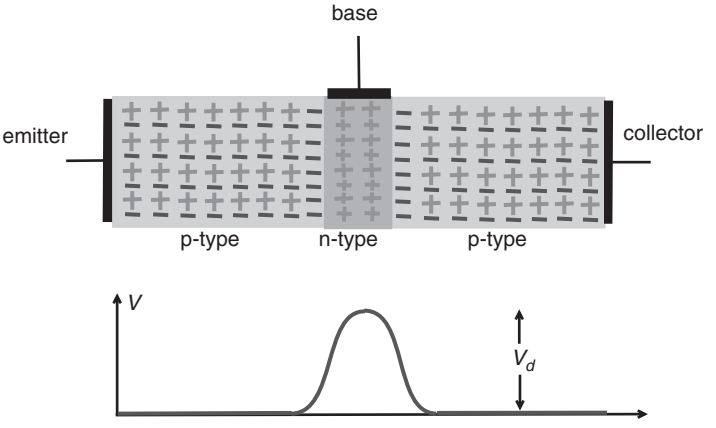


Fig. 4.11 The bipolar junction transistor in an unbiased state.

point contacts with an n-type semiconductor. Shockley, however, was able to improve on this by growing a crystal with an appropriate pnp sandwich. Figure 4.11 shows a pnp transistor geometry in its unbiased state and from which it will be noted that the depletion of carriers from the base area has caused a potential barrier to any further diffusion. The transistor looks like two back-to-back pn junctions, but this analogy is far too simplistic. Crucial to the transistor's operation is the very close proximity of the two p-type regions (not a property of two diodes when connected back-to-back) and a very much lower density of impurity in the n-type region. We now consider what happens when the transistor base and collector are biased by negative voltages with respect to the emitter (see Figure 4.12). It is assumed that the bias voltages are sufficient to make the base forward-biased with respect to the emitter and the collector reverse-biased with respect to the base. Positive carriers will now flood into the base, but will find few electrons to combine with due to the small density of impurity. However, they will feel a strong pull from the collector due to the reverse bias and will readily flow into this region. They will do this in preference to flowing out through the base, due to the smaller distance to be covered. Consequently, a current I_C will flow into the collector which is

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

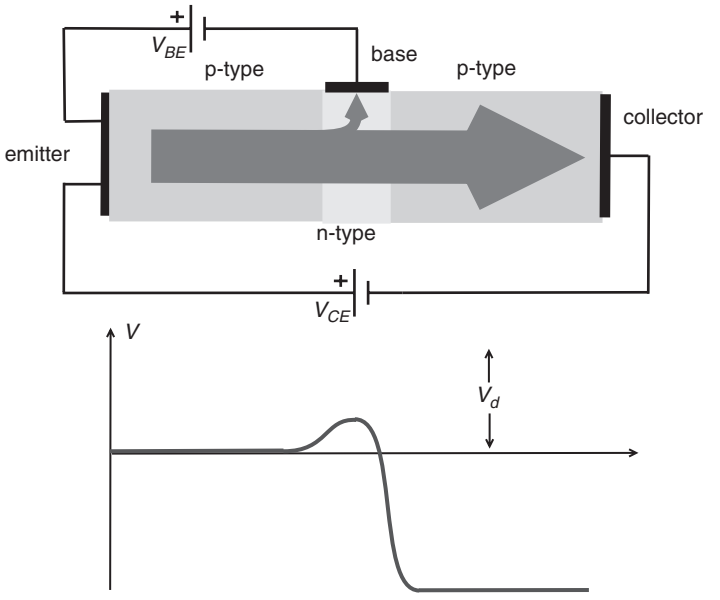


Fig. 4.12 A pnp bipolar junction transistor in a biased state.

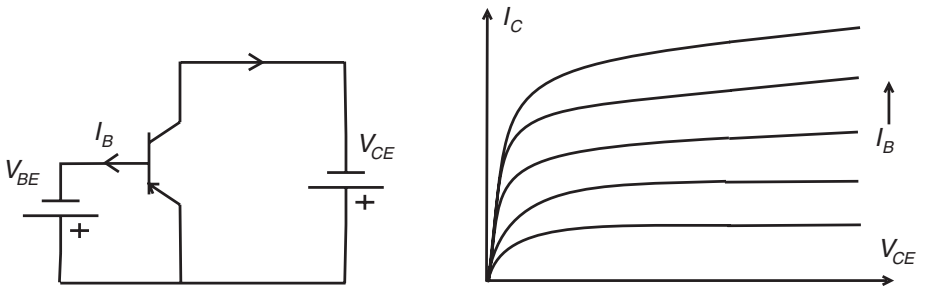
β times the current I_B that flows out through the base, i.e. $I_C = \beta I_B$. The quantity β is known as the *current gain* and can have values of the order of 100. Essentially, the BJT is a current amplifying device. It should be noted, however, that the input current is related to the voltage that is imposed between the emitter and the base. The base–emitter junction is essentially a forward biased diode and so the current that flows out of the base is related to the base–emitter voltage V_{BE} through $I_{BE} = I_s\left(\exp\left(\frac{eV_{BE}}{\eta kT}\right) - 1\right)$. The current I_s is very small and so we will have that

$$I_C \approx \beta I_s \exp\left(\frac{V_{BE}}{V_T}\right), \tag{4.4}$$

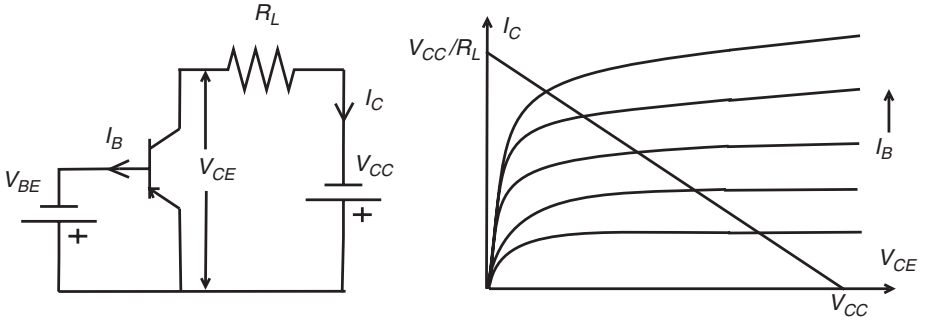
where $V_T = \frac{kT}{e}$ (approximately 25 mV) and we have taken η to be 1.

It is clear that there needs to be a reasonable bias at the collector in order to maintain a reverse bias on the base–collector junction. Figure 4.13a shows characteristic curves (the relationship between collector current and collector voltage) for a variety of base currents. These show that the characteristics become highly nonlinear for low collector voltages. In reality, the transistor will be used to drive a load and this will affect the voltage at the collector of the transistor. Figure 4.13b illustrates the situation through what is known as the *load line*. For a given base current, the intersection of the corresponding characteristic curve and the load line will provide the collector–emitter voltage and the collector current. As the base voltage varies it is clear that the voltage across the load will vary in sympathy and at a much greater amplitude. This property is put to use in the simple receiver circuit of Figure 4.13c. The input signal is demodulated by the diode consisting of the emitter to base junction and then the demodulated signal is amplified in the transistor (the capacitor $C2$ filters out any residual RF frequencies).

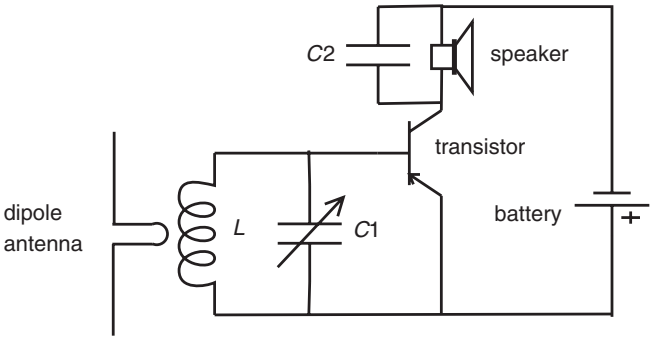
https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press



a) Circuit description of a biased pnp BJT and a typical characteristic.



b) The transistor driving a load.



c) A simple radio receiver using a pnp transistor as both detector and audio amplifier.

Fig. 4.13 Application of pnp bipolar junction transistor to amplification.

The problem with the amplifier of circuit of Figure 4.13b is that it is highly nonlinear. From (4.4), the relationship between the input voltage V_{BE} and the voltage drop V_L across the load is given by

$$V_L \approx R_L \beta I_s \exp\left(\frac{V_{BE}}{V_T}\right). \tag{4.5}$$

As a consequence, a sinusoidal base voltage V_{BE} would result in a voltage drop V_L with the lower half of the sinusoid missing and the remaining half heavily distorted. In the above receiver, this is put to good use in demodulating the incoming signal. On the

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

whole, however, we would like to amplify without signal distortion. Even without a V_{BE} that goes negative, the output can still become distorted. As the output current rises, the value of V_{CE} can fall and there is the possibility that the lower half of the sinusoid will stray into the highly nonlinear region on the left of the characteristic. In general, an amplifier needs to be designed so that the output voltage does not stray into these strongly nonlinear regions. The solution is to create a DC bias at the transistor base so that the input voltage does not stray into negative territory and to choose the load so that the V_{CE} does not stray into the nonlinear region of small V_{CE}. For appropriate choices of bias and load impedance, it is possible to get variations in voltage drop across the load that are very nearly a constant multiple of quite large variations in the base voltage (see Figure 4.14). Obviously, to choose a suitable bias point, one would need to study the manufacturer's characterisation of the transistor.

The pnp geometry is not the only combination of semiconductors that produces a transistor device and an alternative is the npn combination (see Figure 4.15). The behaviour of the npn device is much the same as a pnp device, except that we must change the polarity of the biases. Figure 4.16a shows an amplifier circuit that uses an npn transistor. It will be noted that we have derived the base bias from the collector supply V_{CC} with a suitable voltage divider (this is a common arrangement). It will be further noted that there is also an emitter resistor that has a bypass capacitor C_{BP} (the value of this capacitor is sufficiently high for it to behave as a short circuit at signal frequencies). This resistor might seem unnecessary, but it provides feedback that stabilises the operation of the transistor. The DC bias voltage V_{BE} is chosen such that the *quiescent current* I_Q through the transistor (the current that flows when there is no signal) places the variations in output voltage v_o well away from the regions of nonlinearity. Once the appropriate quiescent collector current I_Q has been ascertained, the emitter resistor R_E needs to be chosen to give an emitter voltage V_E that is the maximum consistent with the operation of the transistor (it must allow a sufficient swing in the output voltage). The voltage divider (R_1 and R_2) needs to be chosen to give a voltage $V_{BE} + V_E$ at the base and, in

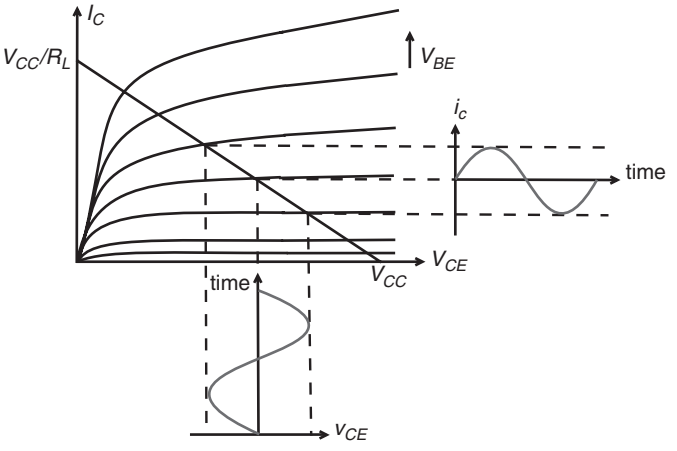


Fig. 4.14 Input/output relationship for a biased amplifier.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

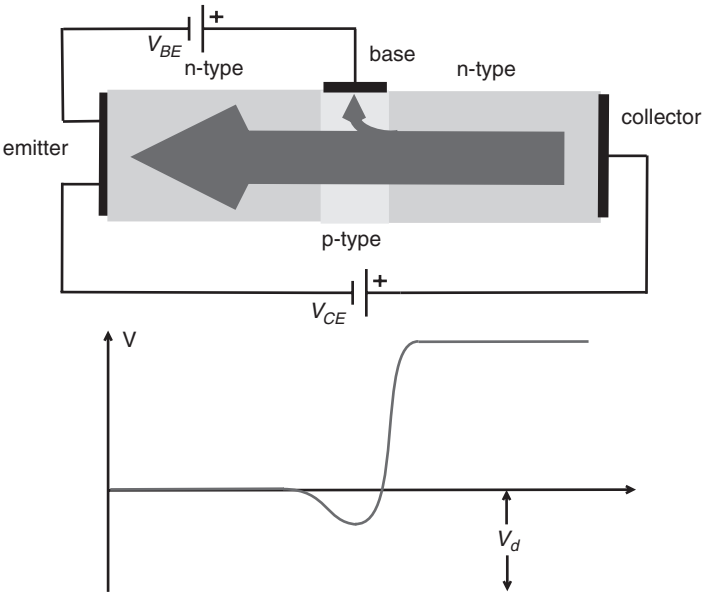


Fig. 4.15 A npn bipolar junction transistor in a biased state.

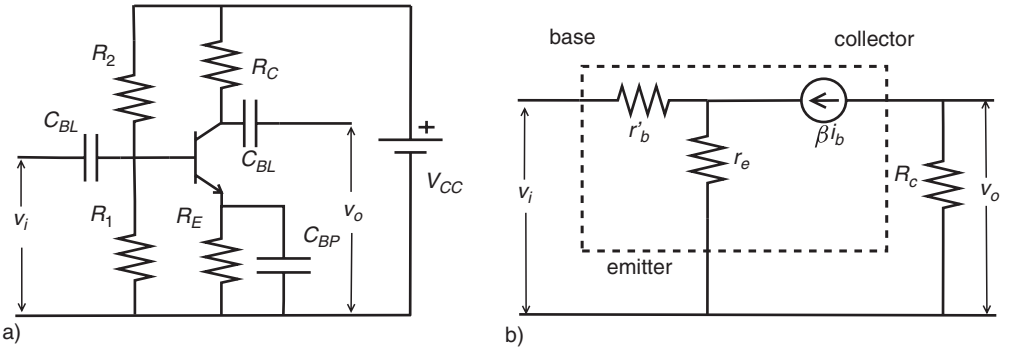


Fig. 4.16 A npn bipolar junction transistor in a biased state and a simple small-signal model.

most circumstance, V_{BE} is sufficiently approximated by the diffusion potential V_d (0.3 volts for germanium transistors and 0.8 volts for silicon). Divider resistances need to be chosen large enough that they do not overload the source (i.e. have a combined resistance well above that of the source), but small enough that they themselves are not overloaded by the transistor.

We now turn to the reason for the emitter resistance R_E. From (4.5) it can be seen that the current gain β, and the temperature T, have a big impact on the quiescent current once V_{BE} has been set. Obviously, temperature will vary considerably, even under normal conditions. Furthermore, there can be a considerable spread of β in manufactured devices. If current rises due to a change in temperature, the voltage across R_E will rise. As a consequence, the voltage V_{BE} will fall and, along with it, the

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

current. Consequently, R_E provides feedback that stabilises a device against fluctuations in temperature. Further, this feedback also helps to stabilise against fluctuations in β. Larger values of R_E provide better feedback, but the value will need to be moderated by the other requirements of the design.

Once we have fixed the DC components of current through the bias circuits, we can ignore this aspect and deal with the RF signal alone, assumed to be a small fluctuation about the quiescent state. For such an analysis, the bypass C_{BP} and blocking capacitors C_{BL} are replaced by short circuits. (The purpose of the blocking capacitors is to prevent the bias of one stage of a more complex circuit affecting that of another stage and their values need to be chosen to be large enough to behave as a short circuit at signal frequencies.) Further, the DC supply is treated as a short circuit (note that we normally place a bypass capacitor across the supply in order to ensure this). The circuit in Figure 4.16a will now reduce to that shown in Figure 4.16b. Note that the transistor has been replaced by a linear model (the circuit within the broken rectangle) and this is valid for *small signals*. In this model, the transistor is effectively a current source with the current proportional to the base current. Resistance r'_b is known as the *base spreading resistance* and r_e as the *emitter resistance*. From the circuit of Figure 4.16b, we find that

$$v_i = i_b r'_b + (\beta + 1) i_b r_e \tag{4.6}$$

and

$$v_o = -\beta i_b R_C, \tag{4.7}$$

where we have used lower case letters to denote RF voltages and currents. Eliminating i_b between (4.6) and (4.7), we obtain that

$$v_o = -R_C \frac{\beta}{r'_b + (\beta + 1) r_e} v_i = -R_C g_m v_i, \tag{4.8}$$

where g_m is known as the *transconductance* of the amplifier. Under normal circumstances, r'_b is very small and can be ignored. On the other hand, β is usually large and so $g_m \approx 1/r_e$.

An alternative form of small-signal model is shown in Figure 4.17. To apply the model, however, we need values for the transconductance g_m, the input resistance r_π and the output resistance r_0. The transconductance is given by $g_m = \mathrm{d}I_C/\mathrm{d}V_{BE}$ when

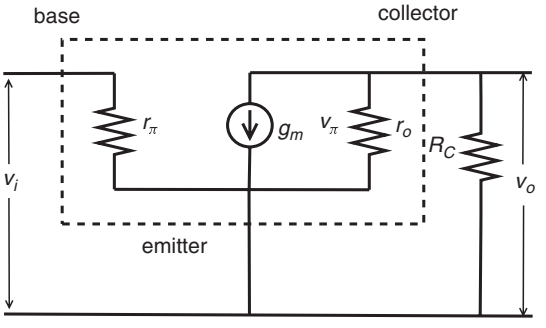


Fig. 4.17 An improved small-signal model.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

evaluated at the quiescent current I_Q (i.e. $g_m = I_Q/V_T$) and the input resistance by $r_\pi = r_b' + (\beta + 1)/g_m$ from (4.6). Output resistance r_o is due to the small slope of the characteristic curves at large V_{CE} and can normally be ignored (it has a value of the order of 100 Kiloohms). Referring to Figure 4.16, we have that $v_o = -g_m(R_C \parallel r_0)v_i$, i.e. the voltage gain of the amplifier is $-g_m(R_C \parallel r_0)$. Further, to a load, the amplifier looks like a voltage source with impedance $R_C \parallel r_o$.

4.4 The Field-Effect Transistor

The *field-effect transistor* (FET) is a transistor device with very different physics from that of a BJT. In a BJT the conduction is interrupted by a layer of different-type semiconductor that controls the current through the device by means of a current through a contact that is connected to this intermediate layer. In the FET, however, the flow through a continuous semiconductor is controlled by an electric field that is applied through a contact known as a gate. The gate current is negligible and so the device is voltage-controlled rather than current-controlled. Such transistors can be traced back to 1925 when Julius Lilienfeld took out a patent on such a device. Efforts to create a working transistor were frustrated by the available technology and it was only with developments that came through the BJT that a successful device could be constructed.

There are two major varieties of the device: the junction FET and the insulated gate FET. A junction FET (or JFET) consists of semiconductor channel (the ends of which are known as the source and drain), but with a diode junction on the side of the channel (known as the gate). The diode is reverse-biased and so there is a depletion region whose size is then controlled by the voltage between the source and gate. In order to enhance the size of the depletion region within the channel, the channel is much more lightly doped than the gate region. The size of this depletion region will then control the current that flows through the device (see Figure 4.18a). A typical characteristic behaviour of

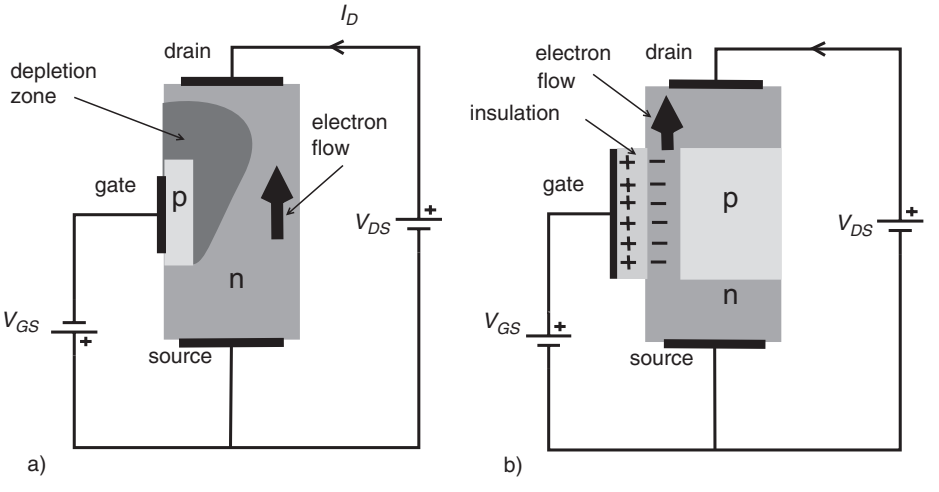


Fig. 4.18 N channel JFET and IGFET.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

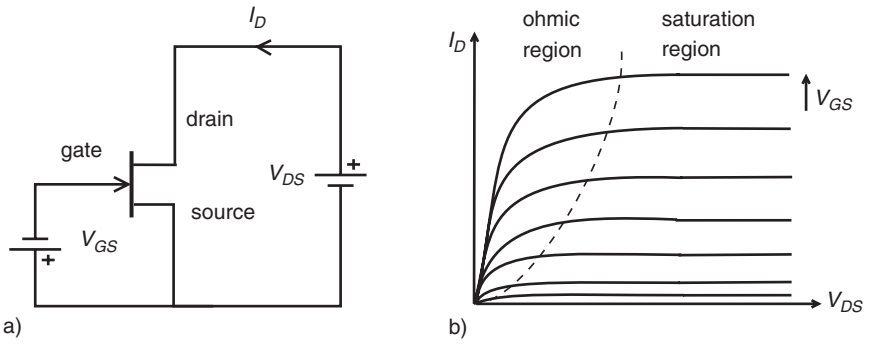


Fig. 4.19 JFET circuit and characteristic.

an n-channel JFET is shown in Figure 4.19, along with a circuit description. Consider a fixed V_{GS} (source to gate voltage), for small V_{DS} (source to drain voltage) the channel will initially behave as a simple resistor with V_{GS} controlling the channel resistance. As the drain voltage V_{DS} increases, the current I_D will steadily increase. However, as V_{DS} gets larger, the depletion region will also grow towards the drain and start to constrict the flow. Eventually, a point is reached where an increase in flow through the channel resistor is countered by an increase in channel resistance due to the channel constriction. At this point, we have reached *saturation* and the current flow remains constant for any increase in drain voltage. The channel is now effectively pinched-off and the drain voltage at which this happens is known as the *pinch-off voltage*. For its operation, the device requires the gate channel junction to be reverse-biased. There will, however, be a threshold V_T below which there will be no flow through the device due to the depletion region covering the total width of the device. For operation in the saturated region, the drain current is related to the gate voltage through the relation

$$I_D = K(V_{GS} - V_T)^2, \tag{4.9}$$

where K is a constant that depends on the device construction.

As with the BJT, we will need to keep the operating range well away from the nonlinear regions of the transistor characteristic if we are going to achieve linear amplification. We can do this by choosing a suitable quiescent drain current I_{DQ} through a biasing network of the sort shown in Figure 4.20a. It will be noted that, since the input impedance of a FET is very high (of the order of megaohms), the resistors of the voltage divider (R_1 and R_2) can also be very large. As with the BJT, manufactured devices can exhibit a great range of material parameters, in particular V_T, and the source resistor R_S provides the feedback that guards against these variations. If ΔV_T is the variation expected from the devices, and ΔI_D is the allowable variation in drain current, we will need to choose R_S greater than $\Delta V_T / \Delta I_D$.

As with the BJT, once we have fixed the quiescent component of current through the bias circuits, we can ignore this aspect and deal with the signal fluctuations alone. To do this, we will need a transistor model of the sort shown Figure 4.20b, valid for small signals. Like the BJT, the FET behaves like a current source. However, the input

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

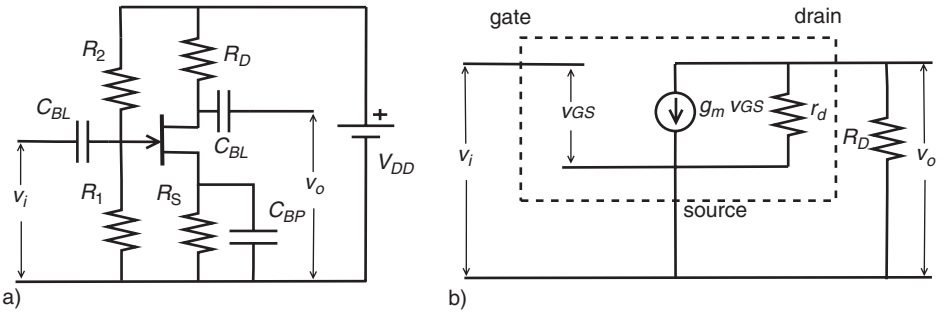


Fig. 4.20 A JFET common-source amplifier and JFET small-signal model.

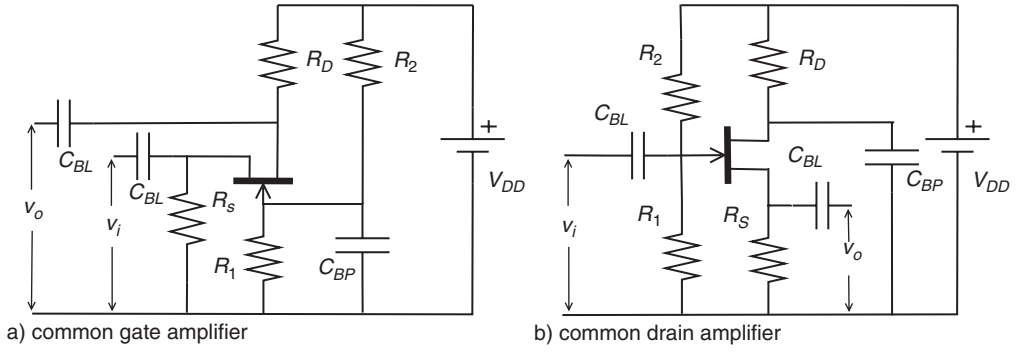


Fig. 4.21 Other varieties of FET amplifier.

resistance is now so high that we can ignore it. As with the output resistance of the BJT, the value of r_d is fairly large and can be ignored to a first approximation. The transconductance is given by $g_m = \mathrm{d}I_D/\mathrm{d}V_{GS}$ and is evaluated at the quiescent current I_{DQ}. Consequently, from 4.9, we obtain that $g_m = 2\sqrt{KI_{DQ}}$. Referring to Figure 4.20, we have that $v_o = -g_m(R_D \parallel r_d)v_i$, i.e. the voltage gain of the amplifier is $-g_m(R_D \parallel r_d)$. Further, to a load, the amplifier looks like a voltage source with impedance $R_D \parallel r_d$.

Up to now we have concentrated on what is commonly known as the common-source amplifier, but for small-signal amplifiers there are other options. Figure 4.21 shows two other varieties of amplifier (common-gate and common-drain amplifiers), both with very different properties from the common-source amplifier. The output of the common-gate amplifier behaves very much like the output of the common-source amplifier and is a voltage source $v_o = -g_m R_D v_i$ with impedance $R_D \parallel r_d$. Looking into the input, things are a little different and the relatively high-input impedance of the common-source amplifier gives way to a relatively low input impedance of $R_S \parallel g_m^{-1}$. For the common-drain amplifier, the input looks like that of the common-source amplifier, but everything changes at the output. The amplifier looks like a voltage source v_i (i.e. a voltage gain of 1) with an impedance of $R_S \parallel g_m^{-1}$. The common-gate and common-drain amplifiers have their counterparts in the BJT world, i.e. the common-base and common-collector

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

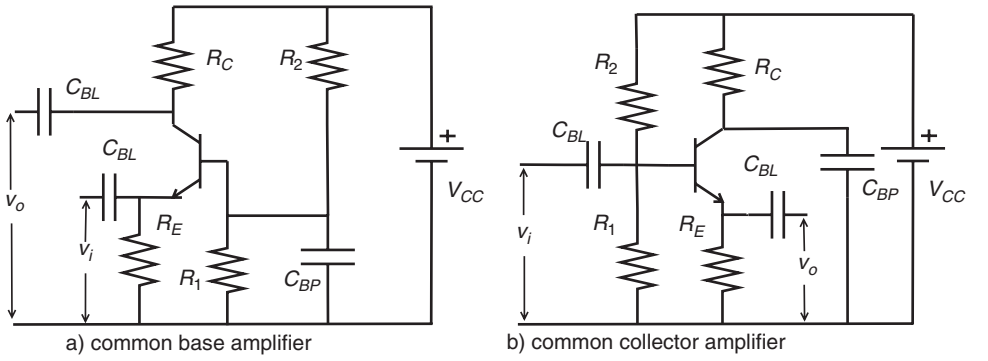


Fig. 4.22 Other varieties of BJT amplifier.

amplifiers (see Figure 4.22). Further, given the similarity of the BJT small-signal model to that of the FET, they will possess similar properties.

An insulated-gate FET (or IGFET) is a semiconductor bar (assumed to be p-type for the present) in which two junctions have been formed (the source and the drain) by infusing n-type semiconductor. The junctions are joined by an n-type layer, so thin that almost no current flows between the junctions (see Figure 4.18b). Placed over this thin region is a metal electrode that is separated from it by an insulating layer. If a positive voltage is now applied to this electrode, it behaves like a capacitor and positive charge accumulates. This charge is then counterbalanced by electrons that are drawn into the thin layer. This enhances the current-carrying capacity of the thin layer by effectively widening it.

The characteristics for the IGFET are similar to those of the JFET and can be explained as follows. We consider the capacitor that is formed between the gate electrode and the thin n-type layer. The average voltage on the channel side will be $V_{DS}/2$ and so we have $Q = C(V_{GS} - V_{DS}/2)$ negative charge formed on the channel side (C is the capacity of the gate channel capacitor). However, current does not start flowing until all the remaining holes in the thin channel are filled up, i.e. CV_T of the above negative charge is not available as carriers where V_T is the gate voltage at which charge starts to flow. We have charge $Q_c = C(V_{GS} - V_T - V_{DS}/2)$ that is available as carriers. Most of the voltage drop across the device will occur in the thin region and so the electric field to which the charge is subjected will be $E_c = V_{DS}/L$ where L is the length of the thin channel. The electron field will cause the available charge to accelerate, but this will be moderated by collisions and the net effect will be a drift velocity $v_d = \mu_n E_c$, where μ_n is a quantity known as the electron mobility. As a consequence there will be a current $\mu_n Q_c E_C/L$, i.e

$$I_D = 2K(V_{GS} - V_T - V_{DS}/2)V_{DS}, \tag{4.10}$$

where K is a constant that depends on the device construction. In reality, the voltage in the gate will not be uniform across the gate, but increase linearly from source to drain. This will mean that the channel width will decrease as we move from the source to drain. Further, as V_{DS} increases, there will eventually be a point where the channel is

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

pinched-off at the drain end. At this point there will be no further increase in current with V_{DS}. At the point where we have pinch-off we will have $V_{DS} = V_{GS} - V_T$. Consequently,

$$I_D = K(V_{GS} - V_T)^2 \tag{4.11}$$

and this remains the same throughout the saturation region. Due to the ever-decreasing size of devices, mainly driven by the ever-increasing complexity of integrated circuits (ICs), the channel length L has continued to be reduced and this has resulted in ever-larger values of channel field E_c. However, at a certain level of the field, the carriers will attain enough energy to set the semiconductor lattice into a quantum mechanical mode of vibration and hence lose energy to this mechanism. In other words, the velocity of the carriers will saturate (at a level of about 10^5 m/s for silicon). The saturation velocity v_{sat} will correspond to a field level $E_{\text{sat}} = v_{\text{sat}}/\mu_n$. If this saturation occurs before pinch-off, it will result in a modification to (4.11) of the form

$$I_D = \frac{K(V_{GS} - V_T)^2}{1 + \frac{V_{GS} - V_T}{LE_{\text{sat}}}}. \tag{4.12}$$

In Section 4.7, we will find that such nonlinearity in transistor behaviour can have far-reaching consequences for radio receiving systems.

The gate of an IGFET is frequently made by using a thin layer of oxide to insulate the gate electrode from the semiconductor and, for this reason, these devices are often known as MOS (metal oxide silicon) transistors. The above n-type channel FET is then known as an NMOS transistor. A PMOS transistor is an alternative that has a p-type channel and an n-type substrate. Typical common-source amplifiers for these two transistors are shown in Figure 4.23 and from which it will be noted that the PMOS drain needs to have a negative bias (in addition to a negative bias on the gate). The different bias requirements of the various types of FETs are summarised in Figure 4.24 which shows the characteristics of these devices in saturation. MOS FETs were invented by Dawon Kahng and Martin Atalla in 1959 and have now largely superseded the JFET due to their ease of construction.

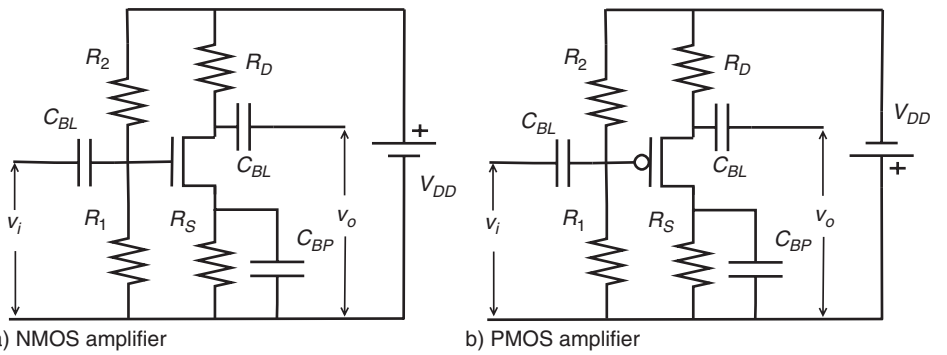


Fig. 4.23 NMOS and PMOS amplifiers.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

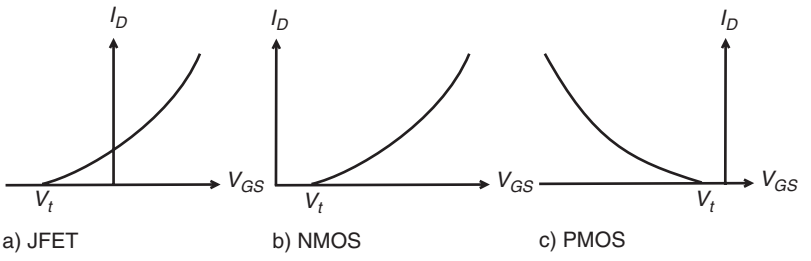


Fig. 4.24 Characteristic curves for various FETs in saturation.

4.5 Radio Frequency Amplifiers

At high frequencies, a major problem arises due to the internal capacitances of a transistor. There is obviously capacitance across the junctions of a BJT and a JFET. Furthermore, CMOS depends on the capacitance between gate and channel. As a consequence, at RF frequencies, we need a model of transistor behaviour of the sort shown in Figure 4.25a (this is appropriate for a FET, but a similar model will also apply to the BJT). The problem with this model is that the analysis of such circuits is difficult due to the capacitive feedback that is inherent in the transistor model. It turns out that the capacitances C_{GS} and C_{DS} are the least problematic and so we ignore them for the moment. Fortunately, we can further simplify feedback using a result that is known as *Miller's theorem* (Miller, 1920). We first note that the source in the FET model can be changed from a current source to an equivalent voltage source as shown in Figure 4.25b. Consequently, we consider the amplifier shown in the first circuit of Figure 4.26. It turns out that the feedback impedance Z can be replaced by impedances Z_1 and Z_2 at the input and output of the transistor (see Figure 4.26), providing they are suitably chosen. From the Kirchhoff voltage law we have,

$$V_o = AV_i + I_f r_o \text{ and } V_i = V_o + I_f Z \tag{4.13}$$

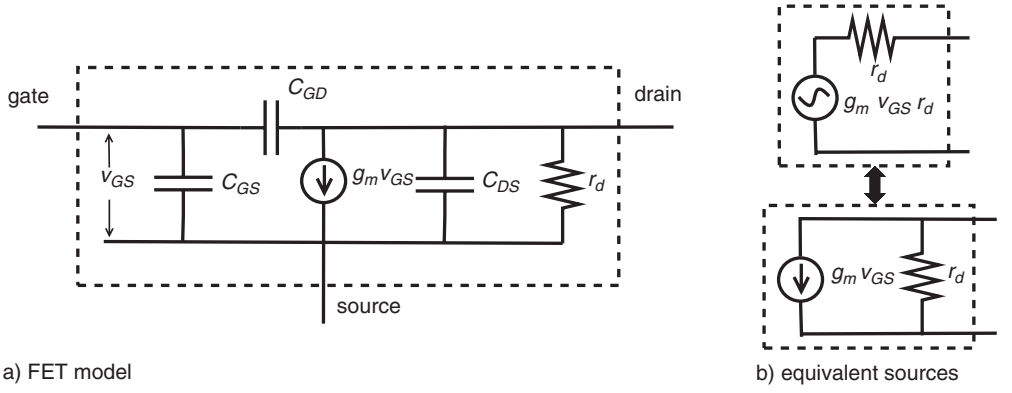


Fig. 4.25 High-frequency model of a FET.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

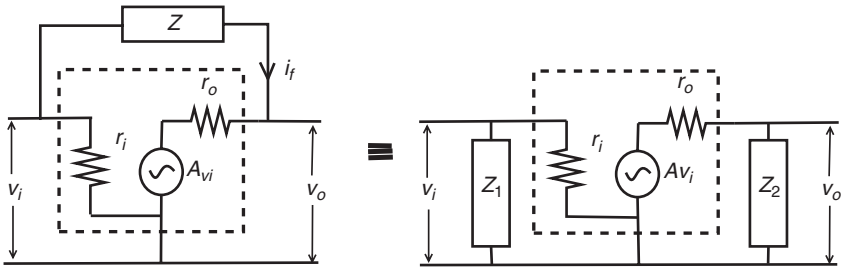


Fig. 4.26 Miller's result.

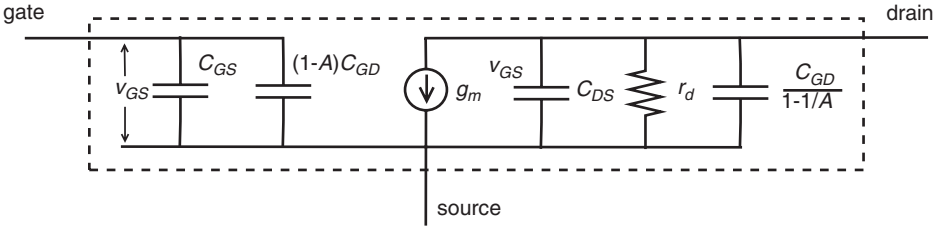


Fig. 4.27 Alternative high-frequency model of a FET based on Miller's theorem.

where $v_o = \Re\{V_o \exp(j\omega t)\}$, $v_i = \Re\{V_i \exp(j\omega t)\}$ and $i_f = \Re\{I_f \exp(j\omega t)\}$ (i.e. we assume time-harmonic signals). From Figure 4.25b, we need Z_1 to draw the same current as the feedback circuit, i.e. $I_f = V_i/Z_1$. Further, from (4.13), we have that $I_f/V_i = (1-A)/(r_o + Z)$ and so

$$Z_1 = \frac{r_o + Z}{1 - A}. \tag{4.14}$$

Similar arguments at the output also imply that

$$Z_2 = \frac{AZ + r_o}{A - 1}. \tag{4.15}$$

For an ideal amplifier ($r_o = 0$ and $r_i = \infty$), we will have $Z_1 = Z/(1 - A)$ and $Z_2 = Z/(1 - 1/A)$. For the model of Figure 4.25, the feedback is the capacitance C_{GD} and this, and the other capacitances, are usually only a few picofarad. As a consequence, the additional impedance at the output is of the same order as existing impedance. At the input, however, there is a different story. Assuming we can treat the amplifier as ideal, the additional impedance will be capacitative (we assume that we have a common-source or common-emitter amplifier) with the value of the capacitor $(1-A)C_{GD}$ (see Figure 4.27). For an amplifier with reasonable gain we have a large shunt capacitance at the input, hence negating the gain of the amplifier by grounding the input at high frequencies. This is known as the *Miller effect* and we need to find strategies for countering it.

One possible approach is to cancel out the input capacitance with a parallel tuned circuit. Figure 4.28 shows the circuit of a simple receiver in which this approach has been used. In this receiver there is a second amplifier which also acts as a detector. The operating point of the second amplifier is set to be just above that at which the

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

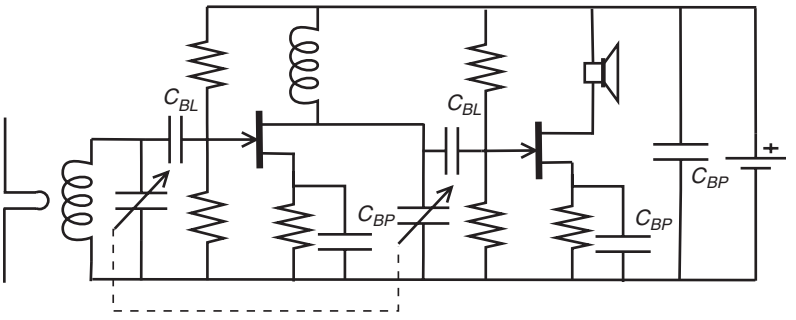


Fig. 4.28 Tuned radio frequency (TRF) receiver.

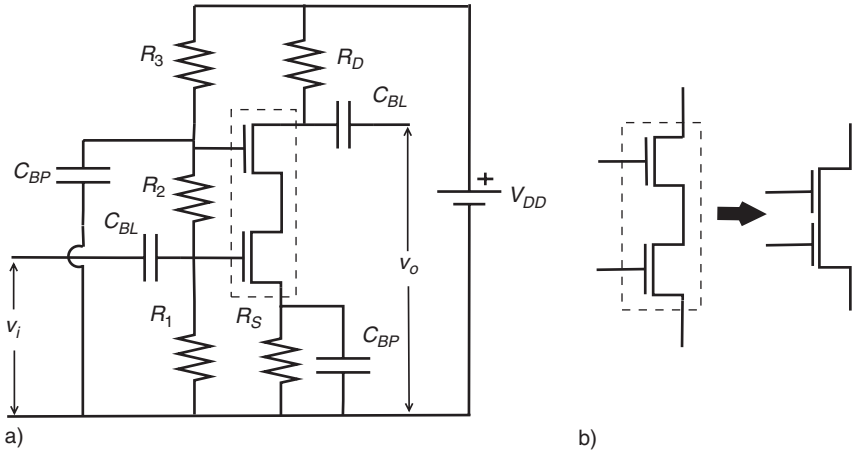


Fig. 4.29 Cascode amplifier and dual gate FET.

transistor starts to conduct and so the transistor will only conduct on the positive part of the RF cycle. Consequently, the average current at the drain will consist of the amplitude modulation of the signal. The important thing to be noted is that, besides overcoming the Miller effect, the additional tuned circuits will increase the selectivity of the receiver, i.e. they will better reject unwanted signals on nearby frequencies. Later we will see that the ability to reject nearby signals is crucial to the performance of a radio receiver. The receiver shown in Figure 4.28 is known as a tuned radio frequency (TRF) receiver and was popular (in its valve form) during the decade after the first world war. The drawback, however, is that all the tuned circuits need to be adjusted together (the dotted line in the Figure indicates a 'ganged' tuning capacitor) if the frequency of reception is to be changed.

The Miller effect can be overcome without the use of tuned circuits through the cascode amplifier (see Figure 4.29a). This circuit is essentially a common-source amplifier that is followed by a common-gate amplifier (a common-emitter amplifier followed by a common-base amplifier is used in the case of BJTs). The input impedance of the

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

common-gate amplifier will be g_m^{-1} and so this will be the load of the common-source amplifier, i.e. the gain of the common-source amplifier will be 1. As a consequence, the common-source amplifier will have no gain and hence no Miller effect. On the other hand, the common-gate amplifier will have a very low input impedance and so the Miller effect will have to be extreme in order to have an effect. Furthermore, the total gain of the amplifier will be that of the common-gate amplifier. By combining two transistors, we now have an amplifier with the same behaviour as the common-source amplifier, but without the Miller effect. Since the drain of the bottom FET and the source of the upper FET are directly connected, the two transistors can be combined into a single device known as a dual-gate FET. (In the case of valve amplifiers, a similar trick to avoid the Miller effect consists of adding an additional grid to form a *tetrode valve*.)

A further example of amplifier topology is the differential pair, as shown in Figure 4.30a. Consider the RF components of current and voltage. There is a constant-bias current I_B that is set by V_{bias} and so $i_D^a + i_D^b = 0$. Further, $v_o^a = -R_D i_D^a$ and $v_o^b = -R_D i_D^b$ and, from the small-signal model of a FET, $i_D^a = g_m v_{GS}^a$ and $i_D^b = g_m v_{GS}^b$. Bringing these relations together, we find that

$$v_o^b - v_o^a = -g_m R_D (v_i^b - v_i^a), \tag{4.16}$$

i.e. the difference in output voltages is a multiple of the difference in input voltages.

Interestingly, the same topology can be used to produce another amplifier that can overcome the Miller effect. Consider the amplifier shown in Figure 4.30b. In this case the transistor loads are constructed from transistors (the load on the left has resistance g_m^{-1} and the load on the right has resistance r_d). This is a useful approach in integrated circuit technology where transistors are far easier to fabricate than resistors. The input transistor has unity gain and is thus immune from the Miller effect. In the case of the output transistor, the base voltage is fixed and so the Miller effect is irrelevant. As with the cascode amplifier, this alternative amplifier will have a gain of $-g_m R_D$.

The amplifier configuration of Figure 4.30a can also be used to make a two-input, one-output, amplifier that is commonly known as an *operational amplifier*. If a buffer

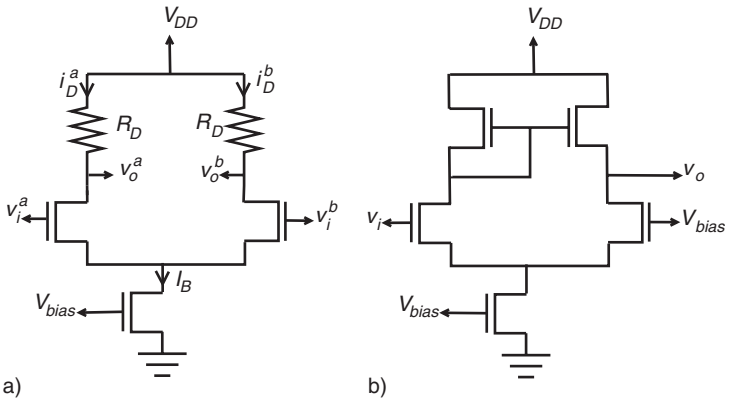


Fig. 4.30 An CMOS differential amplifier and alternate with active loads.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

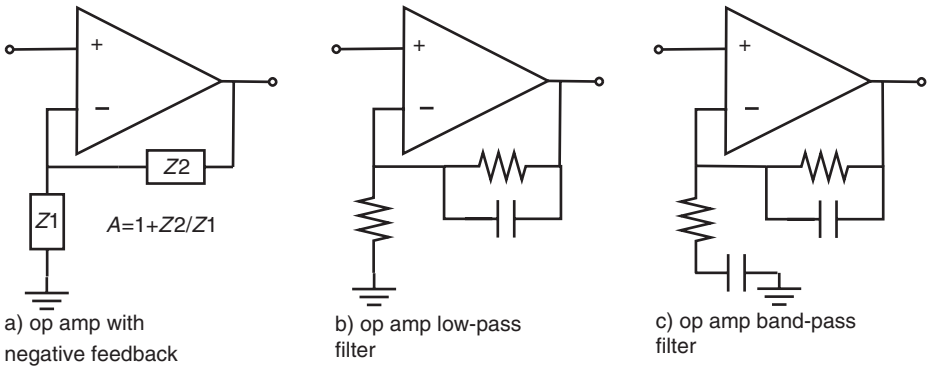


Fig. 4.31 General configuration of an op amp together with low- and band-pass filters.

amplifier is added at the output, we have a high-gain amplifier with high-input impedances and low output impedances. Importantly, for one of the inputs (the one normally labelled $-$), the output is phase-shifted by 180°. Figure 4.31a shows an operational amplifier with a general negative-feedback configuration (the voltage gain of the amplifier is given by $A = 1 + Z2/Z1$) and Figures 4.31b and 4.31c show the feedback configured so that the amplifier acts as low-pass and band-pass filters respectively.

4.6 Noise

Besides the desired signal, there will always be *noise*, i.e. competing unwanted signals. The electronics of a radio can itself be the source of considerable noise, usually known as *internal noise* in order to distinguish it from that which comes in through the antenna. Even the humble resistor will create noise due to the thermal motion of its electrons (*Johnson noise*). In reality, a resistor should be modelled as an ideal resistance in series with a random noise source (see Figure 4.32a) of mean square voltage

$$\overline{v_n^2} = 4kTBR, \tag{4.17}$$

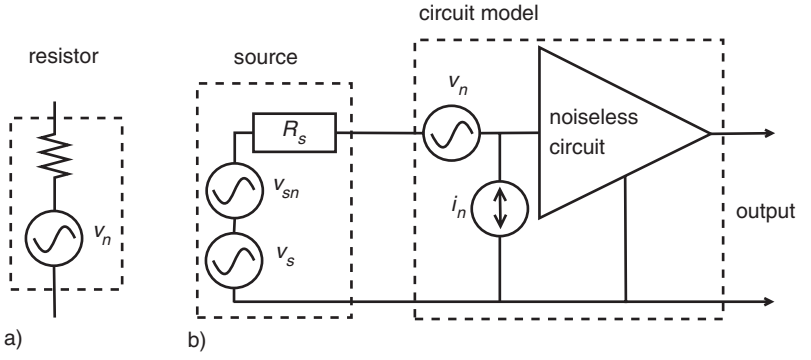


Fig. 4.32 Noise models (a resistor and a general circuit).

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

where T (in kelvins) is the absolute temperature, B (in hertz) is the bandwidth of the receiver, R (in ohms) is the resistance and k is the Boltzmann constant (1.38×10^{-23} joules per kelvin). (A general impedance Z will behave as the noise source of its resistive part in series with the ideal impedance Z.) Semiconductors are also the source of a considerable variety of noises (popcorn noise and flicker noise to name but two). In particular, *shot noise* occurs at semiconductor junctions due to the discrete nature of the carriers and their independence. (For this noise, the mean-square current is given by $\overline{i_n^2} = 2eI_0B$ where I_0 is the quiescent current.) As a consequence, complex electronic circuits can be quite difficult to analyse from a noise perspective. Fortunately, it can be shown that a complex circuit can be modelled as a noise-free circuit with a series voltage noise source at its input and current noise source across the input (see Figure 4.32b). In the case of a BJT amplifier, the main sources of noise are the shot noises in the junctions and the noise in the base resistance r_b'. At radio frequencies, these noises can be approximated by

$$\overline{v_n^2} = 4kTB\left(r_b' + \frac{r_e}{2}\right) \text{ and } \overline{i_n^2} = 2eI_bB\left(1 + \beta\frac{\omega^2}{\omega_T^2}\right), \tag{4.18}$$

where ω_T is the cut-off frequency of the transistor (the frequency at which the internal capacitances of a BJT have reduced its current gain to unity). In the case of a FET amplifier, the dominant sources of noise are due to the channel resistance and shot noise at the gate in the case of a JFET. At radio frequencies, the noise can be approximated by

$$\overline{v_n^2} = \frac{8kTB}{3g_m} \text{ and } \overline{i_n^2} \approx 2eI_gB + \frac{4}{9}\frac{8kTB}{3g_m}\omega^2C_{GS}^2, \tag{4.19}$$

where I_g is the gate current (zero in the case of CMOS). It will be noted that, for a BJT, the collector current is a key element in deciding the level of noise and so one might ask whether there is an optimum quiescent collector current I_C from the viewpoint of noise. The available equivalent input noise power for a common-source BJT amplifier will be

$$N_i = 4kTB + \frac{4kTB}{R_S}\left(r_b' + \frac{V_T}{2I_C}\right) + 2e\frac{I_C}{\beta}BR_S, \tag{4.20}$$

where R_S is the source resistance (we assume the frequency is well below ω_T). For N_I to be minimum, we require that $\partial N_i/\partial I_C = 0$ and from this we find that $I_C = \sqrt{\beta}V_T/R_S$. This will usually imply a small quiescent current for minimum noise and this can be at odds with other requirements of the amplifier, linearity requirements in particular. (The reader should consult van der Ziel (1986) for a more detailed discussion of noise in semiconductor devices.)

In a realistic radio system, *external noise* (that which comes through the antenna) is just as important as the internal variety. This can arise from man-made sources (computers and ignition interference for example), natural extraterrestrial sources (galactic noise) and natural terrestrial sources (lightning for example). Lightning as a source of noise is complex since, for frequencies below 30 MHz, the noise at a single point can be the accumulation of the effect of lightning strikes across the globe. This arises due to propagation via the ionospheric duct. If an antenna has an effective resistance R_A, the noise coming in through the antenna is often described in terms of *antenna temperature*

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

T_A, i.e. the temperature that a resistor R_A would need to have in order to produce the same noise as that coming in through the antenna. External noise is the ultimate constraint on a receiver system and it is usual to design a radio receiver to be *externally noise limited* (i.e. the internal noise is below the level of external noise).

The signal-to-noise ratio (SNR) is the crucial parameter in quantifying the effect of noise upon a receiver system and is defined by

$$\text{SNR} = \frac{\text{signal power}}{\text{noise power}}. \tag{4.21}$$

How a stage of a receiver affects the SNR is usually measured in terms of its *noise factor* F (known as the *noise figure* when expressed in dB terms). This is defined by

$$\text{F} = \frac{\text{available output noise power}}{\text{available output noise power due to the source alone}}. \tag{4.22}$$

The noise factor needs to be referenced to a well-defined noise source, usually a resistor at ambient temperature T (defined to be 290 K for convenience). In this case, the noise can be represented as a noise source of power $FkTB$ at the input of a noise-free circuit. If the circuit has a noise source at a temperature T_S other than ambient (an antenna for example), the total noise power at the input to the noise-free circuit will then be

$$F = kT_SB + (F-1)kTB. \tag{4.23}$$

If we apply definition (4.22) to the model of noise shown in Figure 4.32b, we obtain that

$$\text{F} = 1 + \frac{\overline{v_n^2} + \overline{i_n^2}R_s^2}{4kTBR_s}, \tag{4.24}$$

where R_s is the source resistance (note that we have assumed there to be no correlation between the equivalent voltage- and current-noise sources). Obviously, we would like to minimise the noise factor. Since the source impedance is under our control, we could ask the question as to what source impedance would provide the lowest noise. We need to find the R_s for which $\partial F/\partial R_s = 0$ and this turns out to be when $R_s^2 = \overline{v_n^2}/\overline{i_n^2}$. Unfortunately, this is not necessarily the source impedance that would give a power match. Fortuitously, however, it is the value that gives the best SNR. To see this consider the situation shown in Figure 4.33 in which the source is connected to the input of the circuit through a transformer with turns ratio N. Noting that impedance is scaled by N^2 through the transformer and voltage by N, we find that

$$\text{SNR} = \frac{\frac{v_s^2}{R_s}}{\frac{\overline{v_{sn}^2}}{R_s} + \frac{\overline{v_n^2}}{N^2R_s} + \overline{i_n^2}R_sN^2}. \tag{4.25}$$

Minimising SNR with respect to N (i.e. we solve for $\partial\text{SNR}/\partial N = 0$), we find that $R_sN^2 = \sqrt{\overline{v_n^2}/\overline{i_n^2}}$, i.e the transformer needs to change the source impedance to $\sqrt{\overline{v_n^2}/\overline{i_n^2}}$.

As a signal passes through the circuits of the receiver, the SNR will degrade through contributions from various noise sources within the circuits. Further, in many circumstances, the received signal will be very weak and therefore need to pass through several stages of amplification. Consequently, we need to know how the noise factor will

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

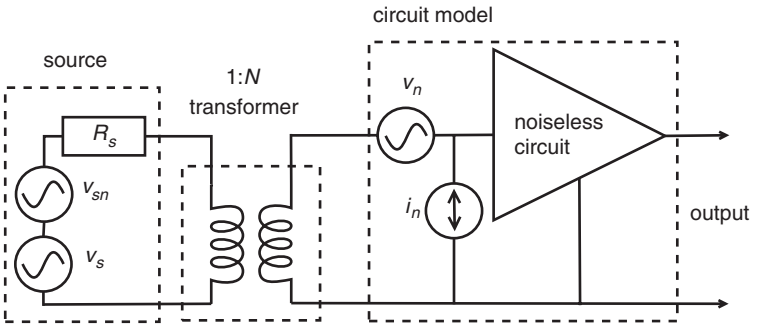


Fig. 4.33 Noise model with source-matching circuit.

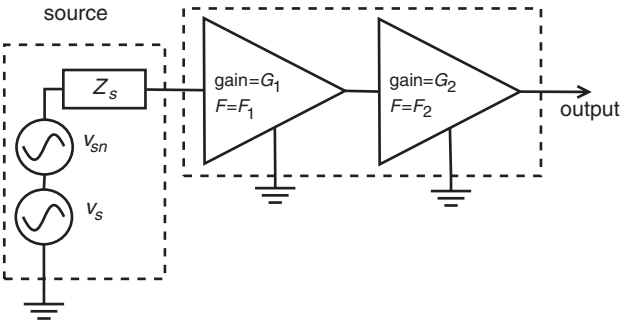


Fig. 4.34 Noise figure of cascaded circuits.

change through a cascade of circuits. Firstly, if a circuit is driven by a source at ambient temperature T, we have from the definition of the noise factor F that the noise power added by the circuit will be an equivalent noise power $(F-1)kTB$ at the circuit input. Now consider the cascaded circuits shown in Figure 4.34. The noise at the output of the first circuit will be $G_1kTB + G_1(F_1-1)kTB$. In the case of the second circuit, the noise power added will be $(F_2-1)kTB$ at the input. Consequently, the total noise at the output of the second circuit will be $G_1G_2kTB + G_1G_2(F_1-1)kTB + G_2(F_2-1)kTB$. The noise due to the source alone will be G_1G_2kTB and so, from the definition of the noise factor,

$$\begin{aligned} \mathrm{F} &= \frac{G_1G_2kTB + G_1G_2(F_1-1)kTB + G_2(F_2-1)kTB}{G_1G_2kTB} \\ &= F_1 + \frac{F_2-1}{G_1}, \end{aligned} \tag{4.26}$$

where F_1 and F_2 are the noise factors of the separate amplifiers and G_1 and G_2 are their power gains. An important deduction from this result is that the first amplifier in a radio receiver always needs to be the one with the best noise figure.

An important factor in the characterisation of a radio receiver is its sensitivity. This is dictated by the level of noise with which the incoming signal must compete and is often described in terms of the *noise floor* N_f. For a receiver with noise factor F, and bandwidth

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

B, the noise floor is total equivalent noise at the input of a noiseless receiver, i.e.

$$N_f = F = kT_A B + (F-1)kTB, \tag{4.27}$$

where T_A is the antenna temperature and T is usually assumed to be the ambient noise temperature, i.e. 290 K. Another measure of sensitivity the *minimum detectable signal* (MDS) which is defined to be the level of signal that is equal to the noise (i.e. it is the same number as the noise floor). The sensitivity of a receiver can sometimes be quoted as a voltage $v_f = \sqrt{N_f R_I}$ where R_I is the input resistance of the receiver.

The relation between SNR and detection was quantified by Claude Shannon of Bell Laboratories in 1948. Shannon showed that C, the maximum capacity of a communication channel in bits per second, was related to the SNR through (Shannon, 1947)

$$C = B\log_2(1+\text{SNR}), \tag{4.28}$$

where B is the channel bandwidth in Hertz and $\log_2$ indicates a logarithm to the base 2 ($\log_2 x = 1.4427 \ln x$ in terms of natural logarithms). Although stated in terms of the language of digital communication (bits per second), this is a very general result that sets an ultimate limit on the rate of communications for a given bandwidth and SNR.

4.7 The Effect of Nonlinearity

As we have seen, in the case of both FET and BJT amplifiers, the relationship between input and output voltages is by no means linear. This made it necessary to bias circuits to operate within the most linear part of the circuit characteristic. Unfortunately, it is not possible to remove all nonlinearity and so we need to investigate the consequences of the nonlinearity for circuit performance. We will assume that the output voltage v_o is related to the input voltage v_i through the relationship

$$v_o = k_o + k_1 v_i + k_2 v_i^2 + k_3 v_i^3 + \cdots \tag{4.29}$$

If we assume a sinusoidal input voltage of the form

$$v_i = V\cos\omega t \tag{4.30}$$

the output voltage will take the form

$$v_o = (k_o + \frac{k_2}{2}V^2) + (k_1 + \frac{3k_3}{4}V^2)V\cos\omega t + \frac{k_2}{2}V^2\cos 2\omega t + \frac{k_3}{4}V^3\cos 3\omega t + \cdots \tag{4.31}$$

after the application of some standard trigonometric identities. The output contains harmonics (signals at multiples of the input frequency), a constant offset voltage and an amplification term with voltage gain $k_1 + \frac{3k_3}{4}V^2$. The harmonics can be removed with suitable filtering and so pose no particular problem. The amplification term, however, is problematic due to the dependence of the gain upon the level of input voltage. The value of k_3 is usually negative and this will lead to a reduction in gain with increasing input voltage, i.e. we have *gain compression*. Gain compression is an important consequence

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

of nonlinearity and is normally measured in terms of the 1 dB compression point P_{1dB}, i.e. the input power level at which the amplifier gain is reduced by 1 dB.

Whilst gain compression is a problem for power amplification, even weak signals can suffer from degradation due to nonlinearity. Consider an input signal v_i that is a combination of two sinusoidal signals

$$v_i = V_1 \cos\omega_1 t + V_2 \cos\omega_2 t. \tag{4.32}$$

We will ignore the effects of coefficients higher than k_3 in (4.29), then

$$\begin{aligned} v_o(t) = k_0 &+ \frac{k_2}{2}(V_1^2 + V_2^2) \\ &+ (k_1 V_1 + \frac{3}{4}k_3 V_1^3 + \frac{3}{2}k_3 V_1 V_2^2)\cos(\omega_1 t) \\ &+ (k_1 V_2 + \frac{3}{4}k_3 V_2^3 + \frac{3}{2}k_3 V_1^2 V_2)\cos(\omega_2 t) \\ &+ \frac{3k_3}{4}V_1 V_2^2 \cos((2\omega_2 - \omega_1)t) + \frac{3k_3}{4}V_1^2 V_2 \cos((2\omega_1 - \omega_2)t) \\ &+ \frac{k_2}{2}V_1^2 \cos(2\omega_1 t) + \frac{k_2}{2}V_2^2 \cos(2\omega_2 t) \\ &+ \frac{k_3}{4}V_1^3 \cos(3\omega_1 t) + \frac{k_3}{4}V_2^3 \cos(3\omega_2 t) \\ &+ k_2 V_1 V_2 \cos((\omega_1 + \omega_2)t) + k_2 V_1 V_2 \cos((\omega_1 - \omega_2)t) \\ &+ \frac{3k_3}{4}V_1 V_2^2 \cos((2\omega_2 + \omega_1)t) + \frac{3k_3}{4}V_1^2 V_2 \cos((2\omega_1 + \omega_2)t) \end{aligned} \tag{4.33}$$

after the application of some standard trigonometric identities. We will consider frequencies ω_1 and ω_2 that are within the pass band of the amplifier. The last four lines of the above expression, and the zero frequency terms of the first line, represent products that can be easily removed by filtering. The terms on lines two and three, however, represent the desired amplified frequencies. Besides gain compression, the main thing to note is that a strong undesired signal at frequency ω_2 could cause a reduced response at a desired frequency ω_1 when k_3 is negative. Such *desensitisation* can be quite a problem for weak signals in a strong signal environment. The terms of line four represent another problem that is caused by a non-zero k_3. If the signals at frequencies ω_1 and ω_2 are strong and undesired, the third-order inter-modulation will produce components at frequencies $|2\omega_1 - \omega_2|$ and $|2\omega_2 - \omega_1|$. In a crowded radio environment, it is possible that such components could be coincident with a desired frequency and so be a source of interference.

A measure of this effect is the *inter-modulation distortion* (IMD) which is the ratio of the output power at the frequency $|2\omega_1 - \omega_2|$ to the output power at the fundamental frequency ω_1,

$$\text{IMD} = \left(\frac{3k_3 V_1 V_2}{4k_1}\right)^2. \tag{4.34}$$

A frequently used measure of third-order effects is the *third-order intercept point IIP*3 (see Figure 4.35). If the two unwanted signals have equal amplitude ($V_1 = V_2$), this is

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

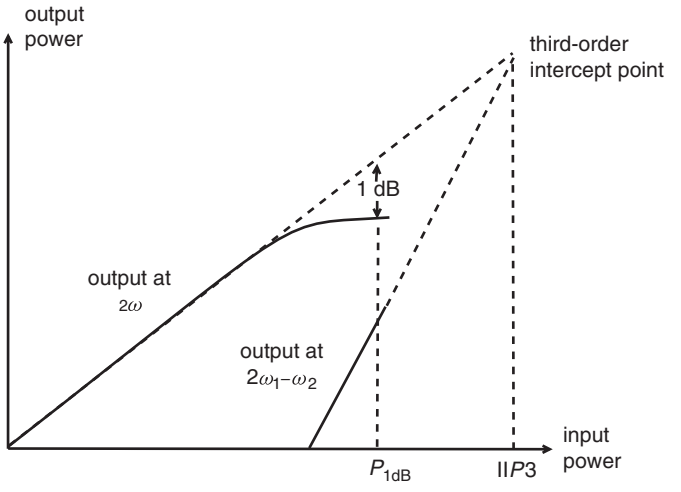


Fig. 4.35 P_{1dB} and third-order intercept point.

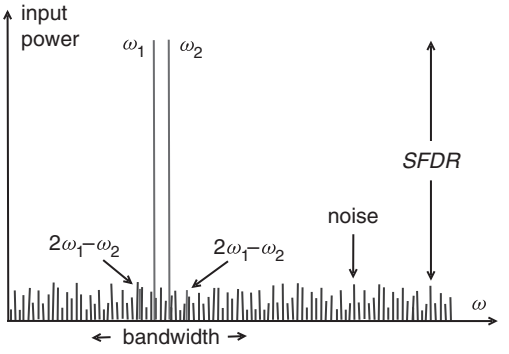


Fig. 4.36 Spurious-free dynamic range (SFDR).

defined to be the input power for which the IMD has value 1 ($IIP3 = |2k_1/3k_3R|$ where R is the input impedance of the receiver). As mentioned in the previous section, we can improve the SNR of a receiver by adding a low-noise preamplifier. We could now ask what impact this will have on the strong-signal behaviour of the receiver, i.e. its IIP3. It turns out that if the receiver consists of two cascaded stages with intercept point $IIP3_1$ for the first stage and $IIP3_2$ for the second stage, the total third-order intercept point $IIP3$ will be given by

$$\frac{1}{IIP3} = \frac{1}{IIP3_1} + \frac{G_1}{IIP3_2}, \tag{4.35}$$

where G_1 is the power gain of the first stage, i.e. the $IIP3$ of the later stages of a receiver has more influence than the earlier stages. As a consequence, whilst the first amplifier of a receiver needs to be designed for minimum noise, later stages should be designed to minimise $IIP3$.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press

In the case of a receiver, an important measure of nonlinearity is its *spurious-free dynamic range* (SFDR). This is the signal whose third-order distortion just reaches the level of the noise when scaled upon the noise. It can be shown that

$$\text{SFDR} = \left(\frac{IIP3}{N_f}\right)^{2/3}, \tag{4.36}$$

where N_f is the noise floor (measured at the receiver input). The SFDR can be regarded as the range of input powers for which the receiver imperfections will remain hidden by the noise.

4.8 Conclusion

In the present chapter, we have considered the active devices that make the amplification of radio signals possible (i.e. valves and transistors). However, these devices have their limitations which we have discussed in detail, along with techniques for mitigating the worst effects. Whilst amplifiers (both audio and radio frequency) can significantly improve a radio based upon the simple crystal detector, active devices can be used in more imaginative ways to produce high-quality radio receivers and transmitters and this is the subject of the next chapter.

https://doi.org/10.1017/9781108684514.005 Published online by Cambridge University Press