3 Semiconductors and Valves

The development of semiconductor technology has had a profound impact on daily life by facilitating unprecedented progress in all areas of electronic engineering and telecommunications. Since the first transistors were made in the late 1940s, the number and variety of semiconductor devices have increased rapidly with the application of advanced technology and new materials. The devices now available range from the humble silicon rectifier diode used in power supplies to specialised GaAsFETs (gallium arsenide field effect transistors) used for low-noise microwave amplification. Digital ULSI ICs (ultra large-scale integrated circuits), involving the fabrication of millions of tiny transistors on a single chip of silicon, are used for the low-power microprocessor and memory functions which have made powerful desktop personal computers a reality.

Within amateur radio, there are now virtually no items of electronic equipment that cannot be based entirely on semiconductor, or 'solid-state', engineering. However, thermionic valves may still be found in some high-power linear amplifiers used for transmission (see the Building Blocks chapter).

SILICON

Although the first transistors were actually made using the semiconductor germanium (atomic symbol Ge), this has now been almost entirely replaced by silicon (Si). The silicon atom, shown pictorially in **Fig 3.1**, has a central nucleus consisting of 14 protons, which carry a positive electrical charge, and also 14 neutrons. The neutrons have no electrical charge, but they do possess the same mass (atomic weight) as the protons. Arranged around the nucleus are 14 electrons. The electrons, which are much lighter particles, carry a negative charge. This balances the positive charge of the protons, making the atom electrically neutral and therefore stable.

The 14 electrons are arranged within three groups, or shells. The innermost shell contains two electrons, the middle shell eight, and the outer shell four. The shells are separated by forbidden regions, known as energy band gaps, into which individual electrons cannot normally travel. However, at temperatures above absolute zero (0 degrees Kelvin or minus 273 degrees Celsius) the electrons will move around within their respective shells due to thermal excitation. The speed, and therefore the energy, of the excited electrons increases as temperature rises.

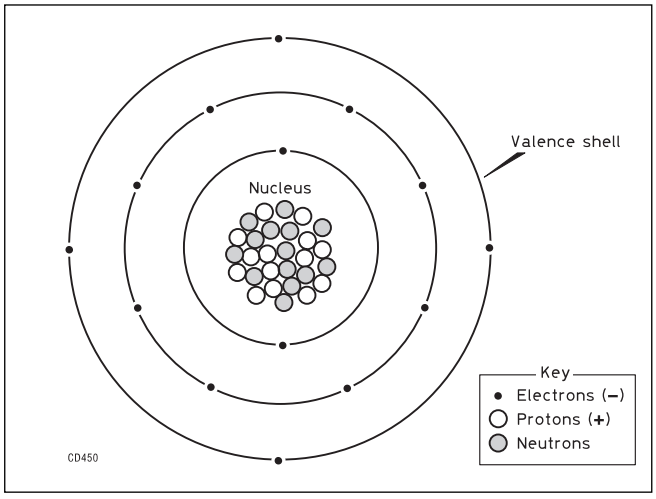


Fig 3.1: The silicon atom, which has an atomic mass of 28.086. 26% of the Earth's crust is composed of silicon, which occurs naturally in the form of silicates (oxides of silicon), eg quartz. Bulk silicon is steel grey in colour, opaque and has a shiny surface

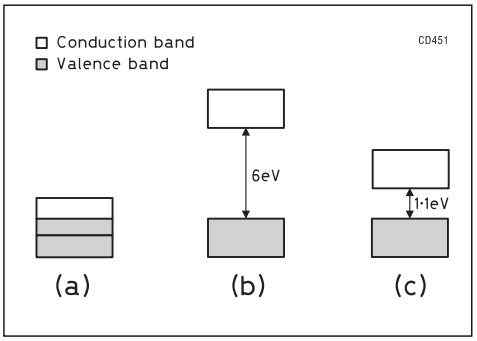


Fig 3.2: Band gaps of conductors (a), a typical insulator (b) and silicon (c)

It is the outermost shell, or valence band, which is of importance when considering whether silicon should be classified as an electrical insulator or a conductor. In conductors (see **Fig 3.2(a)**), such as the metals copper, silver and aluminium, the outermost (valence) electrons are free to move from one atom to another, thus making it possible for the material to sustain electron (current) flow. The region within which this exchange of electrons can occur is known, not surprisingly, as the conduction band, and in a conductor it effectively overlaps the valence band. In insulating materials (eg most plastics, glass and ceramics), however, there is a forbidden region (bandgap) which separates the valence band from the conduction band (**Fig 3.2(b)**). The width of the bandgap is measured in electron volts (eV), 1eV being the energy imparted to an electron as it passes through a potential of 1V. The magnitude of the bandgap serves as an indication of how good the insulator is and a typical insulator will have a bandgap of around 6eV. The only way to force an insulator to conduct an electrical current is to subject it to a very high potential difference, ie many thousands of volts. Under such extreme conditions, the potential may succeed in imparting sufficient energy to the valence electrons to cause them to jump into the conduction band. When this happens, and current flow is instigated, the insulator is said to have broken down. In practice, the breakdown of an insulator normally results in its destruction. It may seem odd that such a high potential is necessary to force the insulator into conduction when the bandgap amounts to only a few volts. However, even a very thin slice of insulating material will have a width of many thousand atoms, and a potential equal to the band gap must exist across each of these atoms before current flow can take place.

Fig 3.2(c) shows the relationship between the valence and conduction bands for intrinsic (very pure) silicon. As can be seen, there is a bandgap of around 1.1eV at room temperature. This means that intrinsic silicon will not under normal circumstances serve as an electrical conductor, and it is best described as a narrow bandgap insulator. The relevance of the term 'semiconductor', which might imply a state somewhere between conduction and insulation, or alternatively the intriguing possibility of being able to move from one to the other, is partly explained by the methods that have been developed to modify the electrical behaviour of materials like silicon.

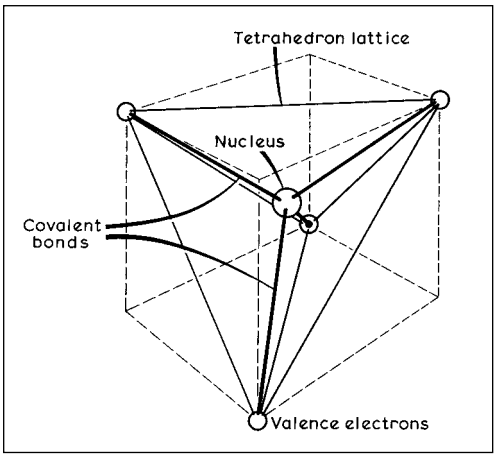


Fig 3.3: A three-dimensional view of the crystal lattice filter

Doping

The atoms within a piece of silicon form themselves into a criss-cross structure known as a tetrahedral crystal lattice - this is illustrated three-dimensionally by **Fig 3.3**. The lattice is held together by a phenomenon called covalent bonding, where each atom shares its valence electrons with those of its four nearest neighbours by establishing electron pairs. For greater clarity, **Fig 3.4** provides a simplified, two-dimensional, representation of the crystal lattice.

It is possible to alter slightly the crystal lattice structure of silicon, and through doing so modify its conductivity, by adding small numbers of atoms of other substances. These substances, termed dopants (also rather misleadingly referred to as 'impurities'), fall into two distinct categories:

- Group III - these are substances that have just three valence electrons: one less than silicon. Boron (atomic symbol B) is the most commonly used. Material doped in this way is called P-type.
- Group V - substances with five valence electrons: one more than silicon. Examples are phosphorus (P) and arsenic (As). This produces N-type material.

The level of doping concentration is chosen to suit the requirements of particular semiconductor devices, but in many cases the quantities involved are amazingly small. There may be only around one dopant atom for every 10,000,000 atoms of silicon (1 in 10^7). Silicon modified in this way retains its normal structure because the Group III and Group V atoms are able to fit themselves into the crystal lattice.

In the case of N-type silicon (**Fig 3.5(a)**, each dopant atom has one spare electron that cannot partake in covalent bonding.

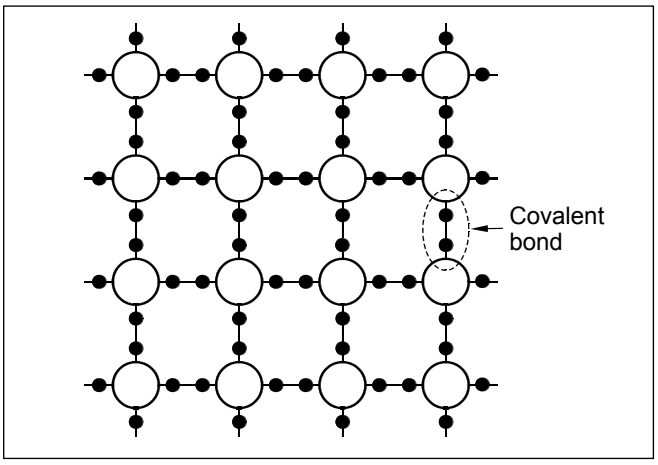


Fig 3.4: Covalent bonding in the crystal lattice filter

These 'untethered' electrons, are free to move into the conduction band and can therefore act as current carriers. The dopant atom is known as a donor, because it has 'donated' an electron to the conduction band. In consequence, N-type silicon has a far higher conductivity than intrinsic silicon. The conductivity of silicon is also enhanced by P-type doping, although the reason for this is less obvious. **Fig 3.5(b)** shows that when a Group III atom fits itself into the crystal lattice an additional electron is 'accepted' by the dopant atom (acceptor), creating a positively charged vacant electron position in the valence band known as a hole. Holes are able to facilitate electron (current) flow because they exert a considerable force of attraction for any free electrons. Semiconductors which utilise both free electrons and holes for current flow are known as bipolar devices. Electrons, however, are more mobile than holes and so semiconductor devices designed to operate at the highest frequencies will usually rely on electrons as their main current carriers.

THE PN JUNCTION DIODE

The PN junction diode is the simplest bipolar device. It serves a vital role both in its own right (in rectification and switching applications), and as an important building block in more complex devices, such as transistors. **Fig 3.6** shows that the junction consists of a sandwich of P- and N-type silicon. This representation suggests that a diode can be made by simply bonding together small blocks of P- and N-type material. Although this is now technically possible, such a technique is not in widespread use.

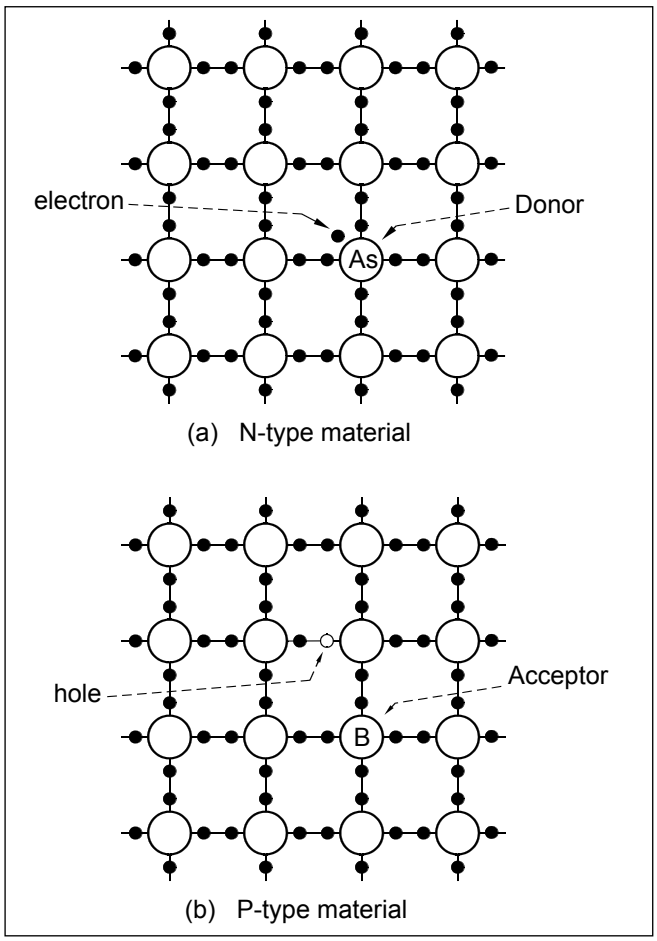


Fig 3.5: (a) The introduction of a dopant atom having five valence electrons (arsenic) into the crystal lattice. (b) Doping with an atom having three valence electrons (boron) creates a hole

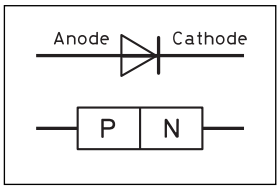


Fig 3.6: The PN junction and diode symbol

The process generally used to fabricate junction diodes is shown in **Fig 3.7**. This is known as the planar process, and variations of this technique are employed in the manufacture of many other semiconductor devices. The starting point is a thin slice of N-type silicon which is called the substrate (see **Fig 3.7(a)**). An insulating layer of silicon dioxide (SiO_2) is thermally grown on the top surface of the substrate in order to protect the silicon underneath. The next step (**Fig 3.7(b)**) is to selectively etch a small hole, referred to as a window, into the oxide layer - often using dilute hydrofluoric acid. A P-type region is then formed by introducing boron dopant atoms into the silicon substrate through the oxide window, using either a diffusion or ion implantation process (**Fig 3.7(c)**). This means, of course, that a region of the N-type substrate has been converted to P-type material. However, the process does not involve removing any of the N-type dopant, as it is simply necessary to introduce a higher concentration of P-type atoms than there are N-type already present. Ohmic contacts (ie contacts in which the current flows equally in either direction) are then formed to both the substrate and window (**Fig 3.7(d)**). In practice, large numbers of identical diodes will be fabricated on a single slice (wafer) of silicon which is then cut into 'chips', each containing a single diode. Finally, metal leads are bonded onto the ohmic contacts before the diode is encapsulated in epoxy resin, plastic or glass.

The diode's operation is largely determined by what happens at the junction between the P-type and N-type silicon. In the absence of any external electric field or potential difference, the large carrier concentration gradients at the junction cause carrier diffusion. Electrons from the N-type silicon diffuse into the P side, and holes from the P-type silicon diffuse into the N-side. As electrons diffuse from the N-side, uncompensated positive donor ions are left behind near the junction, and as holes leave the P-side, uncompensated negative acceptor ions remain. As **Fig 3.8(a)** shows, a negative space charge forms near the P-side of the junction (represented by '-' symbols), and a positive space charge forms near the N-side (represented by '+' symbols). Therefore this diffusion of carriers results in an electrostatic potential difference across the junction within an area known as the depletion region. The magnitude of this built-in potential depends on the doping concentration, but for a typical silicon diode it will be around 0.7V. It is important to realise, however, that this potential cannot be measured by connecting a voltmeter externally across the diode because there is no net flow of current through the device.

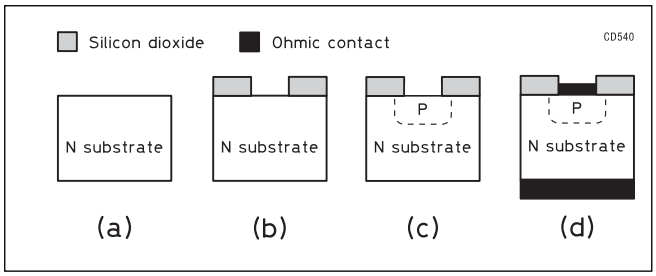


Fig 3.7: Fabrication of a PN junction diode

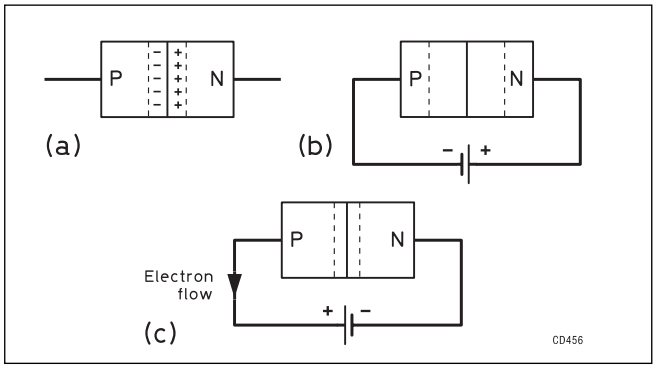


Fig 3.8: Behaviour of the junction diode under zero bias (a), reverse bias (b) and forward bias (c)

If a battery is connected to the diode as in **Fig 3.8(b)**, the external applied voltage will increase the electrostatic potential across the depletion region, causing it to widen (because the polarity of the external voltage is the same as that of the built-in potential). Under these conditions practically no current will flow through the diode, which is said to be reverse biased. However, if the external bias is raised above a certain voltage the diode may 'break down', resulting in reverse current flow.

Reversing the polarity of the battery (**Fig 3.8(c)**) causes the external applied voltage to reduce the electrostatic potential across the depletion region (by opposing the built-in potential), therefore narrowing the depletion region. Under this condition, known as forward bias, current is able to flow through the diode by diffusion of current carriers (holes and electrons) across the depletion region.

The diode's electrical characteristics are summarised graphically in **Fig 3.9**, where a linear scale is used for both current and voltage. In Region A, where only a small forward bias voltage is applied (less than about 0·7V), the current flow is dominated by carrier diffusion, and in fact rises exponentially as the forward bias voltage is increased. Region B commences at a point often termed the 'knee' of the forward bias curve, and here series resistance (both internal and external) becomes the major factor in determining current flow. There remains, however, a small voltage drop which, in the case of an ordinary silicon diode, is around 0·7V .

Under reverse bias (region C) only a very small leakage current flows, far less than 1µA for a silicon diode at room temperature. This current is typically due to generation of electron-hole pairs in the reverse-biased depletion region, and does not vary with reverse voltage. In Region D, where the reverse bias is increased above the diode's breakdown voltage, significant reverse current begins to flow. This is often due to a mechanism

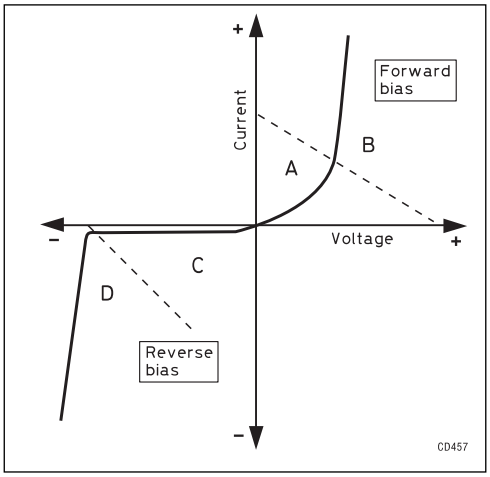


Fig 3.9: Characteristic curve of a PN junction diode

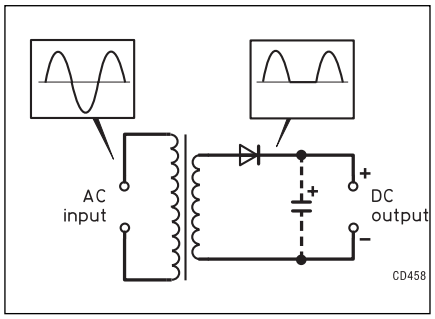


Fig 3.10: The junction diode used as a half-wave rectifier

of avalanche multiplication of electron-hole pairs in the high electric field in the depletion region. Although this breakdown process is not inherently destructive, the maximum current must be limited by an external circuit to avoid excessive heating of the diode.

Diodes have a very wide range of applications in radio equipment, and there are many types available with characteristics tailored to suit their intended use. Fig 3.10 shows a diode used as a rectifier of alternating current in a simple power supply. The triangular part of the diode symbol represents the connection to the P-type material and is referred to as the anode. The single line is the N-type connection, or cathode (Fig 3.6). Small diodes will have a ring painted at one end of their body to indicate the cathode connection (see the lower diode in Fig 3.11). The rectifying action allows current to flow during positive half-cycles of the input waveform, while preventing flow during negative half-cycles. This is known as half-wave rectification because only 50% of the input cycle contributes to the DC output. Fig 3.12 shows how a more efficient power supply may be produced using a bridge rectifier, which utilises four diodes. Diodes A and B conduct during positive half-cycles, while diodes C and D conduct during negative cycles. Notice that C and D are arranged so as to effectively reverse the connections to the transformer's secondary winding during negative cycles, thus enabling the negative half-cycles to contribute to the 'positive' output. Because this arrangement is used extensively in equipment power supplies, bridge rectifier packages containing four interconnected diodes are readily available. When designing power supplies it is necessary to take account of the voltages and currents that the diodes will be subjected to. All rectifiers have a specified maximum forward current at a given case temperature. This is because the diode's forward voltage drop gives rise to power dissipation within the device - for instance, if the voltage across a rectifier diode is 0·8V at a current of 5A, the diode will dissipate 4W (0·8 x 5 = 4). This power will be converted to heat, thus raising the diode's temperature. Consequently, the rectifier diodes of high-current power supplies may need to be mounted on heatsinks. In power supplies especially, diodes must not be allowed to conduct in the reverse direction. For this reason, the maximum reverse voltage that can be tolerated before breakdown is likely to occur must be known. The term peak inverse voltage (PIV) is used to describe this characteristic (see the chapter on power supplies).

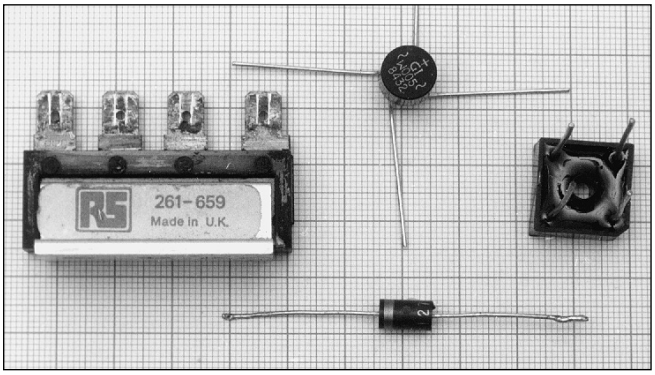


Fig 3.11: Rectifiers. From the left, clockwise: 35A, 100PIV bridge, 1.5A 50PIV bridge, 6A 100PIV bridge, 3A 100PIV diode. Note that the small squares are 1mm across

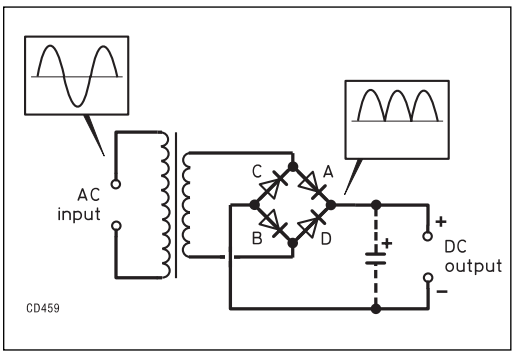


Fig 3.12: A full-wave, or bridge, rectifier using four unijunction diodes

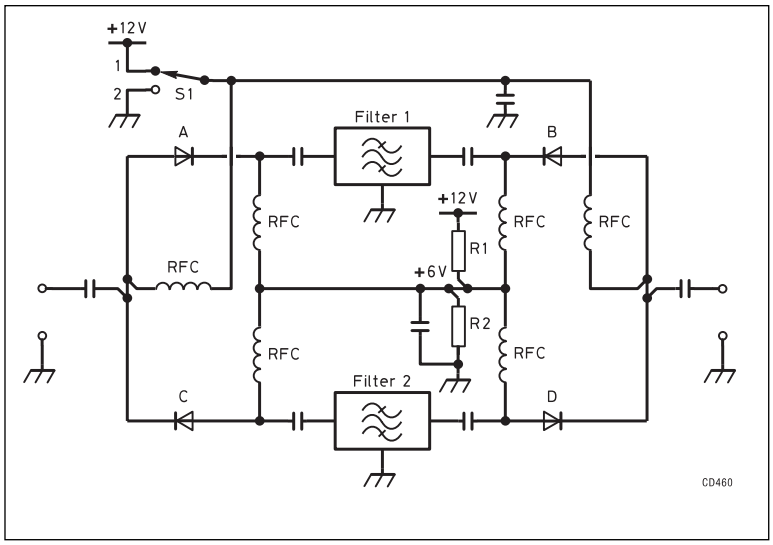


Fig 3.13: Diodes are frequently used as switches in the signal circuitry of transceivers. The setting of S1 determines which of the two filters is selected. R1 and R2 will both have a value of around 2.2kΩ (this determines the forward current of the switching diodes). For HF applications the RFCs are 100µH

Miniature low-current diodes, often referred to as small-signal types, are very useful as switching elements in transceiver circuitry. Fig 3.13 shows an arrangement of four diodes used to select one of two band-pass filters, depending on the setting of switch S1. When S1 is set to position 1, diodes A and B are forward biased, thus bringing filter 1 into circuit. Diodes C and D, however, are reverse biased, which takes filter 2 out of circuit. Setting S1 at position 2 reverses the situation. Notice how the potential divider R1/R2 is used to develop a voltage equal to half that of the supply rail. This makes it easy to arrange for a forward bias of 6V, or a reverse bias of the same magnitude, to appear across the diodes. Providing that the peak level of signals at the filter terminations does not reach 6V under any circumstances, the diodes will behave as almost perfect switches. One of the main advantages of diode switching is that signal paths can be kept short, as the leads (or PCB tracks) forming connections to the front-panel switches carry only a DC potential which is isolated from the signal circuitry using the RF chokes and decoupling capacitors shown.

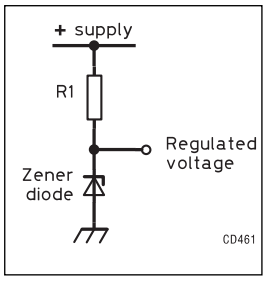


Fig 3.14: A zener diode used as a simple voltage regulator

Zener Diodes

These diodes make use of the reverse breakdown characteristic discussed previously. The voltage at which a diode begins to conduct when reverse biased depends on the doping concentration. As the doping level is increased, the breakdown voltage drops. This fact can be exploited during the manufacture of the diodes, enabling the manufacturer to specify the breakdown voltage for a given component. Zener diodes with breakdown voltages in the range 2·7V to over 150V are available and can be used to provide reference voltages for power supplies and bias generators.

Fig 3.14 shows how a zener diode can be used in conjunction with a resistor to provide voltage regulation (note the use of a slightly different circuit symbol for the zener diode). When power is initially applied, the zener diode will start to conduct as the input voltage is higher than the diode's reverse breakdown value. However, as the diode begins to pass current, an increasingly high potential difference will appear across resistor R1. This potential will tend to rise until the voltage across R1 becomes equal to the difference between the input voltage and the zener's breakdown voltage. The net result is that the output voltage will be forced to settle at a level close to the diode's reverse breakdown potential. The value of R1 is chosen so as to limit the zener current to a safe value (the maximum allowable power dissipation for small zener diodes is around 400mW), while ensuring that the maximum current to be drawn from the regulated supply will not increase the voltage across R1 to a level greater than the difference between the zener voltage and the minimum expected input voltage (see the chapter on power supplies).

Varactor Diodes

The varactor, or variable capacitance, diode makes use of the fact that a reverse biased PN junction behaves like a parallel-plate capacitor, where the depletion region acts as the spacing between the two capacitor plates. As the reverse bias is increased, the depletion region becomes wider. This produces the same effect as moving the plates of a capacitor further apart - the capacitance is reduced (see the fundamentals chapter). The varactor can therefore be used as a voltage-controlled variable capacitor, as demonstrated by the graph in Fig 3.15. The capacitance is governed by the diode's junction area (ie the area of the capacitor plates), and also by the width of the depletion region for a given value of reverse bias (which is a function of the doping concentration). Varactors are available covering a wide range of capacitance spreads, from around 0·5-10pF up to 20-400pF. The voltages at which the stated maximum and minimum capacitances are obtained will be quoted in the manufacturer's literature, but they normally fall in the range 2-20V. The maximum reverse bias voltage should not be exceeded as this could result in breakdown.

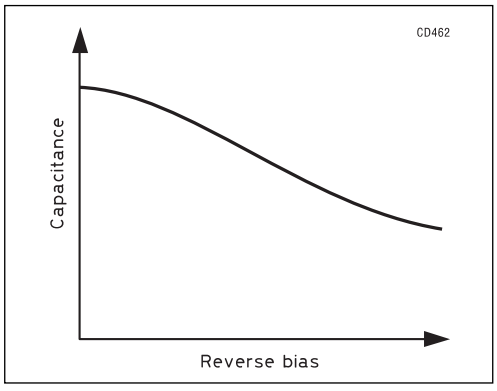


Fig 3.15: Relationship between capacitance and reverse bias of a varactor (variable capacitance diode

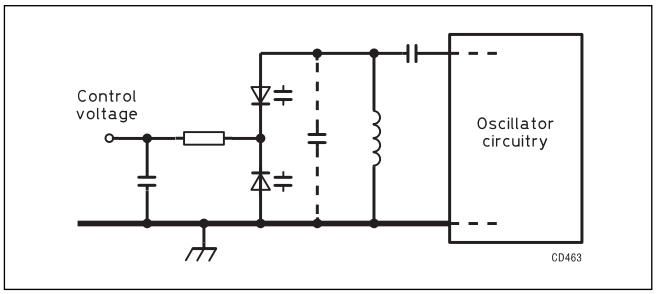


Fig 3.16: Two varactor diodes used to tune a voltage-controlled oscillator. The capacitor drawn with dotted lines represents the additional component which may be added to the tank circuit in order to modify the LC ratio and tuning range

Varactors are commonly used to achieve voltage control of oscillator frequency in frequency synthesisers. Fig 3.16 shows a typical arrangement where two varactors are connected 'back to back' and form part of the tank circuit of a voltage-controlled oscillator (VCO). The use of two diodes prevents the alternating RF voltage appearing across the tuned circuit from driving the varactors into forward conduction, which is most likely to happen when the control voltage is low. Because the varactor capacitances appear in series, the maximum capacitance swing is half that obtainable when using a single diode. Three-lead packages containing dual diodes internally connected in this way are readily available. It is also possible to obtain multiple-diode packages containing two or three matched diodes but with separate connections. These are used to produce voltage-controlled versions of two-gang or three-gang variable capacitors.

When used as a capacitive circuit element, the varactor's Q may be significantly lower than that of a conventional capacitor. This factor must be taken into account when designing high-performance frequency synthesisers (see the chapter on oscillators).

PIN Diodes

A PIN diode, Fig 3.17, is a device that operates as a variable resistor at RF and even into microwave frequencies. Its resistance is determined only by its DC excitation. It can also be used to switch quite large RF signals using smaller levels of DC excitation.

The PIN diode chip consists of a chip of pure (intrinsic or I-type) silicon with a layer of P-type silicon on one side and a layer of N-type on the other. The thickness of the I-region (W) is only a little smaller than the thickness of the original wafer from which the chip was cut.

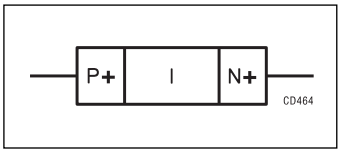


Fig 3.17: The PIN diode

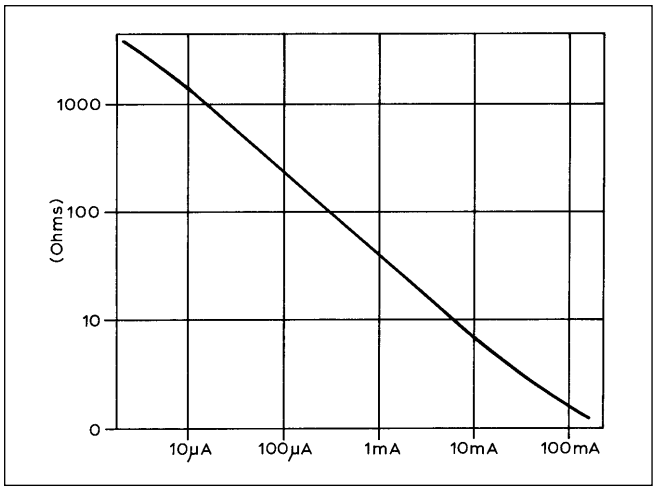


Fig 3.18: Relationship between forward current and resistance for a PIN dode

Forward-biased PIN diodes

When forward-biased electrons and holes are injected into the I-region from the N- and P-regions respectively, these charges don't recombine immediately but a finite charge remains stored in the I-layer.

The quantity of stored charge, Q, depends on the recombination time (t, usually called the carrier lifetime) and the forward current I as:

$$Q = It$$

The resistance of the I-layer is proportional to the square of its thickness (W) and inversely proportional to Q, and is given by:

$$R = \frac{W^2}{(N+p)Q}$$

Combining these two equations, we get:

$$R = \frac{W^2}{(N+p)It}$$

Thus the resistance is inversely proportional to the DC excitation I. For a typical PIN diode, R varies from 0·1Ω at 1A to 10kΩ at 1µA. This is shown in **Fig 3.18**.

This resistance-current relationship is valid over a wide frequency range, the limits at low resistance being set by the parasitic inductance of the leads, and at high resistance by the junction capacitance. The latter is small owing to the thickness of the I-layer which also ensures that there is always sufficient charge so that the layer presents a constant resistance over the RF cycle.

Reverse-biased PIN diodes

At high RF, a reverse-biased PIN diode behaves as a capacitor independent of the reverse bias with a value of:

$$C = \varepsilon A/W \quad \text{farads}$$

where ε is the permittivity of silicon. Note that although the dielectric constant is quite high, the capacitance of the diode is small because of the small area of the diode and the thickness of the intrinsic layer. Therefore as a switch the isolation in the reverse bias state is good but gets slightly worse as the frequency increases. This is not as serious as it seems since nearly all PIN diode switches at VHF are operated in 50Ω circuits and the reactance is high compared to 50Ω.

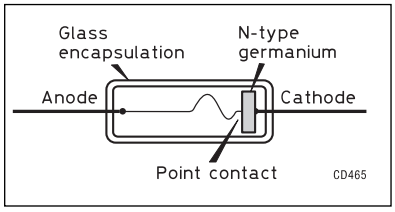


Fig 3.19: The germanium point-contact diode

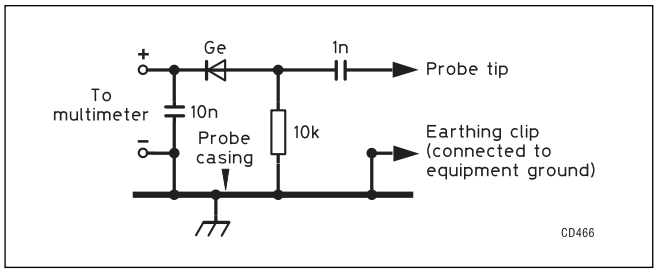


Fig 3.20: A simple RF probe using a germanium (Ge) point-contact diode

Germanium Point-contact Diodes

Ironically, one of the earliest semiconductor devices to find widespread use in telecommunications actually pre-dates the thermionic valve. The first broadcast receivers employed a form of envelope detector (RF rectifier) known colloquially as a cat's whisker. This consisted of a spring made from a metal such as bronze or brass (the 'whisker'), the pointed end of which was delicately bought into contact with the surface of a crystal having semiconducting properties, such as galena, zincite or carborundum.

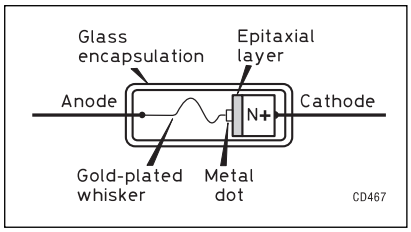
The germanium point-contact diode is a modern equivalent of the cat's whisker and consists of a fine tungsten spring which is held in contact with the surface of an N-type germanium crystal (**Fig 3.19**). During manufacture, a minute region of P-type material is formed at the point where the spring touches the crystal. The point contact therefore functions as a PN diode. In most respects the performance of this device is markedly inferior to that of the silicon PN junction diode. The current flow under reverse bias is much higher - typically 5mA, the highest obtainable PIV is only about 70V and the maximum forward current is limited by the delicate nature of the point contact.

Nevertheless, this device has a number of saving graces. The forward voltage drop is considerably lower than that of a silicon junction diode - typically 0·2V - and the reverse capacitance is also very small. There is also an improved version known as the gold-bonded diode, where the tungsten spring is replaced by one made of gold. **Fig 3.20** shows a simple multimeter probe which is used to rectify low RF voltages. The peak value can then be read with the meter switched to a normal DC range. The low forward voltage drop of the point-contact diode leads to more accurate readings.

The Schottky Barrier Diode

Ordinary PN junction diodes suffer from a deficiency known as charge storage, which has the effect of increasing the time taken for a diode to switch from forward conduction to reverse cut-off when the polarity of the applied voltage is reversed. This reduces the efficiency of the diode at high frequencies. Charge storage occurs because holes, which are less mobile than electrons, require a finite time to migrate back from the N-doped cathode material as the depletion region widens under the influence of reverse bias. The fact that in most diodes the P-type anode is more heavily doped than the N-type cathode tends to exacerbate matters.

The hot-carrier, or Schottky barrier, diode overcomes the problem of charge storage by utilising electrons as its main current carriers. It is constructed (**Fig 3.21**) in a similar fashion to the



(above) Fig 3.21: Hot-carrier (Schottky) diode

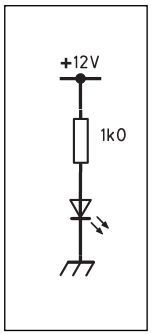


Fig 3.22: An LED emits light when forward biased. The series resistor limits the forward current to a safe value

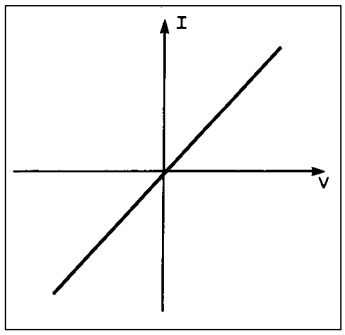


Fig 3.23: Voltage-current through a resistor

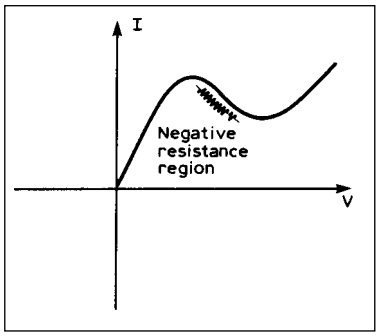


Fig 3.24: Voltage-current through a Gunn diode

germanium point-contact type, but there are a few important differences. The semiconductor used is N-type silicon which is modified by growing a layer (the epitaxial region) of more lightly doped material onto the substrate during manufacture. The device is characterised by its high switching speed and low capacitance. It is also considerably more rugged than the germanium point-contact diode and generates less noise.

Hot-carrier diodes are used in high-performance mixers of the switching, or commutating, type capable of operating into the microwave region.

Light-emitting diodes (LEDs)

The LED consists of a PN junction formed from a compound semiconductor material such as gallium arsenide (GaAs) or gallium phosphide (GaP). As gallium has a valency of three, and arsenic and phosphorus five, these materials are often referred to as Group III-V semiconductors.

When electrons recombine with holes across the energy gap of a semiconductor, as happens around the depletion layer of a forward-biased PN junction, particles of light energy known as photons are released. The energy, and therefore wavelength, of the photons is determined by the semiconductor band gap. Pure gallium arsenide has a band gap of about 1·43eV, which produces photons with a wavelength of 880 nanometres (nm). This lies at the infrared end of the spectrum. Adding aluminium to the gallium arsenide has the effect of increasing the band gap to 1·96eV which shortens the light wavelength to 633nm. This lies in the red part of the visible spectrum. Red LEDs can also be made by adding phosphorus to gallium arsenide. Green LEDs (wavelength 560nm) are normally made from gallium phosphide.

The LEDs used as front-panel indicators consist of a PN junction encapsulated within translucent plastic. At a current of 10mA they will generate a useful amount of light without overheating. The forward voltage drop at this current is about 1·8V. **Fig 3.22** shows a LED operating from a 12V power rail (note the two arrows representing rays of light which differentiates the LED circuit symbol from that of a normal diode). The series resistor determines the forward current and so its inclusion is mandatory. LEDs are also used in more complex indicators, such the numeric (seven-segment) displays employed in frequency counters (see the chapter on test equipment).

The Gunn diode

The Gunn diode, named after its inventor, B J Gunn, comprises little more than a block of N-type gallium arsenide. It is not, in fact, a diode in the normally accepted sense because there is no P-N junction and consequently no rectifying action. It is properly called the Gunn effect device. However, 'diode' has become accepted by common usage and merely explains that it has two connections.

It consists of a slice of low-resistivity N-type gallium arsenide on which is grown a thin epitaxial layer, the active part, of high-resistivity gallium arsenide with a further thicker layer of low-resistivity gallium arsenide on top of that. Since the active layer is very thin, a low voltage across it will produce a high electric field strength. The electrons in gallium arsenide can be in one of two conduction bands. In one they have a much higher mobility than in the other and they are initially in this band. As the electric field increases, more and more are scattered into the lower mobility band and the average velocity decreases. The field at which this happens is called the threshold field and is 320V/mm. Since current is proportional to electron velocity and voltage is proportional to electric field, the device has a region of negative resistance. This odd concept is explained by the definition of resistance as the slope of the voltage-current graph and a pure resistor has a linear relationship (**Fig 3.23**). On the other hand, the Gunn diode has a roughly reversed 'S'-shaped curve (**Fig 3.24**) and the negative resistance region is shown by the hatching. The current through the device takes the form of a steady DC with superimposed pulses (**Fig 3.25**) and their fre-

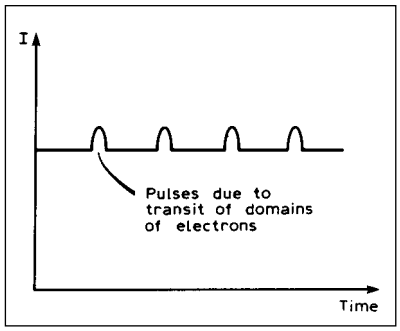


Fig 3.25: Current in Gunn diode

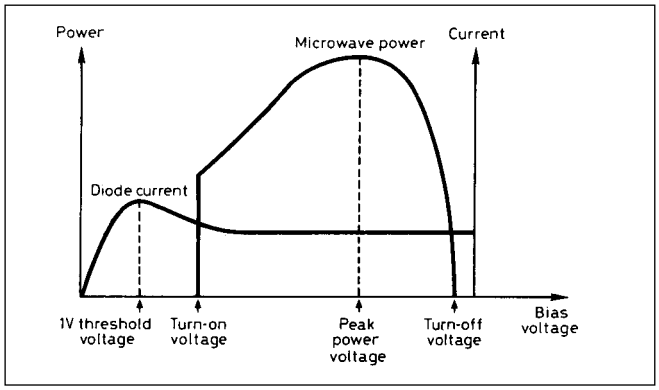


Fig 3.26: Characteristic shape of a Gunn oscillator's bias power curve

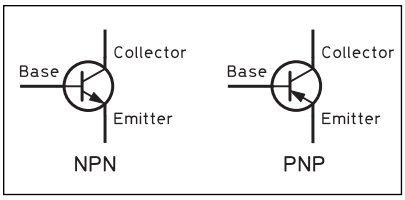


Fig 3.27: Circuit symbols for the bipolar transistor

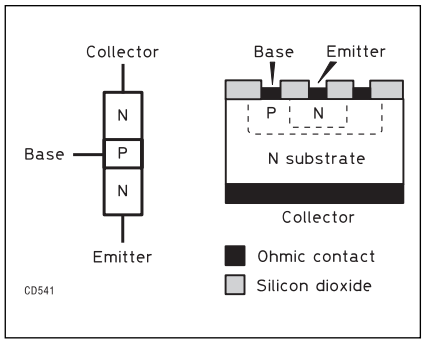


Fig 3.28: Construction of an NPN transistor

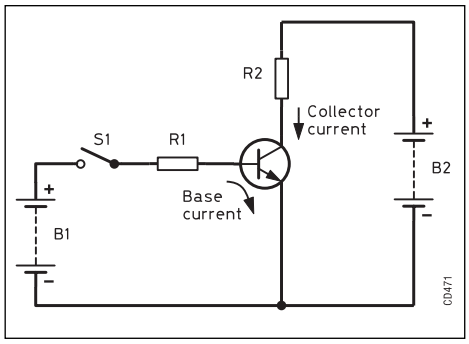


Fig 3.29: Applying forward bias to the base-emitter junction of a bipolar transistor causes a much larger current to flow between the collector and emitter. The arrows indicate conventional current flow, which is opposite in direction to electron flow

quency is determined by the thickness of the active epitaxial layer. As each pulse reaches the anode, a further pulse is generated and a new domain starts from the cathode. Thus the rate of pulse formation depends on the transit time of these domains through the epitaxial layer. In a 10GHz Gunn diode, the layer is about 10µm thick, and a voltage of somewhat greater than 3·5V gives the high field state and microwave pulses are generated. The current through the device shows a peak before the threshold voltage followed by a plateau. See **Fig 3.26**. The power output reaches a peak at between 7·0 and 9·0V for 10GHz devices.

The Gunn diode is inherently a wide-band device so it is operated in a high-Q cavity (tuned circuit) and this determines the exact frequency. It may be tuned over a narrow range by altering the cavity with a metallic screw, a dielectric (PTFE or Nylon) screw or by loading the cavity with a varactor. With a 10GHz device, the whole of the 10GHz amateur band can be covered with reasonable efficiency. Gunn diodes are not suitable for narrow-band operation since they are of low stability and have relatively wide noise sidebands. The noise generated has two components, thermal noise and a low frequency 'flicker' noise. Analysis of the former shows that it is inversely proportional to the loaded Q of the cavity and that FM noise close to the carrier is directly proportional to the oscillator's voltage pushing, ie to the variation of frequency caused by small variations of voltage. Clearly, the oscillator should be operated where this is a minimum and that is often near the maximum safe bias.

THE BIPOLAR TRANSISTOR

The very first bipolar transistor, a point-contact type, was made by John Bardeen and Walter Brattain at Bell Laboratories in the USA during December 1947. A much-improved version, the bipolar junction transistor, arrived in 1950 following work done by another member of the same team, William Schockley. In recognition of their pioneering work in developing the first practical transistor, the three were awarded the Nobel Prize for physics in 1956.

The bipolar transistor is a three-layer device which exists in two forms, NPN and PNP. The circuit symbols for both types are shown in **Fig 3.27**. Note that the only difference between the symbols is the direction of the arrow drawn at the emitter connection.

Fig 3.28 shows the three-layer sandwich of an NPN transistor and also how this structure may be realised in a practical device. The emitter region is the most heavily doped.

In **Fig 3.29** the NPN transistor has been connected into a simple circuit to allow its operation to be described. The PN junction between the base and emitter forms a diode which is forward biased by battery B1 when S1 is closed. Resistor R1 has been included to control the level of current that will inevitably flow, and R2 provides a collector load. A voltmeter connected between the base and emitter will indicate the normal forward voltage drop of approximately 0·6V typical for a silicon PN junction.

The collector-base junction also forms the equivalent of a PN diode, but one that is reverse biased by battery B2. This suggests that no current will flow between the collector and the emitter. This is indeed true for the case where S1 is open, and no current is flowing through the base-emitter junction. However, when S1 is closed, current does flow between the collector and emitter. Due to the forward biasing of the base-emitter junction, a large number of electrons will be injected into the P-type base from the N-type emitter. Crucially, the width of the base is made less than the diffusion length of electrons, and so most of these injected carriers will reach the reverse biased collector-base junction. The electric field at this junction is such that these electrons will be swept across the depletion region into the collector, thus causing significant current to flow between the emitter and collector.

However, not all the electrons injected into the base will reach the collector, as some will recombine with holes in the base. Also, the base-emitter junction forward bias causes holes to be injected from the P-type base to the N-type emitter, and electrons and holes will recombine in the base-emitter depletion region. These are the three main processes which give rise to the small base current. The collector current will be significantly larger than the base current, and the ratio between the two (known as the transistor's ß) is related to the doping concentration of the emitter divided by the doping concentration of the base. The PNP transistor operates in a similar fashion except that the polarities of the applied voltages, and also the roles of electrons and holes, are reversed. The graph at **Fig 3.30** shows

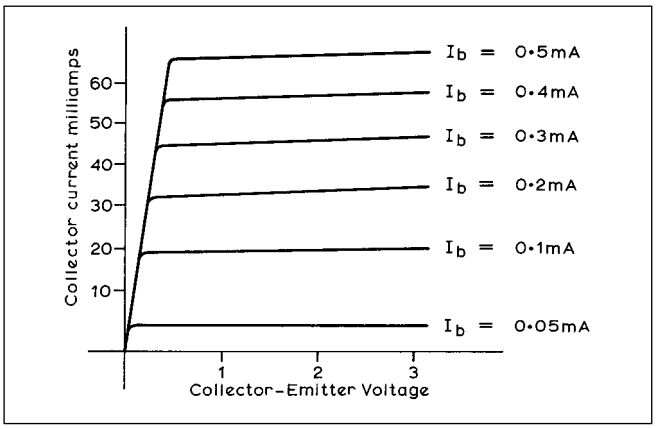


Fig 3.30: The relationship between base current and collector current for a bipolar transistor

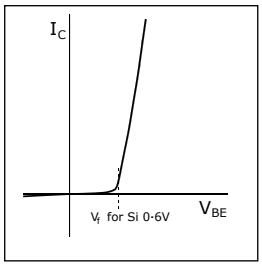


Fig 3.31: The relationship between base-emitter voltage and collector current for a bipolar transistor

the relationship between base current (I_B)and collector current (I_C)for a typical bipolar transistor. The point to note is that the collector current is very much determined by the base current (I_C = ßI_B) and the collector-emitter voltage, V_{CE}, has comparatively little effect. The relationship between base-emitter voltage, V_{BE}, and collector current I_C is shown in **Fig 3.31**. The graph should be compared with the diode characteristics previously shown in Fig 3.9. The graph is almost identical but the current axis is ß times greater due to the current gain of the transistor. It should be noted that the graph of transistor base current against applied voltage I_B/V_{BE} is simply the graph of the base-emitter 'diode'.

The circuit shown in Fig 3.29 therefore provides the basis for an amplifier. Small variations in base current will result in much larger variations in collector current. The DC current gain (ß or h_{FE}), of a typical bipolar transistor at a collector current of 1mA will be between 50 and 500, ie the change in base current required to cause a 1mA change in collector current lies in the range 2 to 20µA. The value of ß is temperature dependent, and so is the base-emitter voltage drop (V_{BE}), which will fall by approximately 2mV for every 1°C increase in ambient temperature.

A Transistor Amplifier

If the battery (B1) used to supply base current to the transistor in Fig 3.29 is replaced with a signal generator set to give a sine-wave output, the collector current will vary in sympathy with the input waveform as shown in **Fig 3.32**. A similar, although inverted, curve could be obtained by plotting the transistor's collector voltage.

Two problems are immediately apparent. First, because the base-emitter junction only conducts during positive half-cycles of the input waveform, the negative half-cycles do not appear at the output. Second, the barrier height of the base-emitter junction results in the base current falling exponentially as the amplitude of the input waveform drops below +0·6V. This causes significant distortion of the positive half-cycles, and it is clear that if the amplitude of the input waveform were to be significantly reduced then the collector current would hardly rise at all.

The circuit can be made far more useful by adding bias. **Fig 3.33** shows the circuit of a practical common-emitter amplifier

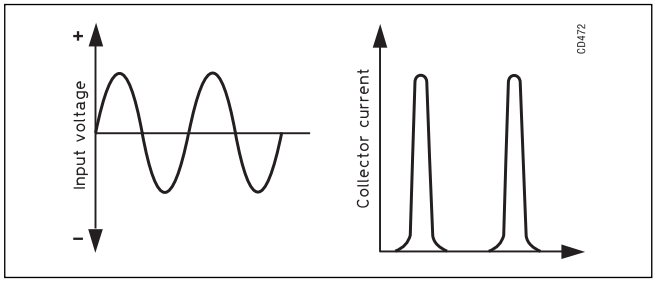


Fig 3.32: The circuit shown in Fig 3.29 would not make a very good amplifier

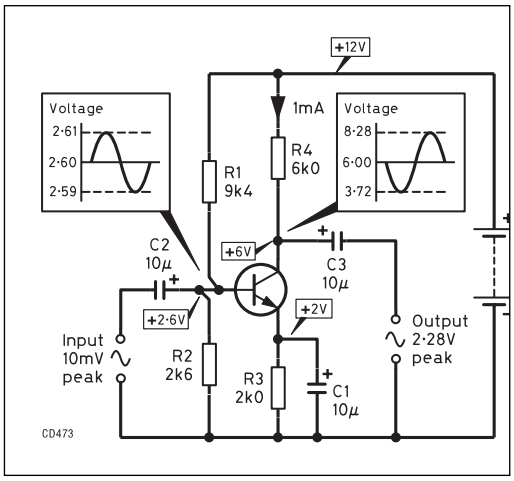


Fig 3.33: A practical common-emitter amplifier

operating in Class A. The term 'common emitter' indicates that the emitter connection is common to both the input and output circuits. Resistors R1 and R2 form a potential divider which establishes a positive bias voltage at the base of the transistor. The values of R1 and R2 are chosen so that they will pass a current at least 10 times greater than that flowing into the base. This ensures that the bias voltage will not alter as a result of variations in the base current. R3 is added in order to stabilise the bias point, and its value has been calculated so that a potential of 2V will appear across it when, in the absence of an input signal, the desired standing collector current of 1mA (0·001A) flows (0·001A x 2000 = 2V). For this reason, the ratio between the values of R1 and R2 has been chosen to establish a voltage of 2·6V at the base of the transistor. This allows for the expected forward voltage drop of around 0·6V due to the barrier height of the base-emitter junction. Should the collector current attempt to rise, for instance because of an increase in ambient temperature causing V_{BE} to fall, the voltage drop across R3 will increase, thus reducing the base-emitter voltage and preventing the collector current rising. Capacitor C1 provides a low-impedance path for alternating currents so that the signal is unaffected by R3, and coupling capacitors C2 and C3 prevent DC potentials appearing at the input and output. The value of the collector load resistor (R4) is chosen so that in the absence of a signal the collector voltage will be roughly 6V - ie half that of the supply rail. Where a transistor is used as an RF or IF amplifier, the collector load resistor may be replaced by a parallel tuned circuit (see the building blocks chapters).

An alternating input signal will now modulate the base voltage, causing it to rise slightly during the positive half-cycles, and fall during negative half-cycles. The base current will be similarly modulated and this, in turn, will cause far larger variations in the collector current. Assuming the transistor has a ß of 100, the voltage amplification obtained can gauged as follows:

In order to calculate the effect of a given input voltage, it is necessary to develop a value for the base resistance. This can be approximated for a low-frequency amplifier using the formula:

$$\text{Base resistance} = \frac{26 \times \beta}{\text{Emitter current in mA}}$$

The emitter current is the sum of the base and collector currents but, as the base current is so much smaller than the collector current, it is acceptable to use just the collector current in rough calculations:

$$\text{Base resistance} = \frac{26 \times 100}{1} = 2600\Omega$$

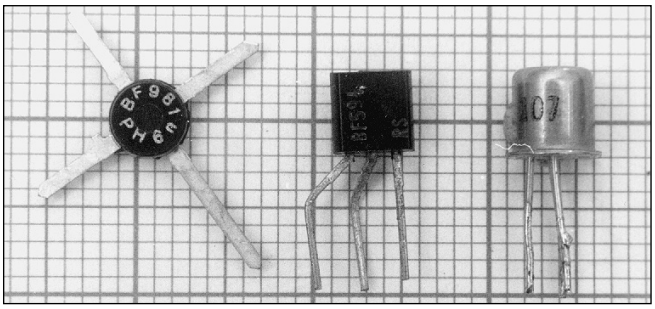


Fig 3.34: Power devices. From the left, clockwise: NPN bipolar power Darlington transistor, power audio amplifier IC, NPN bipolar power transistor. The small squares are 1mm across

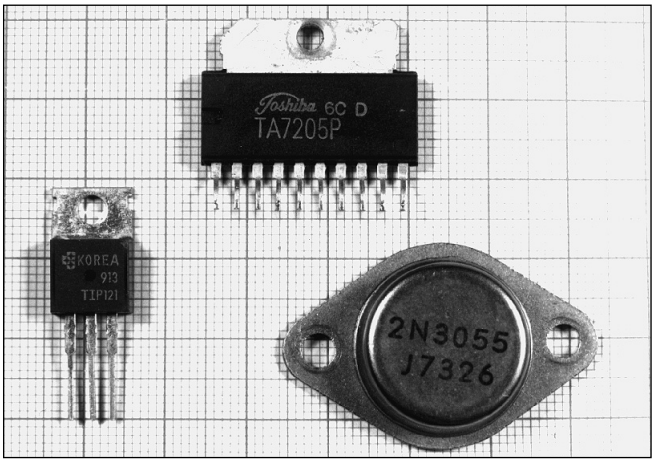


Fig 3.35: Power devices. From the left, clockwise: NPN bipolar power Darlington transistor, power audio amplifier IC, NPN bipolar power transistor. The small squares are 1mm across

Using Ohm's Law, and assuming that the peak amplitude of the input signal is 10mV (0·01V), the change in base current will be:

$$\frac{0\cdot01}{2600} = 3\cdot8\mu A$$

This will produce a change in collector current 100 times larger, that is 380µA.

The change in output voltage (again using Ohm's Law) is:

380µA x 6kΩ (R4) = 2·28V

(3·8 x 10-4 6 103 = 2·28)

The voltage gain obtained is therefore:

$$\frac{2\cdot28}{0\cdot01} = 228 \quad \text{or } 47\text{dB}$$

Note that the output voltage developed at the junction between R4 and the collector is phase reversed with respect to the input. As operating frequency is increased the amplifier's gain will start to fall. There are a number of factors which cause this, but one of the most significant is charge storage in the base region. For this reason, junction transistors designed to operate at high frequencies will be fabricated with the narrowest possible base width. A guide to the maximum frequency at which a transistor can be operated is given by the parameter f_T. This is the frequency at which the ß falls to unity (1). Most general-purpose junction transistors will have an f_T of around 150MHz, but specialised devices intended for use at UHF and microwave frequencies will have an f_T of 5GHz or even higher. An approximation of a transistor's current gain at frequencies below f_T can be obtained by:

ß = f_T ÷ operating frequency

For example, a device with an f_T of 250MHz will probably exhibit a current gain of around 10 at a frequency of 25MHz.

Maximum Ratings

Transistors may be damaged by the application of excessive voltages, or if made to pass currents that exceed the maximum values recommended in the manufacturers' data sheets. Some, or all, of the following parameters may need to be considered when selecting a transistor for a particular application:

V_{CEO} The maximum voltage that can be applied between the collector and emitter with the base open-circuit (hence the 'O'). In practice the maximum value is dictated by the reverse breakdown voltage of the collector-base junction. In the case of some transistors, for instance many of those intended for use as RF power amplifiers, this rating may seem impracticably low, being little or no higher than the intended supply voltage. However, in practical amplifier circuits, the base will be connected to the emitter via a low-value resistance or coupling coil winding. Under these conditions the collector-base breakdown voltage will be raised considerably (see below).

V_{CBO} The maximum voltage that can be applied between the collector and base with the emitter open-circuit. This provides a better indication of the collector-base reverse breakdown voltage. An RF power transistor with a V_{CEO} of 18V may well have a V_{CBO} rating of 48V. Special high-voltage transistors are manufactured for use in the EHT (extra high tension) generators of television and computer displays which can operate at collector voltages in excess of 1kV.

V_{EBO} The maximum voltage that can be applied between the emitter and base with the collector open-circuit. In the case of an NPN transistor, the emitter will be held at a positive potential with respect to the base. Therefore, it is the reverse breakdown voltage of the emitter-base junction that is being measured. A rating of around 5V can be expected.

I_C The maximum continuous collector current. For a small-signal transistor this is usually limited to around 150mA, but a rugged power transistor may have a rating as high as 30A.

P_D The maximum total power dissipation for the device. This figure is largely meaningless unless stated for a particular case temperature. The more power a transistor dissipates, the hotter it gets. Excessive heating will eventually lead to destruction, and so the power rating is only valid within the safe temperature limits quoted as part of the P_D rating. A reasonable case temperature for manufacturers to use in specifying the power rating is 50°C. It is unfortunate that as a bipolar transistor gets hotter, its V_{BE} drops and its ß increases. Unless the bias voltage is controlled to compensate for this, the collector current may start to rise, which in turn leads to further heating and eventual destruction of the device. This phenomenon is known as thermal runaway.

The possibility of failure due to the destruction of a transistor junction by excessive voltage, current or heating is best avoided by operating the device well within its safe limits at all times (see below). In the case of a small transistor, junction failure, should it occur, is normally absolute, and therefore renders the device useless. Power transistors, however, have a more complex construction. Rather than attempting to increase the junction area

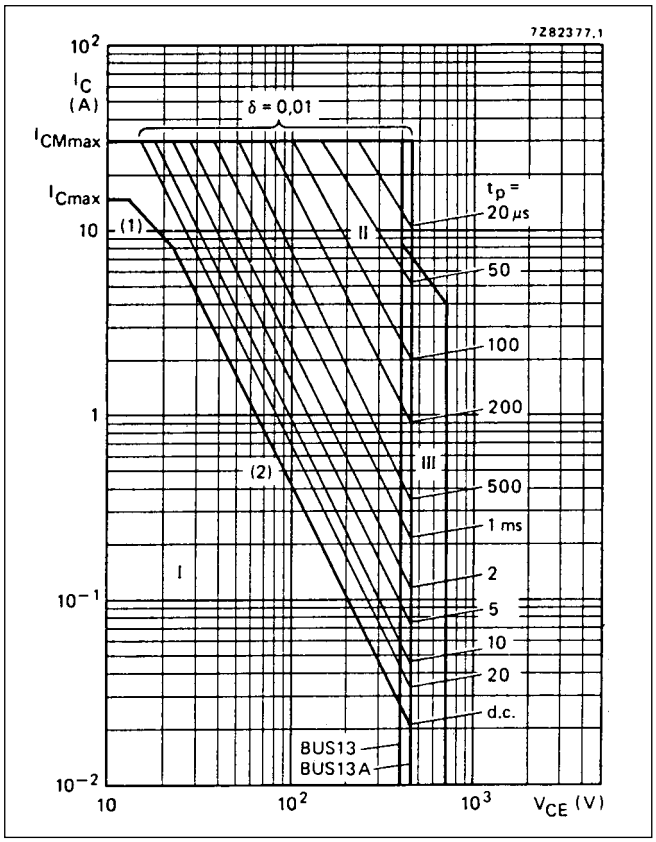


Fig 3.36: The safe operating (SOAR) curves for a BUS13A power bipolar transistor. When any power transistor is used as a pass device in a regulated PSU, it is important to ensure that the applied voltage and current are inside the area on the graph marked 'I - region of permissible DC operation'. Failure to do so may result in the device having a very short life because of secondary breakdown. This is due to the formation of hot spots in the transistor's junction. Note that power FETs do not suffer from this limitation. Key to regions: I - Region of permissible DC operation. II - Permissible extension for repetitive pulse operation. III - Area of permissible operation during turn-on in single-transistor converters, provided $R_{BE} \leq 100\Omega$ and $t_p \leq 0.6\mu s$. (Reproduction courtesy Philips Semiconductors)

of an individual transistor in order to make it more rugged, power transistors normally consist of large numbers of smaller transistors fabricated on a single chip of silicon. These are arranged to operate in parallel, with low-value resistances introduced in series with the emitters to ensure that current is shared equally between the individual transistors. A large RF power transistor may contain as many as 1000 separate transistors. In such a device, the failure of a small number of the individual transistors may not unduly affect its performance. This possibility must sometimes be taken into account when testing circuitry which contains power transistors.

Safe operating area - SOAR

All bipolar transistors can fail if they are over-run and power bipolars are particularly susceptible since, if over-run by excessive power dissipation, hot spots will develop in the transistor's junction, leading to total destruction. To prevent this, manufacturers issue SOAR data, usually in the form of a graph or series of graphs plotted on log-log graph paper with current along one axis and voltage along the other. A typical example is shown in **Fig 3.36**.

Most amateur use will be for analogue operation and should be confined to area I in the diagram (or, of course, to the corresponding area in the diagram of the transistor being considered). For pulse operation, it is possible to stray out into area II.

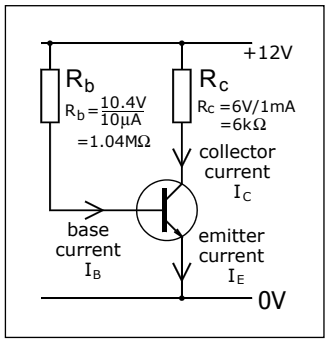


Fig 3.37: Simple bias circuit offering no bias stability

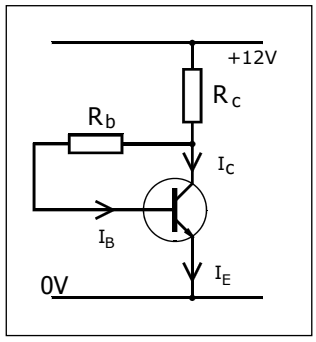


Fig 3.38: Improved bias circuit with some bias stability

How far depends on the height of the current pulses and the duty cycle. Generally speaking, staying well within area I will lead to a long life.

All bipolar transistors have this SOAR but, in the case of small devices, it is not often quoted by the manufacturers and, in any case, is most unlikely to be exceeded.

Other Bias Circuits and Classes of Bias

Fig 3.33 showed a transistor amplifier biased to bring the operating point of the transistor, that is the standing or quiescent voltages and currents, to a mid-range value allowing a signal to vary those currents and voltages both up and down to provide a maximum available range or swing. It is important that these values are predictable and tolerant of the variation in characteristics, particularly gain, from one transistor to another, albeit of the same type and part number.

The circuit in **Fig 3.37** would provide biasing. The base current required is calculated as 1mA/ß (assuming I_c = 1mA) and the resistor R_b chosen accordingly. ß has been assumed to be 100. However the value of ß is not well controlled and may well vary over a 4:1 range. In this circuit, if ß actually was 200, the collector current would double, with 12V across R_c and almost none across the transistor. Clearly unsatisfactory, the design is not protected against the vagaries of the transistor ß.

If the top of R_b was moved from the supply rail to the collector of the transistor, as shown in **Fig 3.38**, then an increase in collector current (due to ß being higher than expected) would cause a fall in collector voltage and a consequent drop in the voltage across R_b. This would reduce the base current into the transistor, offsetting the rise in collector current. The circuit is reason-

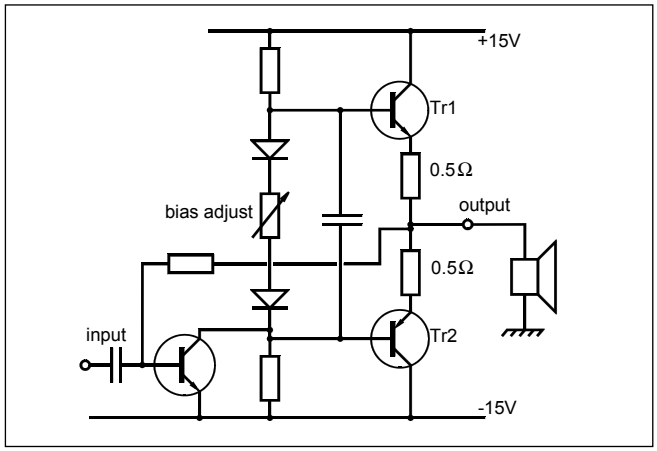


Fig 3.39: A push-pull amplifier using two transistors biased in Class B

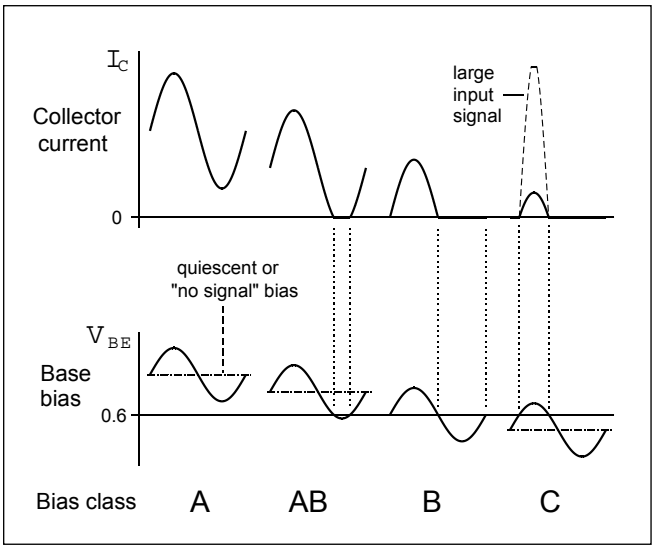


Fig 3.40: Classes of bias relate to the proportion of the signal waveform for which collector current flows

ably stable against variations in gain and is sometimes used although R_b does feed some of the output signal back to the base, which, being inverted or in anti-phase, does reduce the signal gain.

The method of bias stability in Fig 3.33 is better. If the collector current is higher than calculated, perhaps because ß is greater, the potential difference across R3 will rise. However the base voltage is held constant by R1 and R2, so the base-emitter voltage V_{BE} will fall. Reference to Fig 3.31 will show that a reduction in V_{BE} has a marked effect in reducing I_C quickly offsetting the assumed rise.

Bias Classes

The choice of base and collector voltages is normally such that an input signal is amplified without the distortion shown in Fig 3.32. However there are occasions where this is not intended. Take, for example, the circuit in **Fig 3.39**. The two output transistors may be biased just into conduction, drawing a modest quiescent current. If the signal voltage at the collector of the input transistor rises, then the V_{BE} on transistor 1 will tend to increase and transistor 1 will conduct, amplifying the signal. The V_{BE} on transistor 2 will reduce and it will not conduct. Conversely, if the voltage at the collector of the input transistor falls then transistor 1 will not conduct and transistor 2 will amplify the signal.

The purpose of this arrangement, know as a push-pull amplifier, is that it allows the quiescent current in the output transistors to be very low. Current only flows when the transistor is actually amplifying its portion of the signal. By this means the heat dissipation is minimised and a much higher output power can be realised than would be achieved normally. This is also desirable in battery-powered devices to prolong battery life.

The 'normal' bias of Fig 3.33 is known as Class A bias, where collector current flows over the entire of the input signal waveform. The push-pull amplifier has Class B bias, where collector current flows over half the input signal waveform. Class AB bias denotes biasing such that collector current flows for more than half but less than the whole cycle of the input signal. Class C bias tells us that the collector current is flowing for less than half the input waveform. This is illustrated in **Fig 3.40**.

The distortion of class B bias is avoided simply by the two transistors sharing the task. In reality there is a small overlap between the two transistors to minimise any cross-over distortion where conduction swaps from one transistor to the other. It could be argued that, in reality, the transistors are biased in class AB, but close to class B. Some schools of thought use the notation AB1 and AB2 to denote closer to class A and closer to class B respectively. The design issue is distortion versus standing current in the transistor. The standing current having implications for heat dissipation/power handling and, especially in battery powered devices, current drain and battery life.

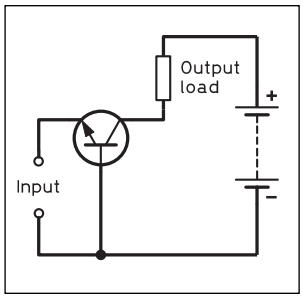


Fig 3.41: The common-base configuration

No such option is available for class C bias and distortion will occur. Class C can only be used at radio frequencies where tuned circuits and filters can be used. It is, however highly efficient, up to 66% or 2/3. That is to say 2/3 of the DC supply is transformed into RF signal.

It will be recalled that any form of waveform or harmonic distortion introduces harmonic frequencies into the signal. Filtering and tuned circuits can remove the harmonics, thereby removing the distortion. This enables a transistor to produce even higher output powers, but, as will be seen in later chapters, the technique is not useable with any form of amplitude modulation of the input signal although a class C RF power amplifier stage can itself be amplitude modulated using collector or anode modulation. This is a technique where the RF input drive is unmodulated and the audio modulating signal is superimposed on the DC supply to the RF power amplifier transistor.

Transistor Configurations

The transistor has been used so far with its emitter connected to the 0V rail and the input to the base. This is not the only way of connecting and using a transistor. The important things to remember are that the transistor is a current driven device, which responds to base current and the potential difference V_{BE}. The output is a variation in collector current, which is normally translated into a voltage by passing that current through a collector resistor. It must also be noted that $I_E=I_B+I_C$ and that $I_C>>I_B$.

The common-emitter configuration, already seen in Fig 3.33, is usually preferred because it provides both current and voltage gain. The input impedance of a common-emitter amplifier using a small, general-purpose transistor will be around 2kΩ at audio frequencies but, assuming that the emitter resistor is bypassed with a capacitor, this will drop to approximately 100Ω at a frequency of a few megahertz. The output impedance will be in the region of 10kΩ. The chapter on Building Blocks provides more information on the design of common-emitter amplifiers.

Common-base configuration

The common-base configuration is shown in **Fig 3.41**. The input and output coupling and bias circuitry has been omitted for the sake of clarity. As the emitter current is the sum of the collector and base currents, the current gain will be very slightly less than unity. There is, however, considerable voltage gain. The input impedance is very low, typically between 10 and 20Ω, assuming a collector current of 2mA. The output impedance is much high-

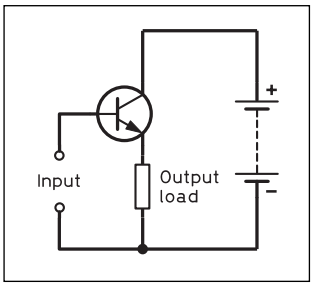


FIg 3.42: The common-collector, or emitter-follower, configuration

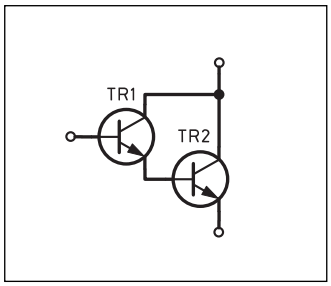


Fig 3.43: A Darlington pair is the equivalent of a single transistor with extremely high current gain

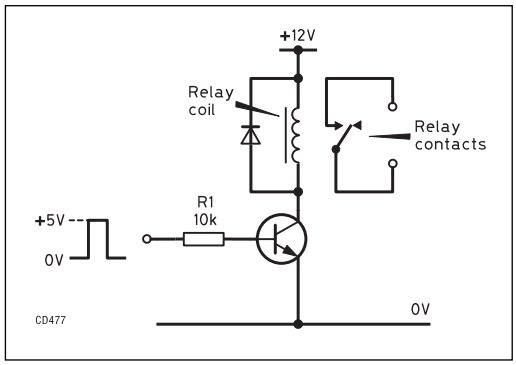


Fig 3.44: A transistor used to switch a relay

er, perhaps 1MΩ at audio frequencies. The output voltage will be in phase with the input. The common-base configuration will sometimes be used to provide amplification where the transistor must operate at a frequency close to its f_T.

Common-collector configuration

The common-collector configuration is normally referred to as the emitter follower, shown in **Fig 3.42**. This circuit provides a current gain equal to the transistor's ß, but the voltage gain is very slightly less than unity. The input impedance is much higher than for the common-emitter configuration and may be approximated by multiplying the transistor's ß by the value of the emitter load impedance. The output impedance, which is much lower, is normally calculated with reference to the impedance of the circuitry which drives the follower, and is approximated by:

$$\text{Output impedance} = \frac{Z_{out}}{\beta}$$

where Z_{out} is the output impedance of the circuit driving the emitter follower.

For example, if the transistor used has a ß of 100 at the frequency of operation and the emitter follower is driven by circuitry with an output impedance (Z_0) of 2kΩ, the output impedance of the follower is roughly 20Ω. As with the common-base circuit, the output voltage is in phase with the input. Emitter followers are used extensively as 'buffers' in order to obtain impedance transformation, and isolation, between stages (see the later chapters).

The Darlington pair

Fig 3.43 shows how two transistors may be connected to produce the equivalent of a single transistor with extremely high current gain (ß). A current flowing into the base of TR1 will cause a much larger current to flow into the base of TR2. TR2 then provides further current gain. Not surprisingly, the overall current gain of the Darlington pair is calculated by multiplying the ß of TR1 by the ß of TR2. Therefore, if each transistor has a ß of 100, the resultant current gain is 10,000.

As an alternative to physically connecting two separate devices, it is possible to obtain 'Darlington transistors'. These contain a pair of transistors, plus the appropriate interconnections, fabricated onto a single chip.

The transistor as a switch

There is often a requirement in electronic equipment to activate relays, solenoids, electric motors and indicator devices etc using control signals generated by circuitry that cannot directly power the device which must be turned on or off. The transistor in **Fig 3.44** solves such a problem by providing an interface between the source of a control voltage (+5V in this example) and a 12V relay coil. If the control voltage is absent, only a minute leakage current flows between the collector and emitter of the switching transistor and so the relay is not energised. When the control voltage is applied, the transistor draws base current through R1 and this results in a larger current flow between its collector and emitter. The relay coil is designed to be connected directly across a potential of 12V and so the resistance between the collector and emitter of the transistor must be reduced to the lowest possible value. The desired effect is therefore the same as that which might otherwise be obtained by closing a pair of switch contacts wired in series with the relay coil.

Assuming that R1 allows sufficient base current to flow, the transistor will be switched on to the fullest extent (a state referred to as saturation). Under these conditions, the voltage at the collector of the transistor will drop to only a few tenths of a volt, thus allowing a potential almost equal to that of the supply rail (12V) to appear across the relay coil, which is therefore properly energised. Under these conditions the transistor will not dissipate much power because, although it is passing considerable current, there is hardly any resistance between the emitter and collector. Assuming that the relay coil draws 30mA when energised from a 12V supply, and that the ß of the transistor is 150 at a collector current of this value, the base current will be:

$$\frac{30\times10^{-3}}{150} = 200\mu A$$

Using Ohm's Law, this suggests that the current limiting resistor R1 should have a value of around 25kΩ. In practice, however, a lower value of 10kΩ would probably be chosen in order to make absolutely sure that the transistor is driven into saturation. The diode connected across the relay coil, which is normally reverse biased, protects the transistor from high voltages by absorbing the coil's back EMF on switching off.

Constant-current generator

Fig 3.45 shows a circuit that will sink a fixed, predetermined current into a load of varying resistance. A constant-current battery charger is an example of a practical application which might use such a circuit. Also, certain low-distortion amplifiers will employ constant-current generators, rather than resistors, to act as collector loads. The base of the PNP transistor is held at a potential

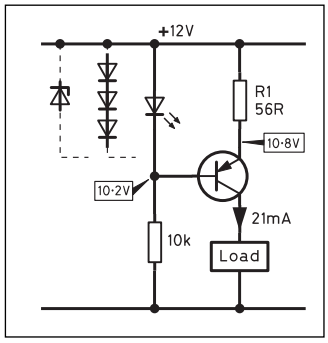


Fig 3.45: The constant-current generator

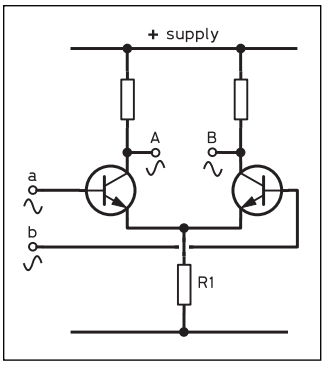


Fig 3.46: A differential amplifier or long-tailed pair

of 10·2V by the forward voltage drop of the LED (12 - 1·8 = 10·2V). Note that although LEDs are normally employed as indicators, they are also sometimes used as reference voltage generators. Allowing for the base-emitter voltage drop of the transistor (approximately 0·6V), the emitter voltage is 10·8V. This means that the potential difference across R1 will be held at 1·2V. Using Ohm's Law, the current flowing through R1 is:

$$\frac{1\cdot 2}{56} = 0\cdot 021\text{A} \quad \text{or } 21\text{mA}$$

The emitter current, and also the collector current, will therefore be 21mA. Should the load attempt to draw a higher current, the voltage across R1 will try and rise. As the base is held at a constant voltage, it is the base-emitter voltage that must drop, which in turn prevents the transistor passing more current.

Also shown in Fig 3.45 are two other ways of generating a reasonably constant reference voltage. A series combination of three forward-biased silicon diodes will provide a voltage drop similar to that obtained from the LED (3 x 0·6 = 1.8V). Alternatively, a zener diode could be used, although as the lowest voltage zener commonly available is 2·7V, the value of R1 must be recalculated to take account of the higher potential difference.

The long-tailed pair

The long-tailed pair, or differential amplifier, employs two identical transistors which share a common emitter resistor. Rather than amplifying the voltage applied to a single input, this circuit provides an output which is proportional to the difference between the voltages presented to its two inputs, labelled 'a' and 'b' in **Fig 3.46**.

Providing that the transistors are well matched, variations in V_{BE} and will cause identical changes in the potential at the two outputs, A and B. Therefore, if both A and B are used, it is the voltage difference between them that constitutes the wanted output. The long-tailed pair is very useful as an amplifier of DC potentials, an application where input and output coupling capacitors cannot be used. The circuit can be improved by replacing the emitter resistor (R1) with a constant-current generator, often referred to in this context as a current source. The differential amplifier is used extensively as an input stage in operational amplifiers.

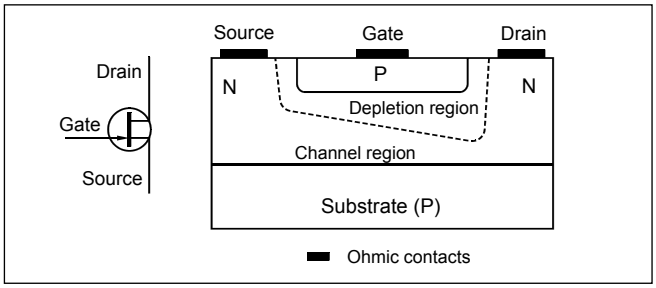


Fig 3.47: The circuit symbol and construction of the JFET

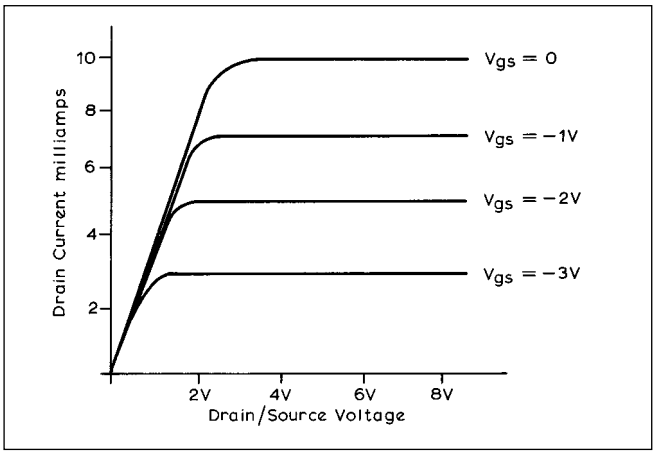


Fig 3.48: The relationship between gate voltage and drain current for a JFET

JUNCTION FIELD EFFECT TRANSISTOR

Like its bipolar counterpart, the junction field effect transistor (JFET) is a three-terminal device which can be used to provide amplification over a wide range of frequencies. **Fig 3.47** shows the circuit symbol and construction of an N-channel JFET. This device differs from the bipolar transistor in that current flow between its drain and source connections is controlled by a voltage applied to the gate. The current flowing to the gate terminal is very small, as it is the reverse bias leakage current of a PN junction. Thus the input (gate) impedance of a JFET is very high, and the concept of current gain, as applied to bipolar devices, is meaningless. The JFET may be referred to as a unipolar device, since the current flow is by majority carriers only, ie electrons in an N-channel transistor. In such a device holes (which have lower mobility than electrons) have no part in the current transport process, thus offering the possibility of very good high-frequency performance.

If both the gate and source are at ground potential (V_{GS}=0), and a positive voltage is applied to the drain (V_{DS}), electrons will flow from source to drain through the N-type channel region (known as the drain-source current, I_{DS}). If V_{DS} is small, the gate junction depletion region width will remain practically independent of V_{DS}, and the channel will act as a resistor. As V_{DS} is increased, the reverse bias of the gate junction near the drain increases, and the average cross-sectional area for current flow is reduced, thus increasing the channel resistance. Eventually V_{DS} is large enough to cause the gate depletion region to expand to fill the channel at the drain end, thus separating the source from the drain. This condition is known as pinch-off. It is important to note that current (I_{DS}) continues to flow across this depletion region, since carriers are injected into it from the channel and accelerated across it by the electric field present. If V_{DS} is increased beyond pinch-off, the edge of the depletion region will move along the channel towards the source. However the voltage drop along the channel between the source and the edge of the depletion region will remain fixed, and thus the drain-source current will remain essentially unchanged. This current saturation is evident from the JFET current-voltage characteristics shown in **Fig 3.48** (V_{GS}=0).

A negative voltage applied to the P-type gate (V_{GS}) establishes a reverse biased depletion region intruding into the N-type channel (see Fig 3.47). Thus for small values of of V_{DS} the channel will again act as a resistor, however its resistance will be larger than when V_{GS}=0 because the cross-sectional area of the channel

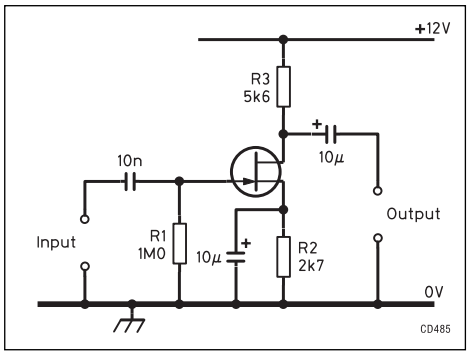


Fig 3.49: A JFET amplifier

has decreased due to the wider depletion region. As V_{DS} is increased the gate depletion region again expands to fill the channel at the drain end, causing pinch-off and current saturation. However, the applied gate voltage reduces the drain voltage required for the onset of pinch-off, thus also reducing the saturation current. This is evident from Fig 3.48 for V_{GS} = -1V, -2V and -3V. Furthermore, Fig 3.48 also suggests that a varying signal voltage applied to the gate will cause proportional variations in drain current. The gain, or transconductance (G_m or Y_{fs}), of a field effect device is expressed in siemens (see Chapter 1). A small general-purpose JFET will have a G_m of around 5 millisiemens.

The circuit of a small-signal amplifier using a JFET is shown in **Fig 3.49**. The potential difference across R2 provides bias by establishing a positive voltage at the source, which has the same effect as making the gate negative with respect to the source. R1 serves to tie the gate at ground potential (0V), and in practice its value will determine the amplifier's input impedance at audio frequencies. The inherently high input impedance of the JFET amplifier is essentially a result of the source-gate junction being reverse biased. If the gate were to be made positive with respect to the source (a condition normally to be avoided), gate current would indeed flow, thus destroying the field effect. The value of R3, the drain load resistor (R_L), dictates the voltage gain obtained for a particular device transconductance (assumed to be 4 millisiemens in this case) as follows:

$$\begin{aligned}\text{Voltage gain} &= G_m \times R_L \\ &= 4\times10^{-3}\times5{\cdot}6\times10^{3} \\ &= 4\times5{\cdot}6 \\ &= 22{\cdot}4 \quad \text{or } 27\text{dB}\end{aligned}$$

The voltage gain obtainable from a common-source JFET amplifier is therefore around 20dB, or a factor of 10, lower than that provided by the equivalent common-emitter bipolar amplifier. Also, the characteristics of general-purpose JFETs are subject to considerable variation, or 'spread', a fact that may cause problems in selecting the correct value of bias resistor for a particular device. However, the JFET does offer the advantage of high input impedance, and this is exploited in the design of stable, variable frequency oscillators. JFETs are also employed in certain types of RF amplifier and switching mixer.

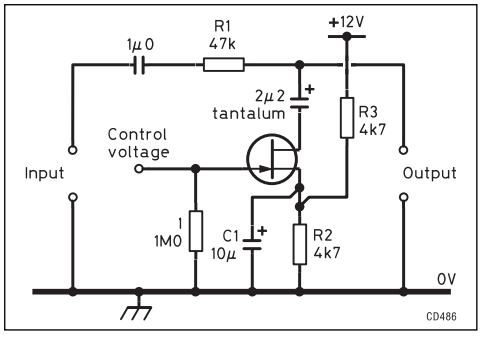


Fig 3.50: The JFET may be used as a voltage-controlled attenuator

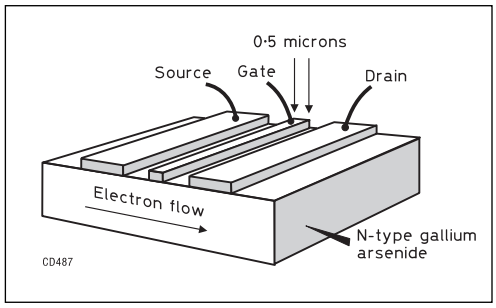


FIg 3.51: Construction of a GaAsFET

The JFET is often used as a voltage-controlled variable resistor in signal gates and attenuators. The channel of the JFET in **Fig 3.50** forms a potential divider working in conjunction with R1. Here R2 and R3 develop a bias voltage which is sufficient to ensure pinch-off. This means that the JFET exhibits a very high resistance between the source and drain and so, providing that the following stage also has a high input impedance, say at least five times greater than the value of R1, the signal will suffer practically no attenuation. Conversely, if a positive voltage is applied to the gate sufficient to overcome the effect of the bias, the channel resistance will drop to the lowest possible value - typically 400Ω for a small-signal JFET. The signal will now be attenuated by a factor nearly equal to the ratio between R1 and the channel resistance - 118, or 41dB (note that C1 serves to bypass R2 at signal frequencies). The circuit is not limited to operation at these two extremes, however, and it is possible to achieve the effect of a variable resistor by adjusting the gate voltage to achieve intermediate values of channel resistance.

GaAsFETs

Although field effect devices are generally fabricated from silicon, it is also possible to use gallium arsenide. GaAsFETs (gallium arsenide field effect transistors) are N-channel field effect transistors designed to exploit the higher electron mobility provided by gallium arsenide (GaAs). The gate terminal differs from that of the standard silicon JFET in that it is made from gold, which is bonded to the top surface of the GaAs channel region (see **Fig 3.51**). The gate is therefore a Schottky barrier junction, as used in the hot-carrier diode. Good high-frequency performance is achieved by minimising the electron transit time between the source and drain. This is achieved by reducing the source drain spacing to around 5 microns and making the gate from a strip of gold only 0·5 microns wide (note that this critical measurement is normally referred to as the gate length because it is the dimension running parallel to the electron flow).

The very small gate is particularly delicate, and it is therefore essential to operate GaAsFETs with sufficient negative bias to ensure that the gate source junction never becomes forward biased. Protection against static discharge and supply line transients is also important.

GaAsFETs are found in very-low-noise receive preamplifiers operating at UHF and microwave frequencies up to around 20GHz. They can also be used in power amplifiers for microwave transmitters.

MOSFETs

The MOSFET (metal oxide field effect transistor), also known as the IGFET (insulated gate field effect transistor), is a very important device, with applications ranging from low-noise preamplification at microwave frequencies to high-power amplifiers in HF and VHF transmitters. Ultra large scale integrated circuits

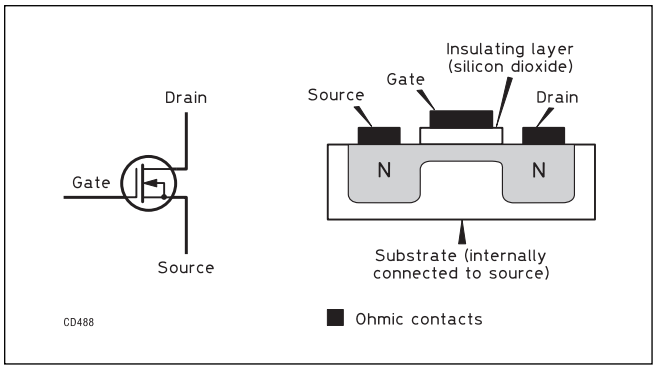


Fig 3.52: Circuit symbol and method of construction of a depletion-mode N-channel MOSFET (metal oxide semiconductor field effect transistor). The circuit symbol for the P-channel type is the same, except that the direction of the arrow is reversed

(ULSICs), including microprocessors and memories, also make extensive use of MOS transistors.

The MOSFET differs from the JFET in having an insulating layer, normally composed of silicon dioxide (SiO_2), interposed between the gate electrode and the silicon channel. The ability to readily grow a high quality insulating material on silicon is a key reason for the dominance of silicon device technology. The source and drain are formed by diffusion into the silicon substrate. This insulation prevents current flowing into, or out of, the gate, which makes the MOSFET easier to bias and guarantees an extremely high input resistance. The insulating layer acts as a dielectric, with the gate electrode forming one plate of a capacitor. Gate capacitance depends on the area of the gate, and its general effect is to lower the impedance seen at the gate as frequency rises. The main disadvantage of this structure is that the very thin insulating layer can be punctured by high voltages appearing on the gate. Therefore, in order to protect these devices against destruction by static discharges, internal zener diodes are normally incorporated. Unfortunately, the protection provided by the zener diodes is not absolute, and all MOS devices should therefore be handled with care.

MOSFETs have either N-type or P-type channel regions, conduction being provided by electrons in N-channel devices, and holes in P-channel devices. However, as electrons have greater mobility than holes, the N-channel device is often favoured because it promises better high-frequency performance.

The current-voltage characteristics of MOSFETs are analagous in form to JFETs. For low values of drain-source voltage (V_{DS}) the channel acts as a resistor, with the source-drain current (I_{DS}) proportional to V_{DS}. As V_{DS} increases, pinch-off eventually occurs, and beyond this point I_{DS} remains essentially constant. The gate-source voltage (V_{GS}) controls current flow in the channel by causing carrier accumulation, carrier depletion or carrier inversion at the silicon channel surface. In accumulation, V_{GS} causes an enhanced concentration of majority carriers (eg electrons in N-type silicon), and in depletion the majority carrier concentration is reduced. In inversion the applied gate voltage is sufficient to cause the number of minority carriers at the silicon surface (ie holes in N-type silicon) to exceed the number of majority carriers. This means that, for example, an N-type silicon surface becomes effectively P-type. The value of V_{GS} required to invert the silicon surface is known as the threshold voltage.

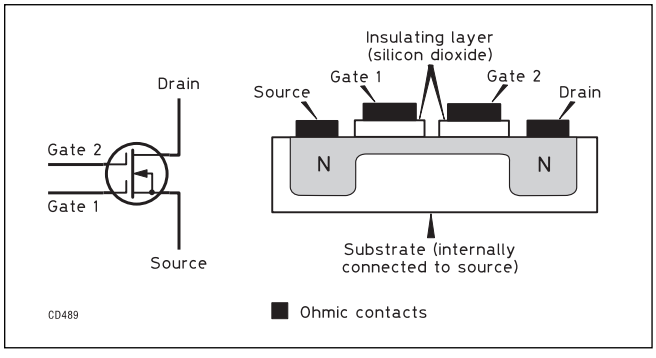


Fig 3.53: A depletion-mode dual-gate MOSFET

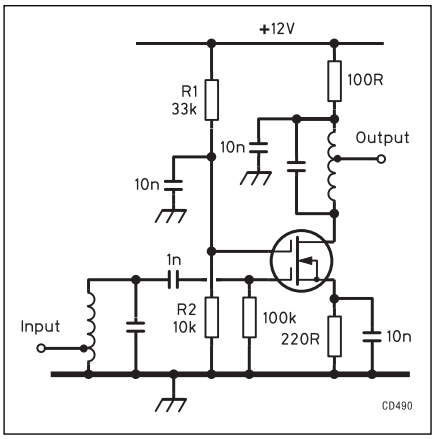


Fig 3.54: An RF amplifier using a dual-gate MOSFET

There are basically two types of N-channel MOSFETs. If at V_{GS}=0 the channel resistance is very high, and a positive gate threshold voltage is required to form the N-channel (thus turning the transistor on), then the device is an enhancement-mode (normally off) MOSFET. If an N-channel exists at V_{GS}=0, and a negative gate threshold voltage is required to invert the channel surface in order to turn the transistor off, then the device is a depletion-mode (normally on) MOSFET. Similarly there are both enhancement and depletion-mode P-channel devices. Adding further to the variety of MOSFETs available, there are also dual-gate types.

Fig 3.52 shows the circuit symbol and construction for a single-gate MOSFET, and **Fig 3.53** features the dual-gate equivalent. Dual-gate MOSFETs perform well as RF and IF amplifiers in receivers. They contribute little noise and provide good dynamic range. Transconductance is also higher than that offered by the JFET, typically between 7 and 15 millisiemens for a general-purpose device.

Fig 3.54 shows the circuit of an RF amplifier using a dual-gate MOSFET. The signal is presented to gate 1, and bias is applied separately to gate 2 by the potential divider comprising R1 and R2. Selectivity is provided by the tuned circuits at the input and

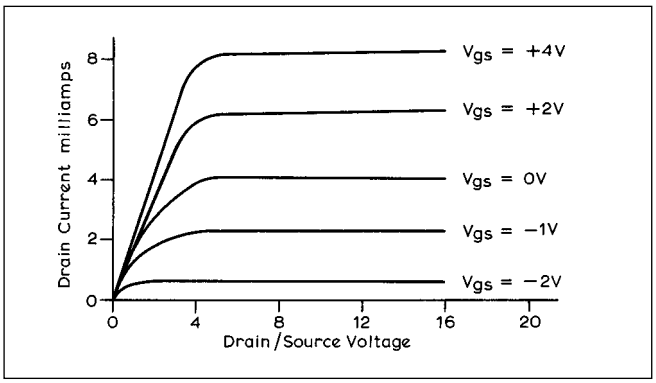


Fig 3.55: The relationship between gate voltage and drain current for an N-channel depletion mode MOSFET

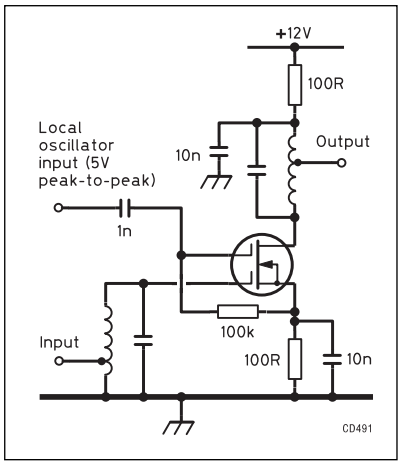


Fig 3.56: A mixer circuit based on a dual-gate MOSFET

output. Care must be taken in the layout of such circuits to prevent instability and oscillation. A useful feature of the dual-gate amplifier is the ability to control its gain by varying the level of the gate 2 bias voltage. This is particularly useful in IF amplifiers, where the AGC voltage is often applied to gate 2. The characteristic curve at **Fig 3.55** shows the effect on drain current of making the bias voltage either negative or positive with respect to the source.

Dual-gate MOSFETs may also be used as mixers. In **Fig 3.56**, the signal is applied to gate 1 and the local oscillator (LO) drive to gate 2. There is a useful degree of isolation between the two gates, and this helps reduce the level of oscillator voltage fed back to the mixer input. For best performance, the LO voltage must be sufficient to turn the MOSFET completely off and on, so that the mixer operates in switching mode. This requires an LO drive of around 5V peak to peak. However, as the gate impedance is high, very little power is required.

VMOS Transistors

As already discussed, MOSFETs exhibit very low gate leakage current, which means that they do not require complex input drive circuitry compared with bipolar devices. In addition, unipolar MOSFETs have a faster switching speed than bipolar transistors. These features make the MOSFET an attractive candidate for power device applications. The VMOS™ (vertical metal oxide semiconductor), also known as the power MOSFET, is constructed in such a way that current flows vertically between the drain, which forms the bottom of the device, to a source terminal at the top (see **Fig 3.57**). The gate occupies either a V- or U-shaped groove etched into the upper surface. VMOS devices feature a four-layer sandwich comprising N+, P, N- and N+ material and

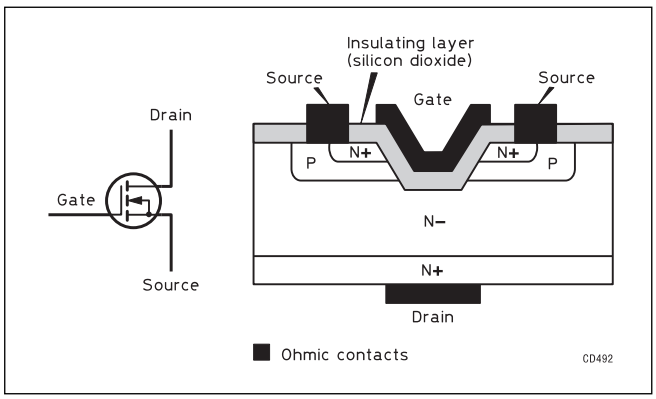


Fig 3.57: Circuit symbol and method of construction of a VMOS transistor

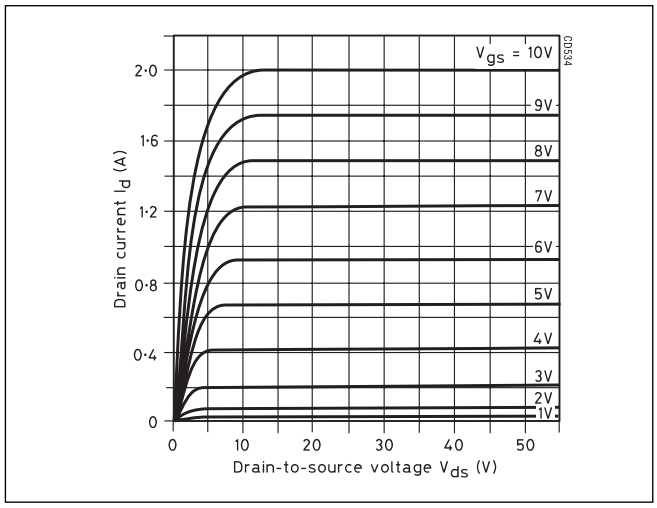


Fig 3.58: Relationship between gate voltage and drain current for a VMOS power transistor

operate in the enhancement mode. The vertical construction produces a rugged device capable of passing considerable drain current and offering a very high switching speed. These qualities are exploited in power control circuits and transmitter output stages. **Fig 3.58** shows the characteristic curves of a typical VMOS transistor. Note that the drain current is controlled almost entirely by the gate voltage, irrespective of drain voltage. Also, above a certain value of gate voltage, the relationship between gate voltage and drain current is highly linear. Power MOSFETs fabricated in the form of large numbers of parallel-connected VMOS transistors are termed HEXFETs™.

Although the resistance of the insulated gate is for all intents and purposes infinite, the large gate area leads to high capacitance. A VMOS transistor intended for RF and high-speed switching use will have a gate capacitance of around 50pF, whereas devices made primarily for audio applications have gate capacitances as high as 1nF. A useful feature of these devices is that the relationship between gate voltage and drain current has a negative temperature coefficient of approximately 0·7% per degree Celsius. This means that as the transistor gets hotter, its drain current will tend to fall, thus preventing the thermal runaway which can destroy bipolar power transistors.

Fig 3.59 shows the circuit of a simple HF linear amplifier using a single VMOS transistor. Forward bias is provided by the poten-

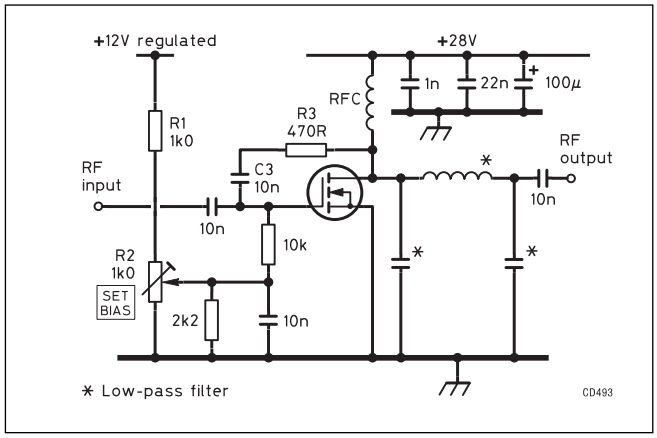


Fig 3.59: A linear amplifier utilising a VMOS power transistor. Assuming a 50-ohm output load, the RFC value is chosen so that it has an inductive reactance (X_L) of approximately 400Ω at the operating requency

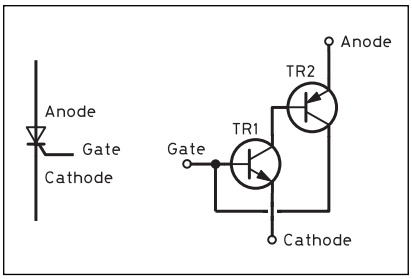


Fig 3.60: The symbol for a thyristor (silicon controlled rectfier) and its equivalent circuit

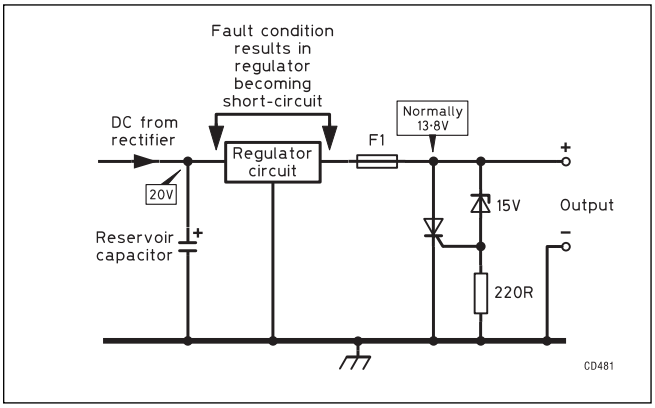


Fig 3.61: Crowbar over-voltage protection implemented wth a thyristor

tial divider R1, R2 so that the amplifier operates in Class AB. In this circuit R3 and C3 provide a small amount of negative feedback to help prevent instability. More complex push-pull amplifiers operating from supplies of around 50V can provide RF outputs in excess of 100W.

THYRISTORS

The thyristor, or silicon controlled rectifier (SCR), is a four-layer PNPN device which has applications in power control and power supply protection systems. The thyristor symbol and its equivalent circuit is shown in **Fig 3.60**. The equivalent circuit consists of two interconnected high-voltage transistors, one NPN and the other PNP. Current flow between the anode and cathode is initiated by applying a positive pulse to the gate terminal, which causes TR1 to conduct. TR2 will now also be switched on because its base is forward biased via TR1's collector. TR1 continues to conduct after the end of the trigger pulse because collector current from TR2 is available to keep its base-emitter junction forward biased. Both transistors have therefore been latched into saturation and will remain so until the voltage between the anode and cathode terminals is reduced to a low value.

Fig 3.61 shows an over-voltage protection circuit for a power supply unit (PSU) based on a thyristor. In the event of regulator failure, the PSU output voltage rises above the nominal 13·8V, a situation that could result in considerable damage to any equipment that is connected to the PSU. As this happens, the 15V zener diode starts to conduct, and in doing so applies a positive potential to the thyristor gate. Within a few microseconds the thyristor is latched on, and the PSU output is effectively short-circuited. This shorting, or crowbar action, will blow fuse F1 and, hopefully, prevent any further harm.

The thyristor will only conduct in one direction, but there is a related device, called the triac (see **Fig 3.62** for symbol), which effectively consists of two parallel thyristors connected anode to cathode. The triac will therefore switch currents in either direction and is used extensively in AC power control circuits, such as the ubiquitous lamp dimmer. In these applications a trigger circuit varies the proportion of each mains cycle for which the triac conducts, thus controlling the average power supplied to a load. Having been latched on at a predetermined point during the AC cycle, the triac will switch off at the next zero crossing point of the waveform, the process being repeated for each following half cycle.

INTEGRATED CIRCUITS

Having developed the techniques used in the fabrication of individual semiconductor devices, the next obvious step for the electronics industry was to work towards the manufacture of complete integrated circuits (ICs) on single chips of silicon. Integrated circuits contain both active devices (eg transistors) and passive devices (eg resistors) formed on and within a single semiconductor substrate, and interconnected by a metallisation pattern.

ICs offer significant advantages over discrete device circuits, principally the ability to densely pack enormous numbers of devices on a single silicon chip, thus achieving previously unattainable functionality at low processing cost per chip. Another advantage of integrated, or 'monolithic', construction is that because all the components are fabricated under exactly the same conditions, the operational characteristics of the transistors and diodes, and also the values of resistors, are inherently well matched. The first rudimentary hybrid IC was made in 1958 by Jack Kilby of Texas Instruments, just eight years after the birth of the bipolar junction transistor. Since then, advances in technology have resulted in a dramatic reduction in the minimum device dimension which may be achieved, and today MOSFET gate lengths of only 50nm (5 x 10^{-6} cm) are possible. This phenomenal rate of progress is set to continue, with 10nm gate lengths expected within the next decade.

The earliest ICs could only contain less than 50 components, but today it is possible to mass-produce ICs containing billions of components on a single chip. For example, a 32-bit microprocessor chip may contain over 42 million components, and a 1Gbit dynamic random access memory (DRAM) chip may contain over 2 billion components.

ICs fall into two broad categories - analogue and digital. Analogue ICs contain circuitry which responds to finite changes in the magnitude of voltages and currents. The most obvious example of an analogue function is amplification. Indeed, virtually all analogue ICs, no matter what their specific purpose may be, contain an amplifier. Conversely, digital ICs respond to only two voltage levels, or states. Transistors within the IC are normally switched either fully on, or fully off. The two states will typically represent the ones and zeros of binary numbers, and the circuitry performs logical and counting functions.

The main silicon technologies used to build these ICs are bipolar (NPN), N-channel MOSFET, and complementary MOS (CMOS) transistors (incorporating P-channel and N-channel MOSFET pairs). While silicon is by far the dominant material for IC production, ICs based around gallium arsenide MESFETs have also been developed for very high frequency applications.

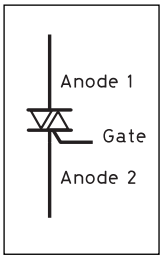


Fig 3.62: The triac or AC thyristor

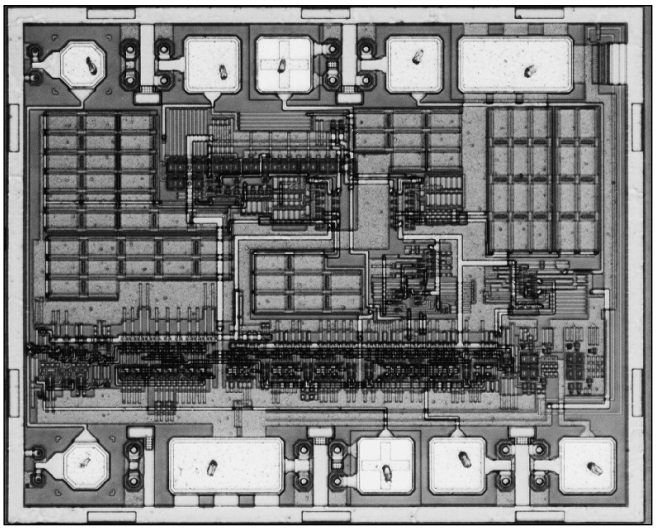


FIg 3.63: Photograph of the SP8714 integrated circuit chip showing pads for bonding wires (GEC-Plessey Semiconductors)

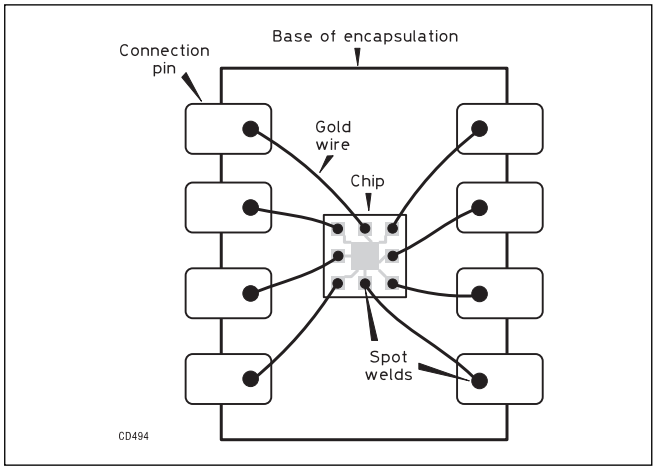


Fig 3.64: Internal construction of an integrated circuit

ICs are produced using a variety of layer growth/deposition, photolithography, etching and doping processes, all carried out under dust-free cleanroom conditions. All chemicals and gases used during processing are purified to remove particulates and ionic contamination. The starting point is a semiconductor wafer, up to 300mm in diameter, which will contain many individual chips. The patterns of conducting tracks and semiconductor junctions are defined by high resolution photolithography and etching. Doping is carried out by diffusion or ion implantation. These processes are repeated a number of times in order to fabricate different patterned layers. Strict process control at each stage minimises yield losses, and ensures that the completed wafer contains a large number of correctly functioning circuits. The wafer is then cut into individual chips and automatically tested to ensure compliance with the design specification. Each selected chip is then fixed to the base of its encapsulation, and very fine gold wires are spot welded between pads located around the chip's periphery and the metal pins which serve as external connections - see **Figs 3.63 and 3.64**. Finally, the top of the encapsulation is bonded to the base, forming a protective seal. Most general-purpose ICs are encapsulated in plastic, but some expensive devices that must operate reliably at high temperatures are housed in ceramic packages.

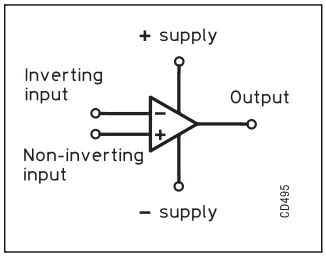


Fig 3.65: Circuit symbol for an operational amplifier (op-amp)

LINEAR INTEGRATED CIRCUITS

Operational Amplifiers

The operational amplifier (op-amp) is a basic building-block IC that can be used in a very wide range of applications requiring low-distortion amplification and buffering. Modern op-amps feature high input impedance, an open-loop gain of at least 90dB (which means that in the absence of gain-reducing negative feedback, the change in output voltage will be at least 30,000 times greater than the change in input voltage), extremely low distortion, and low output impedance. Often two, or even four, separate op-amps will be provided in a single encapsulation. The first operational amplifiers were developed for use in analogue computers and were so named because, with suitable feedback, they can perform mathematical 'operations' such as adding, subtracting, logging, antilogging, differentiating and integrating voltages.

A typical op-amp contains around 20 transistors, a few diodes and perhaps a dozen resistors. The first stage is normally a long-tailed pair and provides two input connections, designated inverting and non-inverting (see the circuit symbol at **Fig 3.65**). The input transistors may be bipolar types, but JFETs or even MOSFETs are also used in some designs in order to obtain very high input impedance. Most op-amps feature a push-pull output stage operating in Class AB which is invariably provided with protection circuitry to guard against short-circuits. The minimum value of output load when operating at maximum supply voltage is normally around 2kΩ, but op-amps capable of driving 500Ω loads are available. Between the input and output circuits there will be one or two stages of voltage amplification. Constant-current generators are used extensively in place of collector load resistors, and also to stabilise the emitter (or source) current of the input long-tailed pair.

In order to obtain an output voltage which is in phase with the input, the non-inverting amplifier circuit shown at **Fig 3.66** is used. Resistors R1 and R2 form a potential divider which feeds a proportion of the output voltage back to the inverting input. In most cases the open-loop gain of an op-amp can be considered infinite. Making this assumption simplifies the calculation of the closed-loop gain obtained in the presence of the negative feedback provided by R1 and R2. For example, if R1 has a value of 9k and R2 is 1k , the voltage gain will be:

$$\frac{R_1 + R_2}{R_2} = \frac{9000 + 1000}{1000} = 10 \text{ or } 20\text{dB}$$

If R2 is omitted, and R1 replaced by a direct connection between the output and the inverting input, the op-amp will function as a unity gain buffer.

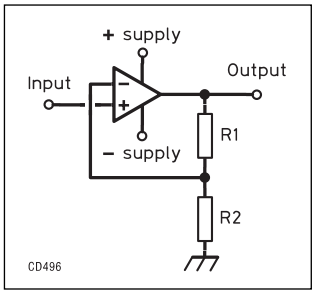


Fig 3.66: A non-inverting amplifier

Fig 3.63 shows an inverting amplifier. In this case the phase of the output voltage will be opposite to that of the input. The closed-loop gain is calculated by simply dividing the value of R1 by R2. Therefore, assum-

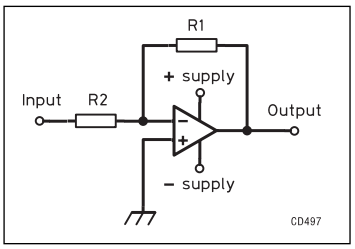


Fig 3.67: An inverting amplifier

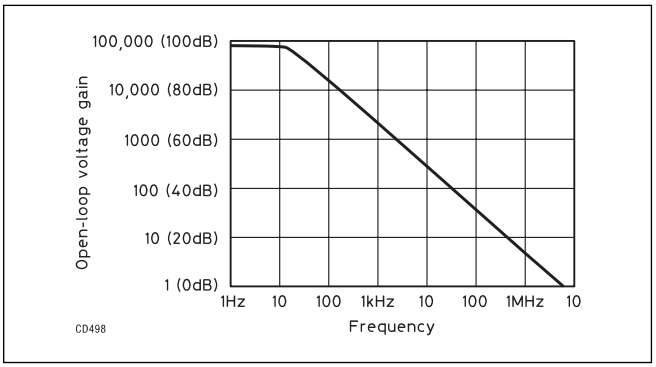


Fig 3.68: The relationship between open loop gain and frequency for a typical internally compensated op-amp

ing the values are the same as used for the non-inverting amplifier:

$$\frac{R_1}{R_2} = \frac{9000}{1000} = 9 \text{ or } 19\text{dB}$$

At low frequencies, the operation of the negative feedback networks shown in **Figs 3.66 and 3.67** is predictable in that the proportion of the output voltage they feed back to the input will be out of phase with the input voltage at the point where the two voltages are combined. However, as the frequency rises, the time taken for signals to travel through the op-amp becomes a significant factor. This delay will introduce a phase shift that increases with rising frequency. Therefore, above a certain frequency - to be precise, where the delay contributes a phase shift of more than 135 degrees - the negative feedback network actually becomes a positive feedback network, and the op-amp will be turned into an oscillator. However, if steps are taken to reduce the open-loop gain of the op-amp to below unity (ie less than 1) at this frequency, oscillation cannot occur. For this reason, most modern op-amps are provided with an internal capacitor which is connected so that it functions as a simple low-pass filter. This measure serves to reduce the open-loop gain of the amplifier by a factor of 6dB per octave (ie for each doubling of frequency the voltage gain drops by a factor of two), and ensures that it falls to unity at a frequency below that at which oscillation might otherwise occur. Op-amps containing such a capacitor are designated as being 'internally compensated'.

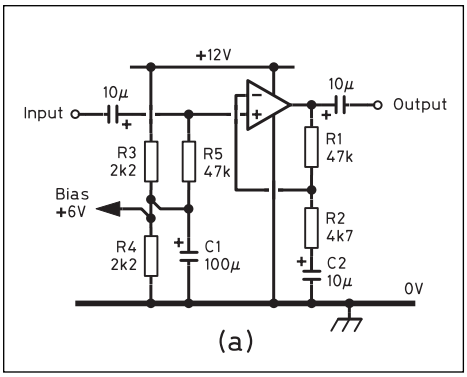


Fig 3.65: Single supply rail versions of the non-inverting (a) and inverting (b) amplifier circuits

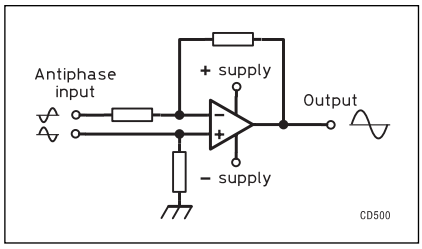


Fig 3.70: A differential amplifier

Compensated op-amps have the advantage of being absolutely stable in normal use. The disadvantage is that the open-loop gain is considerably reduced at high frequencies, as shown by the graph at **Fig 3.68**. This means that general-purpose, internally compensated op-amps are limited to use at frequencies below about 1MHz and they will typically be employed as audio frequency preamplifiers, and in audio filters. There are, however, special high-frequency types available, usually featuring external compensation. An externally compensated op-amp has no internal capacitor, but connections are provided so that the user may add an 'outboard' capacitor of the optimum value for a particular application, thus maximising the gain at high frequencies.

In Figs 3.66 and 3.67 the op-amps are powered from dual-rail supplies. However, in amateur equipment only a single supply rail of around +12V is normally available. Op-amps are quite capable of being operated from such a supply, and **Fig 3.69** shows single-rail versions of the non-inverting and inverting amplifiers with resistor values calculated for slightly different gains. A mid-rail bias supply is generated using a potential divider (R3 and R4 in each case). The decoupling capacitor C1 enables the bias supply to be used for a number of op-amps (in simpler circuits it is often acceptable to bias the op-amp using a potential divider connected directly to the non-inverting input of the op-amp, in which case C1 is omitted). The value of R5 is normally made the same as R1 in order to minimise the op-amp input current, although this is less important in the case of JFET op-amps due to their exceptionally high input resistance. C2 is incorporated to reduce the gain to unity at DC so that the output voltage will settle to the half-rail potential provided by the bias generator in the absence of signals.

Fig 3.70 shows a differential amplifier where the signal drives both inputs in antiphase. An advantage of this balanced arrangement is that interference, including mains hum or ripple, tends to impress voltages of the same phase at each input and so will be eliminated by cancellation. This ability to reject in-phase, or common-mode, signals when operating differentially is known as the common-mode rejection ratio (CMRR). Even an inexpensive op-amp will provide a CMRR of around 90dB, which means that an in-phase signal would have to be 30,000 times greater in magnitude than a differential signal in order to generate the same output voltage.

Further information on op-amps and their circuits is given in the Building Blocks chapter.

Audio Power Amplifiers

The audio power IC is basically just an op-amp with larger output transistors. Devices giving power outputs in the range 250mW to 40W are available, the bigger types being housed in encapsulations featuring metal mounting tabs (TO220 for example) that enable the IC to be bolted directly to a heatsink. **Fig 3.71** shows a 1W audio output stage based on an LM380N device.

The LM380N has internal negative feedback resistors which provide a fixed closed loop voltage gain of approximately 30dB. A bias network for single-rail operation is also provided, and so

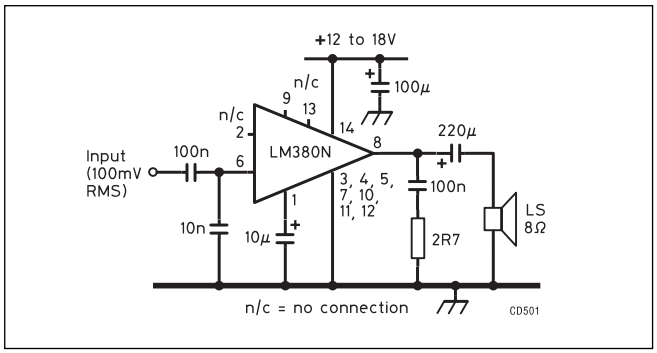


Fig 3.71: An audio output stage using the LM380N IC. The series combination of the 2.7Ω resistor and the 100nF capacitor at the output form a Zobel network. This improves stability by compensating for the inductive reactance of the loudspeaker voice coil at high frequencies

very few external components are required. Not all audio power ICs incorporate the negative feedback and single-rail bias networks 'on-chip', however, and so some, or all, of these components may have to be added externally.

Voltage Regulators

These devices incorporate a voltage reference generator, error amplifier and series pass transistor. Output short-circuit protection and thermal shut-down circuitry are also normally provided. There are two main types of regulator IC - those that generate a fixed output voltage, and also variable types which enable a potentiometer, or a combination of fixed resistors, to be connected externally in order to set the output voltage as required. Devices capable of delivering maximum currents of between 100mA and 5A are readily available, and fixed types offering a wide variety of both negative and positive output voltages may be obtained. **Fig 3.72** shows a simple mains power supply unit (PSU), based on a type 7812 regulator, providing +12V at 1A maximum.

Switched-mode power regulator ICs, which dissipate far less power, are also available. Further information on regulator ICs is given in the chapter on power supplies.

RF Building Blocks

Although it is now possible to fabricate an entire broadcast receiver on a single chip, this level of integration is rarely possible in amateur and professional communications equipment. To achieve the level of performance demanded, and also provide a high degree of operational flexibility, it is invariably necessary to consider each section of a receiver or transceiver separately, and then apply the appropriate technology to achieve the design goals. In order to facilitate this approach, ICs have been developed which perform specific circuit functions such as mixing, RF amplification and IF amplification.

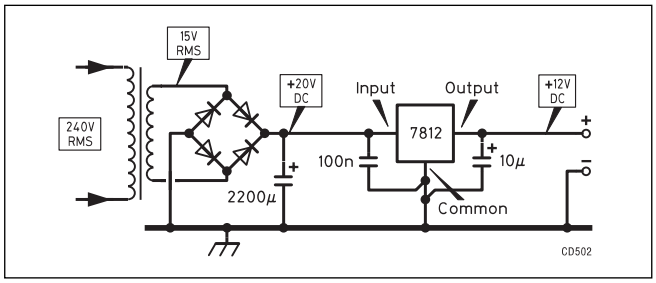


Fig 3.72: A simple 12V power supply unit using a fixed voltage IC regulator type 7812

An example of an IC mixer based on bipolar transistors is the Philips/Signetics NE602AN (featuring very low current consumption as demanded in battery-operated equipment). A useful wide-band RF amplifier is the NE5209D, and there is a similar dual version (ie containing two RF amplifiers within the same encapsulation), the NE5200D. There are also devices offering a higher level of integration - often termed sub-system ICs - which provide more than one block function. For more information on how to use such ICs, see chapters on receivers and transmitters. Amplification at UHF and microwave frequencies calls for special techniques and components. A wide range of devices, known generically as MMICs (microwave monolithic integrated circuits) are available, see the microwaves chapter.

DIGITAL INTEGRATED CIRCUITS

Logic Families

In digital engineering, the term 'logic' generally refers to a class of circuits that perform relatively straightforward gating, latching and counting functions. Historically, logic ICs were developed as a replacement for computer circuitry based on large numbers of individual transistors, and, before the development of the bipolar transistor, valves were employed. Today, very complex ICs are available, such as the microprocessor, which contain most of the circuitry required for a complete computer fabricated on to a single chip. It would be wrong to assume, however, that logic ICs are now obsolescent because there are still a great many low-level functions, many of them associated with microprocessor systems, where they are useful. In amateur radio, there are also applications which do not require the processing capability of a digital computer, but nevertheless depend upon logic - the electronic Morse keyer and certain transmit/receive changeover arrangements, for example.

By far the most successful logic family is TTL (transistor, transistor logic) although CMOS devices have been replacing them for some while. Originally developed in the 'sixties, these circuits have been continuously developed in order to provide more complex functions, increase speed of operation and reduce power consumption. Standard (type 7400) TTL requires a 5V power rail stabilised to within ±250mV. Logic level 0 is defined as a voltage between zero and 800mV, whereas logic 1 is defined as 2·4V or higher. **Fig 3.73** shows the circuit of a TTL NAND gate ('NAND' is an abbreviation for 'NOT AND' and refers to the gate's function, which is to produce an output of logic 0 when both inputs are at logic 1. Really this is an AND function followed by an inverter or

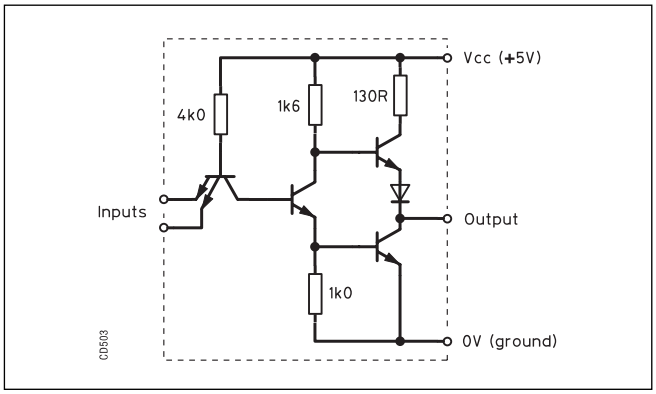


Fig 3.73: The internal circuit of a standard TTL two-input NAND gate. 74LS (low-power Schottky) TTL logic uses higher-value resistors in order to reduce power consumption, and the circuitry is augmented with clamping diodes to increase the switching speed of the transistors

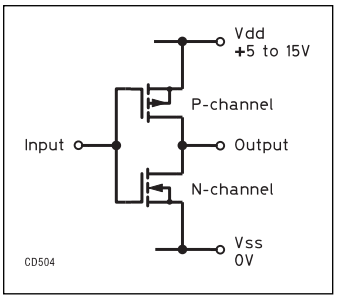


Fig 3.74: Simplified circuit of a CMOS inverter as used in the CD4000 logic family

NOT function. Note the use of a dual-emitter transistor at the input which functions in the same way as two separate transistors connected in parallel.

One of the first improvements made to TTL was the incorporation of Schottky clamping diodes in order to reduce the turn-off time of the transistors - this enhancement produced the 74S series. Latterly, a low-power consumption version of Schottky TTL, known as 74LS, has become very popular. The latest versions of TTL are designed around CMOS (complementary metal oxide silicon) transistors. The 74HC (high-speed CMOS) series is preferred for general use but the 74HCT type must be employed where it is necessary to use a mixture of CMOS and 74LS circuits to implement a design. A 74HC/HCT counter will operate at frequencies up to 25MHz. The 'complementary' in 'CMOS' refers to the use of gates employing a mixture of N-channel and P-channel MOS transistors.

Fig 3.74 shows the simplified circuit of a single CMOS inverter. If the input of the gate is held at a potential close to zero volts (logic 0), the P-channel MOSFET will be turned on, reducing its channel resistance to approximately 400Ω, and the N-channel MOSFET is turned off. This establishes a potential very close to the positive supply rail (V_{dd} - logic 1) at the gate's output. An input of logic 1 will have the opposite effect, the N-channel MOSFET being turned on and the P-channel MOSFET turned off. As one of the transistors is always turned off, and therefore has a channel resistance of about 10,000MΩ, virtually no current flows through the gate under static conditions. However, during transitions from one logic state to another, both transistors will momentarily be turned on at the same time, thus causing a measurable current to flow. The average current consumption will tend to increase as the switching frequency is raised because the gates spend a larger proportion of time in transition between logic levels. The popular CD4000 logic family is based entirely on CMOS technology. These devices can operate with supply voltages from 5 to 15V, and at maximum switching speeds of between 3 and 10MHz.

One of the most useful logic devices is the counter. The simplest form is the binary, or 'divide-by-two' stage shown at **Fig 3.75**. For every two input transitions the counter produces one output transition. Binary counters can be chained together (cascaded) with the output of the first counter connected to the input of the second, and so on. If four such counters are cascaded, there will be one output pulse, or count, for every 16 input pulses. It is also possible to obtain logic circuits which divide by 10 - these are sometimes referred to as BCD (binary coded decimal) counters. Although originally intended to perform the arithmetic function of binary division in computers, the counter can also be used as a frequency divider. For instance, if a single binary counter is driven by a series of pulses which repeat at a frequency of 100kHz, the output frequency will be 50kHz. Counters can therefore be used to generate a range of frequencies that are sub-multiples of a reference input. The crystal calibrator uses this technique, and frequency synthesisers employ a more complex form of counter known as the programmable divider.

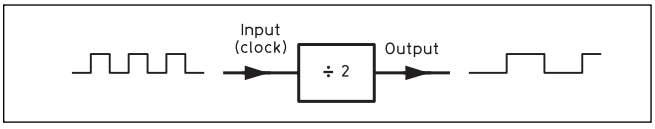


Fig 3.75: A binary counter may be used as a frequency divider

Counters that are required to operate at frequencies above 100MHz use a special type of logic known as ECL (emitter coupled logic). ECL counters achieve their speed by restricting the voltage swing between the levels defined as logic 0 and logic 1 to around 1V. The bipolar transistors used in ECL are therefore never driven into saturation (turned fully on), as this would reduce their switching speed due to charge storage effects. ECL logic is used in frequency synthesisers operating in the VHF to microwave range, and also in frequency counters to perform initial division, or prescaling, of the frequency to be measured.

Memories

There are many applications in radio where it is necessary to store binary data relating to the function of a piece of equipment, or as part of a computer program. These include the spot frequency memory of a synthesised transceiver, for instance, or the memory within a Morse keyer which is used to repeat previously stored messages.

Fig 3.76 provides a diagrammatic overview of the memory IC. Internally, the memory consists of a matrix formed by a number of rows and columns. At each intersection of the matrix is a storage cell which can hold a single binary number (ie either a zero or a one). Access to a particular cell is provided by the memory's address pins. A suitable combination of logic levels, constituting a binary number, presented to the address pins will instruct the IC to connect the addressed cell to the data pin. If the read/write control is set to read, the logic level stored in the cell will appear at the data pin. Conversely, if the memory is set to write, whatever logic level exists on the data pin will be stored in the cell, thus overwriting the previous value. Some memory ICs contain eight separate matrixes, each having their own data pin. This enables a complete binary word (byte) to be stored at each address.

The memory described above is known as a RAM (random access memory) and it is characterised by the fact that data can be retrieved (read) and also stored (written) to individual locations. There are two main types of RAM - SRAM (static RAM) in which data is latched within each memory cell for as long as the power supply remains connected, and DRAM (dynamic RAM)

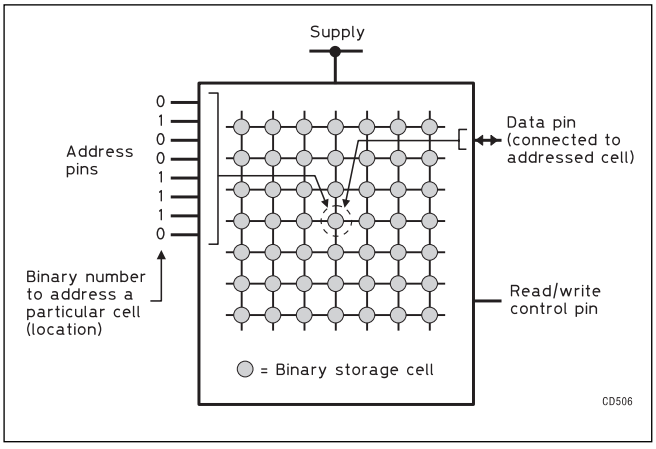


Fig 3.76: Representation of an IC memory

which uses the charge, or the absence of a charge, on a capacitor to store logic levels. The capacitors within a DRAM cannot hold their charge for more than a few milliseconds, and so a process known as refreshing must be carried out by controlling circuitry in order to maintain the stored levels. The method of creating addresses for the memory locations within a DRAM is somewhat complex, in that the rows and columns of the matrix are dealt with separately. DRAM memory chips are used extensively in desktop computers and related equipment because the simple nature of the capacitor memory cell means that a large number of cells can easily be fabricated on a single chip (there are now DRAMs capable of storing 16 million binary digits, or bits). SRAMs fabricated with CMOS transistors are useful for storing data that must be retained while equipment is turned off. The low quiescent current consumption of these devices makes it practicable to power the memory from a small battery located within the equipment, thus providing an uninterrupted source of power.

The ROM (read-only memory) has data permanently written into it, and so there is no read/write pin. ROMs are used to store computer programs and other data that does not need to be changed. The ROM will retain its data indefinitely, irrespective of whether it is connected to a power supply. A special form of ROM known as the EPROM (eraseable programmable ROM) may be written to using a special programmer. Data may later be erased by exposing the chip to ultra-violet light for a prescribed length of time. For this reason, EPROMs have a small quartz window located above the chip which is normally concealed beneath a UV-opaque protective sticker. The EEPROM (electrically eraseable programmable ROM) is similar to the EPROM, but may be erased without using UV light. The PROM (programmable ROM) has memory cells consisting of fusible links. Assuming that logic 0 is represented by the presence of a link, logic 1 may be programmed into a location by feeding a current into a special programming pin which is sufficient to fuse the link at the addressed location. However, once a PROM has been written to in this way, the cells programmed to logic 1 can never be altered.

Analogue-to-digital converters

Analogue-to-digital conversion involves measuring the magnitude of a voltage or current and then generating a numeric value to represent the result. The digital multimeter works in this way, providing an output in decimal format which is presented directly to the human operator via an optical display. The analogue to digital (A/D) converters used in signal processing differ in two important respects. Firstly, the numeric value is generated in binary form so that the result may be manipulated, or 'processed' using digital circuitry. Secondly, in order to 'measure' a signal, as opposed to, say, the voltage of a battery, it is necessary to make many successive conversions so that amplitude changes occurring over time may be captured. In order to digitise speech, for instance, the instantaneous amplitude of the waveform must be ascertained at least 6000 times per second. Each measurement, known as a sample, must then be converted into a separate binary number. The accuracy of the digital representation depends on the number of bits (binary digits) in the numbers - eight digits will give 255 discrete values, whereas 16 bits provides 65,535.

Maximum sampling frequency (ie speed of conversion) and the number of bits used to represent the output are therefore the major parameters to consider when choosing an A/D converter IC. The fastest 8-bit converters available, known as flash types, can operate at sampling frequencies of up to 20MHz, and are used to digitise television signals. 16-bit converters are unfortunately much slower, with maximum sampling rates of around 100kHz. It is also possible to obtain converters offering intermediate levels of precision, such as 10 and 12 bits.

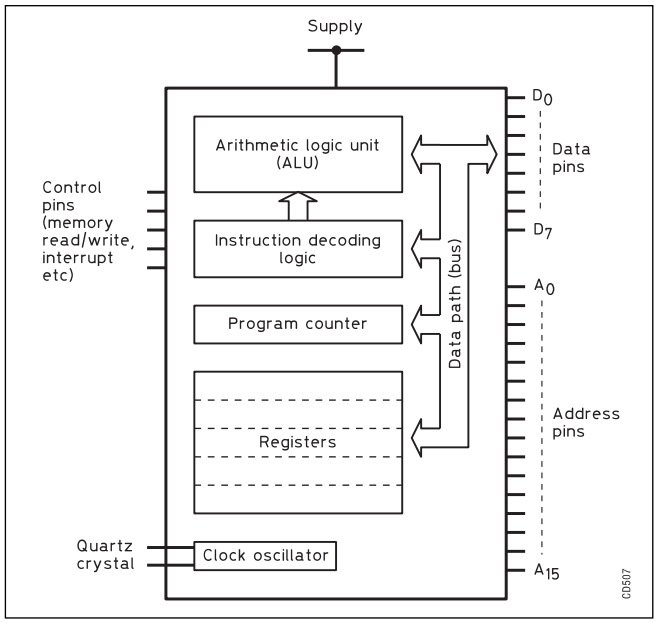


Fig 3.77: An 8-bit microprocessor

Having processed a signal digitally, it is often desirable to convert it back into an analogue form - speech and Morse are obvious examples. There are a variety of techniques which can be used to perform digital-to-analogue conversion, and ICs are available which implement these.

Microprocessors

The microprocessor is different from other digital ICs in that it has no preordained global function. It is, however, capable of performing a variety of relatively straightforward tasks, such as adding two binary numbers together. These tasks are known as instructions, and collectively they constitute the microprocessor's instruction set. In order to make the microprocessor do something useful, it is necessary to list a series of instructions (write a program) and store these as binary codes in a memory IC connected directly to the microprocessor. The microprocessor has both data and address pins (see **Fig 3.77**), and when power is first applied it will generate a pre-determined start address and look for an instruction in the memory location accessed by this. The first instruction in the program will be located at the starting address. Having completed this initial instruction, the microprocessor will fetch the next one, and so on. In order to keep track of the program sequence, and also provide temporary storage for intermediate results of calculations, the microprocessor has a number of internal counters and registers (a register is simply a small amount of memory). The manipulation of binary numbers in order to perform arithmetic calculations is carried out in the arithmetic logic unit (ALU). A clock oscillator, normally crystal controlled, controls the timing of the program-driven events. The microprocessor has a number of control pins, including an interrupt input. This allows normal program execution to be suspended while the microprocessor responds to an external event, such as a keyboard entry.

Microprocessors exist in a bewildering variety of forms. Some deal with data eight bits at a time, which means that if a number is greater than 255 it must be processed using a number of separate instructions. 16-bit and 32-bit and more recently 64-bit microprocessors are therefore generally faster. A special class of microprocessor known as the microcontroller is designed specif-

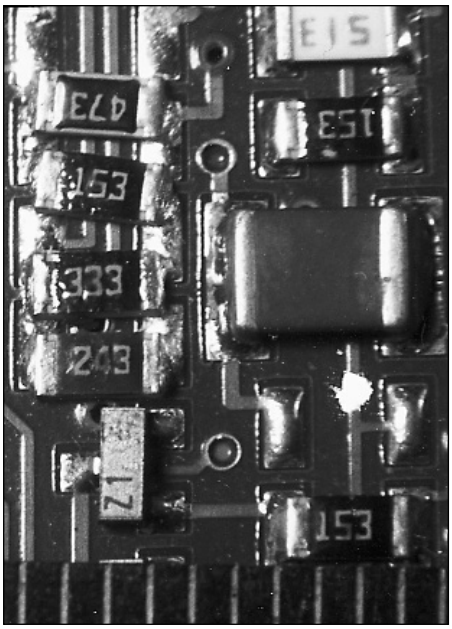


Fig 3.78: Surface mounting components - the light coloured one labelled 'Z1' (bottom left) is a transistor. The scale at the bottom is in millimetres

ically to be built into equipment other than computers. As a result, microcontrollers tend to be more self-sufficient than general microprocessors, and will often be provided with internal ROM, RAM and possibly an A/D converter. Microcontrollers are used extensively in transceivers in order to provide an interface between the frequency synthesiser's programmable divider, the tuning controls - including the memory keypad - and the frequency display.

Following the development of the first microprocessors in the 1970s, manufacturers began to compete with each other by offering devices with increasingly large and ever-more-complex instruction sets. The late 1990s saw something of a backlash against this trend, with the emergence of the RISC (reduced instruction set microprocessor). The rationale behind the RISC architecture is that simpler instructions can be carried out more quickly, and by a processor using a smaller number of transistors.

Computers developed using the 86 series followed briefly by the 168,286 and 386 devices. The 586 series was renamed the Pentium™ and following a ruling that a number could not be trademarked the P1, P2, P3, P4 series had a variety of names for a number of foundries.

Amateurs today make frequent use of the 'PIC' series of microprocessors and these are discussed in more detail elsewhere in the handbook.

DSP (digital signal processing) ICs are special microprocessors which have their instruction sets and internal circuitry optimised for fast execution of the mathematical functions associated with signal processing - in particular the implementation of digital filtering.

SURFACE MOUNT DEVICES

As the semiconductor industry has developed, all packaging has been reduced in size and manufacturing techniques have changed so that devices are mounted on the copper-side surface of the printed circuit board. These SMDs (surface mounting devices) are now the norm and we as amateurs have to live with them. An example is shown in **Fig 3.78** with a section of steel rule with a millimetre scale to show just how small they are. Further details on SMDs and how to handle them are given in the Passive Components, and Construction and Workshop Practice chapters.

ELECTRONIC TUBES AND VALVES

Modern electronic tubes and valves have attained a high degree of reliability and are still available in many forms for a number of the more specialist applications. They have been superseded for virtually all low-power purposes by semiconductors and many of the high power amateur applications. Their use tends to be confined to the highest powers at HF, VHF, UHF and above.

Fundamentals

An electronic valve comprises a number of electrodes in an evacuated glass or ceramic envelope. The current in the valve is simply a large number of electrons emitted from a heated electrode, the cathode, and collected by the anode, which is maintained at a high positive potential. Other electrodes control the characteristics of the device.

Emission

In most types of valve the emission of electrons is produced by heating the cathode, either directly by passing a current through it, or indirectly by using an insulated heater in close proximity. The construction and surface coating of the cathode and the temperature to which it is heated govern the quantity of electrons emitted. This is known as thermionic emission.

Emission may also be produced when electrons impinge on to a surface at a sufficient velocity. For example, electrons emitted from a hot cathode may be accelerated to an anode by the latter's positive potential. If the velocity is high enough, other electrons will be released from the anode. This is known as secondary emission.

The emission of electrons from metals or coated surfaces heated to a certain temperature is a characteristic property of that metal or coated surface. The value of the thermionic emission may be calculated from Richardson's formula:

$$I_S = AT^2 e^{-\frac{b}{T}}$$

where I_S is the emission current in amperes per square centimetre of cathode surface; A and b are constants depending on the material of the emitting surface and T is absolute temperature in Kelvin (K).

Space charge

The thermionic electrons given off by the cathode form a cloud or space charge round the cathode. It tends to repel further electrons leaving the cathode due to its negative charge repelling the negatively charged electrons.

Electron flow

If the anode, see **Fig 3.79** is raised to a positive potential, the electron flow or current will increase to a point where the space charge is completely neutralised and the total emission from the cathode reaches the anode. The flow can be further increased by raising the cathode temperature. **Fig 3.80** shows the rise in anode current up to the saturation point for different heater temperatures.

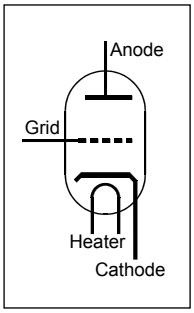


Fig 3.79: A triode valve

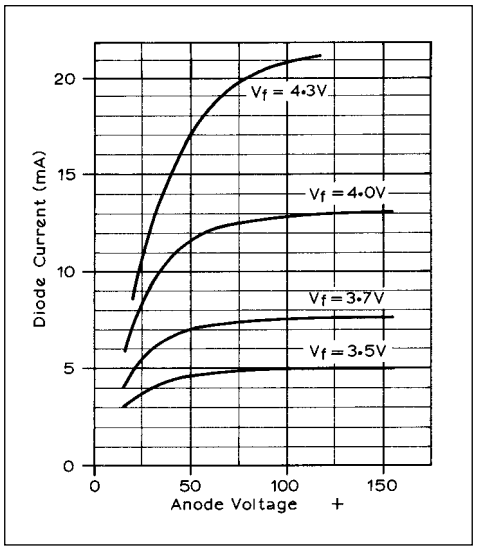


Fig 3.80: Diode saturation curve showing the effect of different filament voltages and hence temperatures

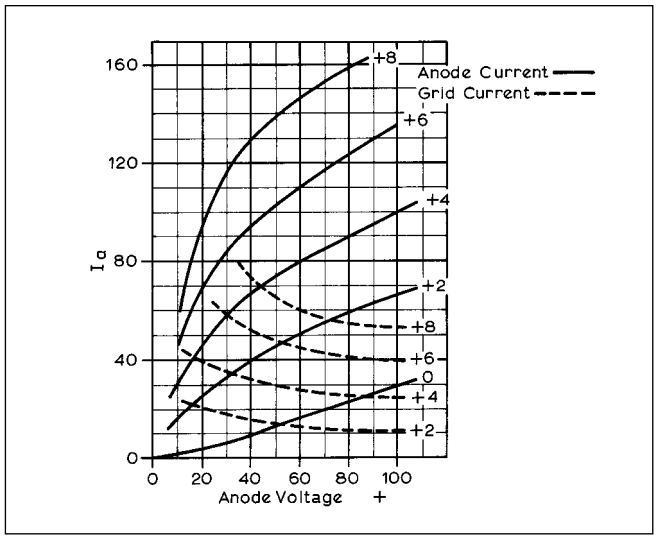


Fig 3.82: Positive-region triode characteristic curves showing how enhanced values of positive grid voltage increase anode current flow

As the electrons traverse the space between one electrode and another they may collide with gas molecules (because no vacuum can be perfect) and such collisions will impede their transit. For this reason the residual gas left inside the evacuated envelope must be minimal. An electronic tube that has been adequately evacuated is termed hard.

However, if a significant amount of gas is present, the collisions between electrons and gas molecules will cause it to ionise. The resultant blue glow between the electrodes indicates that the tube is soft. This blue glow should not be confused with a blue haze which may occur on the inside of the envelope external to the electrode structure: this is caused by bombardment of the glass, and in fact indicates that the tube is very hard.

Diodes

As described so far, we have a diode, a two-electrode valve. Like its semiconductor counterpart it will only permit current flow in one direction, from anode to cathode. The anode cannot emit electrons because it is not heated.

Triodes

By introducing a grid between the cathode and anode of an electronic tube the electron flow may be controlled. This flow may be varied in accordance with the voltage applied to the grid, the required value being determined by the geometry of the grid and the desired valve characteristics. **Fig 3.81** shows the relationship between anode current (I_a) and voltage (V_a) for various negative bias voltages on the grid.

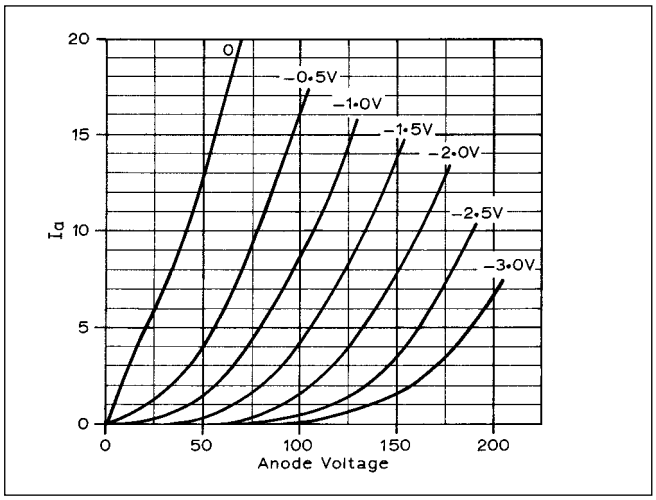


Fig 3.81: Negative-region characteristic curves of a triode showing the reduction in anode current which occurs with increase of negative grid voltage

Normally the grid will operate at a negative potential, limiting the electron flow to the anode. However positive potentials can enhance the anode current but also lead to grid current and possibly more grid dissipation than the designer intended, reducing reliability. Some valves, typically high power devices, are designed to permit grid dissipation and should be used if grid current is expected. **Fig 3.82** shows both the increasing anode current and the grid current for different positive values of grid potential.

The grid voltage (grid bias) for a small general-purpose valve may be obtained by one of several methods:

1. A separate battery - historically common but no longer used.
2. A resistor connected between the cathode and the chassis (earth) so that when current flows the voltage drop across it renders the cathode more positive with respect to the chassis (earth), and the grid circuit return becomes negative with respect to cathode.
3. A resistor connected between the grid and the chassis (earth). When the grid is so driven that appreciable current flows (as in an RF driver, amplifier or multiplier), the grid resistor furnishes a potential difference between grid and chassis (earth), and with cathode connected to chassis a corresponding negative voltage occurs at the grid. A combination of grid resistor and cathode resistor is good practice and provides protection against failure of drive, which a grid resistor alone would not give.

As power amplifiers, triodes have the virtue of simplicity, especially when used in the grounded-grid mode (see later). At VHF and UHF the electrons can have a very short transit time if the valve is built on the planar principle. In this, the cathode, grid and anode are all flat and it is possible to have only a very short distance between them. Also the connection to each electrode has a very low impedance. Both factors promise good operation at HF/VHF/UHF. From this design, it is clear that earthed-grid operation is best when the grid with its disc-shaped connector acts as a screen between the input (cathode) and the output (anode) circuits. (See the section on 'Disc seal valves' later in

this chapter. The 2C39A illustrated there is a good amplifier up to at least 2.3GHz.)

Tetrodes

A tetrode ('four-electrode') valve is basically a triode with an additional grid mounted outside the control grid. That is between the first grid and the anode. When this additional grid is maintained at a steady positive potential a considerable increase in amplification factor occurs compared with the triode state; at the same time the valve impedance is greatly increased.

The reason for this increased amplification lies in the fact that the anode current in the tetrode valve is far less dependent on the anode voltage than it is in the triode. In any amplifier circuit, of course, the voltage on the anode must be expected to vary since the varying anode current produces a varying voltage-drop across the load in the anode circuit. A triode amplifier suffers from the disadvantage that when, for instance, the anode current begins to rise due to a positive half-cycle of grid voltage swing, the anode voltage falls (by an amount equal to the voltage developed across the load) and the effect of the reduction in anode voltage is to diminish the amount by which the anode current would otherwise increase. Conversely, when the grid voltage swings negatively, the anode current falls and the anode voltage rises. Because of this increased anode voltage the anode current is not so low as it would have been if it were independent of anode voltage. This means that the full amplification of the triode cannot be achieved. The introduction of the screen grid, however, almost entirely eliminates the effect of the anode voltage on the anode current, and the amplification obtainable is thus much greater.

A screen functions best when its voltage is below the mean value of the anode voltage. Most of the electrons from the cathode are thereby accelerated towards the anode, but some of them are unavoidably caught by the screen. The resulting screen current serves no useful purpose, and if it becomes excessive it may cause overheating of the screen.

If in low-voltage applications the anode voltage swings down to the screen voltage or lower, the anode current falls rapidly while that of the screen rises due to secondary emission from the anode to the screen. It should be noted that the total cathode current is equal to the sum of the screen and anode currents.

The I_a/V_a characteristics of the tetrode are shown in **Fig 3.83**. It will be noticed that there is a kink in the curve where an increase in anode voltage results in a decrease in anode current. This occurs where the anode voltage is lower than the screen grid but high enough that electrons from the cathode are accelerated fast enough to knock electrons off the anode, which are captured by the higher potential screen grid. As soon as the anode potential is above that of the screen grid, the effect ceases.

Another important effect of introducing the screen grid (G2) is that it considerably reduces the capacitive coupling between the input (control) grid and the anode, making possible the use of stable, high-gain RF amplification. To utilise this facility additional shields are added to the grid (electrically connected) so that the input connection cannot 'see' the anode or its supports. With such a structure it is possible to reduce the unit's capacitance by a factor of almost 1000 compared with the triode. Adequate decoupling of the screen at the operating frequency by the use of a suitable external bypass capacitor is essential.

In another type of tetrode, known as the space-charge grid tetrode, the second grid is positioned between the usual control grid and the cathode. When a positive potential is applied to this space-charge grid, it overcomes the limiting effect of the negative space charge, allowing satisfactory operation to be achieved at very low anode potentials, typically 12-24V.

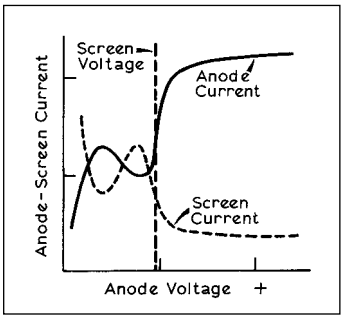


Fig 3.83: Characteristic curves of pure tetrode (often termed screened grid) showing the considerable secondary emission occurring when no suppression is used

Pentodes

To overcome the problem presented by secondary emission in the pure tetrode, a third grid may be introduced between the screen and the anode, and maintained at a low potential or connected to the cathode. Anode secondary emission is overcome and much larger swings of the anode voltage may be realised. This third grid is known as the suppressor grid (G3). Other methods which achieve the same effect are:

- Increasing the space between screen grid and anode;
- Fitting small fins to the inside surface of the anode; or
- Fitting suppressor plates to the cathode to produce what is known as the kinkless tetrode, which is the basis of the beam tetrode suppression system.

In some special types of pentode where it is necessary for application reasons to provide two control grids, the No 3 (sup-

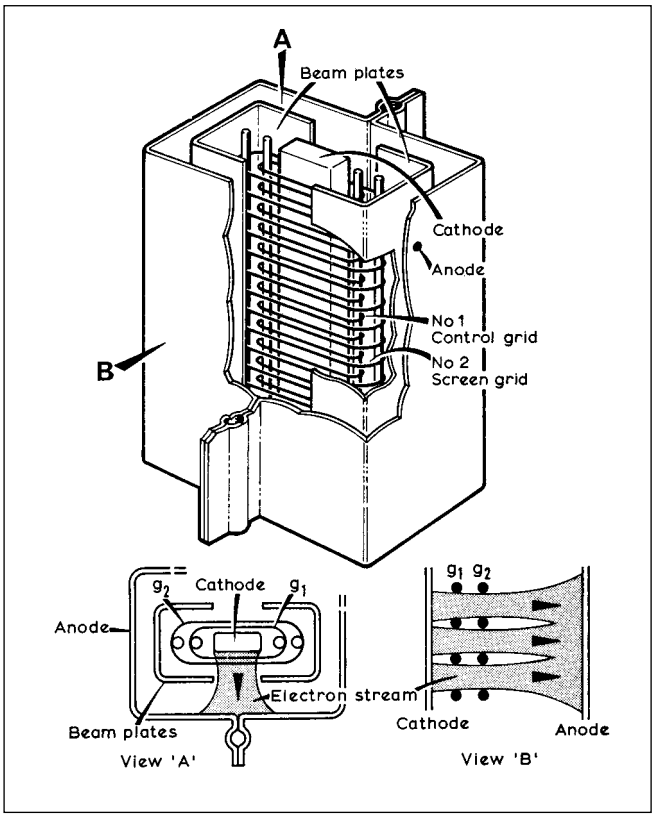


Fig 3.84: The general arrangement of a modern beam tetrode showing the aligned grid winding and the position of the beam forming plates. View 'A': looking vertically into a beam tetrode. View 'B': showing how the aligned electrode structure focuses electrons from cathode to anode

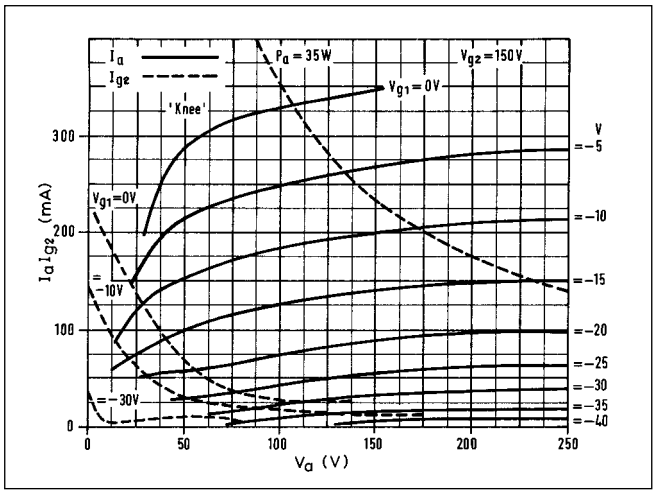


Fig 3.85: Characteristic curves of a beam tetrode. Anode secondary emission is practically eliminated by the shape and position of the suppressor plates

pressor) grid is used as the second and lower-sensitivity control for gating, modulation or mixing purposes. Units of this type need to have a relatively high screen grid (G2) rating to allow for the condition when the anode current is cut off by the suppressor grid (G3).

Beam Tetrodes

A beam tetrode employs principles not found in other types of valve: the electron stream from the cathode is focused (beamed) towards the anode. The control grid and the screen grid are made with the same winding pitch and they are assembled so that the turns in each grid are in optical alignment: see Fig 3.84. The effect of the grid and screen turns being in line is to reduce the screen current compared with a non-beam construction. For example, in a pentode of ordinary construction the screen current is about 20% of the anode current, whereas in a beam valve the figure is 5-10%.

The pair of plates for suppressing secondary emission referred to above is bent round so as to shield the anode from any electrons coming from the regions exposed to the influence of the grid support wires at points where the focusing of the electrons is imperfect. These plates are known as beam-confining or beam-forming plates.

Fig 3.86: 4CX250B showing its air-cooled vanes

Beam valves were originally developed for use as audio-frequency output valves, but the principle has been applied to many types of RF tetrodes, both for receiving and transmitting. Their superiority over pentodes for AF output is due to the fact that the distortion is caused mainly by the second harmonic and only very slightly by the third harmonic, which is the converse of the result obtained with a pentode. Two such valves used in push-pull give a relatively large output with small harmonic distortion because the second harmonic tends to cancel out with push-pull connection.

Fig 3.85 shows the characteristic curves of a beam tetrode. By careful positioning of the beam plates a relatively sharp 'knee' can be produced in the anode current/anode voltage characteristic, at a lower voltage than in the case of a pentode, thus allowing a larger anode voltage swing and greater power output to be achieved. This is a particularly valuable feature where an RF beam tetrode is to be used at relatively low anode voltages.

Valve Construction and Characteristics

Cathodes

Although several types of cathode are used in modern valves, the differences are only in the method of producing thermionic emission. The earliest type is the bright emitter in which a pure tungsten wire is heated to a temperature in the region of 2500-2600K. At such a temperature emission of 4-40mA per watt of heating power may be obtained. These are not normally found in amateur applications.

Oxide-coated cathodes are the most common type of thermionic emitter found in both directly and indirectly heated valves. In this type, the emissive material is usually some form of nickel ribbon, tube or thimble coated with a mixture of barium and strontium carbonate, often with a small percentage of calcium. During manufacture, the coating is reduced to its metallic form and the products of decomposition removed during the evacuation process. The active ingredient is the barium, which provides much greater emission than thoriated tungsten at lower heating powers. Typically, 50-150mA per watt is obtained at temperatures of 950-1050K.

An indirectly heated cathode is a metal tube, sleeve or thimble shape, having a coating of emissive material on the outer surface. The cathode is heated by radiation from a metal filament, called the heater, which is mounted inside the cathode. The heater is electrically insulated from the cathode. The heater is normally made of tungsten or molybdenum-tungsten alloy. The heater or filament voltage should be accurately measured at the valve base and adjusted to the correct value as specified by the makers. This needs great care as it must be done with the stage operating at its rated power.

The life of valves with oxide-coated cathodes is generally good provided the ratings are not exceeded. Occasionally there is some apparent reduction in anode current due to the formation of a resistive layer between the oxide coating and the base metal, which operates as a bias resistor.

Anodes

In most valves the anode takes the form of an open-ended cylinder or box surrounding the other electrodes, and is intended to collect as many as possible of the electrons emitted from the cathode; some electrons will of course be intercepted by the grids interposed between the cathode and the anode.

The material used for the anode of the small general-purpose type of valve is normally bright nickel or some form of metal, coated black to increase its thermal capacity. Power dissipated in the anode is radiated through the glass envelope, a process which is assisted when adequate circulation of air is provided

around the glass surface. In some cases a significant improvement in heat radiation is obtained by attaching to the valve envelope a close-fitting finned metal radiator which is bolted to the equipment chassis so that this functions as a worthwhile heatsink.

Higher-power valves with external anodes are cooled directly by forced air (**Fig 3.86**), by liquid, or by conduction to a heatsink. Forced-air cooling requires a blower, preferably of the turbine type (rather than fan), capable of providing a substantial quantity of air at a pressure high enough to force it through the cooler attached to the anode.

Liquid cooling calls for a suitable cooler jacket to be fitted to the anode; this method is generally confined to large power valves. If water is used as the coolant, care must be taken to ensure that no significant leakage occurs through the water by reason of the high voltage used on the anode. A radiator is then used to cool the water.

In certain UHF disc seal valves a different form of conduction cooling is used, the anode seal being directly attached to an external tuned-line circuit that doubles as the heatsink radiator. Needless to say, it must be suitably isolated electrically from the chassis.

Whatever the type of valve and whatever method is used to cool it, the limiting temperatures quoted by the makers, such as bulb or seal temperatures, should never be exceeded. Underrunning the device in terms of its dissipation will generally greatly extend its life.

Grids

Mechanically, the grid electrode takes many forms, dictated largely by power and the frequency of operation. In small general-purpose valves the grids are usually in the form of a helix (molybdenum or other suitable alloy wire) with two side support rods (copper or nickel) and a cross-section varying from circular to flat rectangular, dependent on cathode shape.

In some UHF valves the grid consists of a single winding of wire or mesh attached to a flat frame fixed directly to a disc seal.

Characteristics

Technical data available from valve and tube manufacturers includes static characteristics and information about typical operating performances obtainable under recommended conditions. Adherence to these recommendations - indeed, to use the valve at lower than the quoted values - will increase life and reliability, which can otherwise easily be jeopardised. In particular, cathodes should always be operated within their rated power recommendations. The following terms customarily occur in manufacturers' data:

Mutual conductance (slope, g_m, transconductance)

This is the ratio of change of anode current to the change of grid voltage at a constant anode voltage. This factor is usually expressed in milliamperes per volt (or sometimes mili-siemens).

Amplification factor (μ)

This is the ratio of change of anode voltage to change of grid voltage for a constant anode current. In the case of triodes classification is customarily in three groups, low μ, where the amplification factor is less than 10, medium μ (10-50) and high μ (greater than 50).

Impedance (r_a, AC resistance, slope resistance)

When the anode voltage is changed while grid voltage remains constant, the anode current will change, an Ohm's Law effect. Consequently, impedance is measured in ohms. The relationship between these three primary characteristics is given by:

$$\text{Impedance}\,(\Omega) = \frac{\text{Amplification factor}}{\text{Mutual conductance}} \times 1000$$

or

$$r_a = \frac{\mu}{g_m} \quad k\Omega$$

where g_m is the mutual conductance in mA/V.

It will be noted that the mutual conductance and the impedance are equal to the slopes of the Ia/Vg and Ia/Va characteristics respectively.

Electrode dissipation

The conversion from anode input power to useful output power will depend upon the tube type and the operating conditions. The difference between these two values, known as the anode dissipation, is radiated as heat. If maximum dissipation is exceeded overheating will cause the release of occluded gas, which will poison the cathode and seriously reduce cathode emission. The input power to be handled by any valve or tube will, in the limiting case, depend on the class of operation. Typical output efficiencies expressed as percentages of the input power are:

Class A	33%
Class AB1	60-65%
Class AB2	60-65%
Class B	65%
Class C	75-80%

Considering a valve with a 10W anode dissipation, the above efficiencies would give outputs as follows (assuming there are no other limiting factors such as peak cathode current):

Class A	5W
Class AB1 and AB2	15-18.5W
Class B	18.5W
Class C	30-40W

Hum

When a cathode is heated by AC, the current generates a magnetic field which can modulate the electron stream, and a modulating voltage is injected into the control grid through the interelectrode capacitance and leakages: additionally there can be emission from the heater in an indirectly heated valve (see below).

When operating directly heated valves such as transmitting valves with thoriated tungsten filaments, the filament supply should be connected to earth by a centre tap or a centre-tapped resistor connected across the filament supply (a hum-bucking resistor). The hum is usually expressed as an equivalent voltage (in microvolts) applied to the control grid. Valve hum should not be confused with hum generated in other circuit components.

Valve Applications

Amplifiers

When an impedance is connected in series with the anode of a valve and the voltage on the grid is varied, the resulting change of anode current will cause a voltage change across the impedance. The curves in **Fig 3.87** illustrate the classifications of valve amplifier operating conditions, showing anode current/grid voltage characteristics and the anode current variations caused by varying the grid voltage.

Class A: The mean anode current is set to the middle of the straight portion of the characteristic curve. If the input signal is allowed either to extend into the curved lower region or to approach zero grid voltage, distortion will occur because grid current is caused to flow by the grid contact potential (usually 0.7-1.0V). Under Class A conditions anode current should show

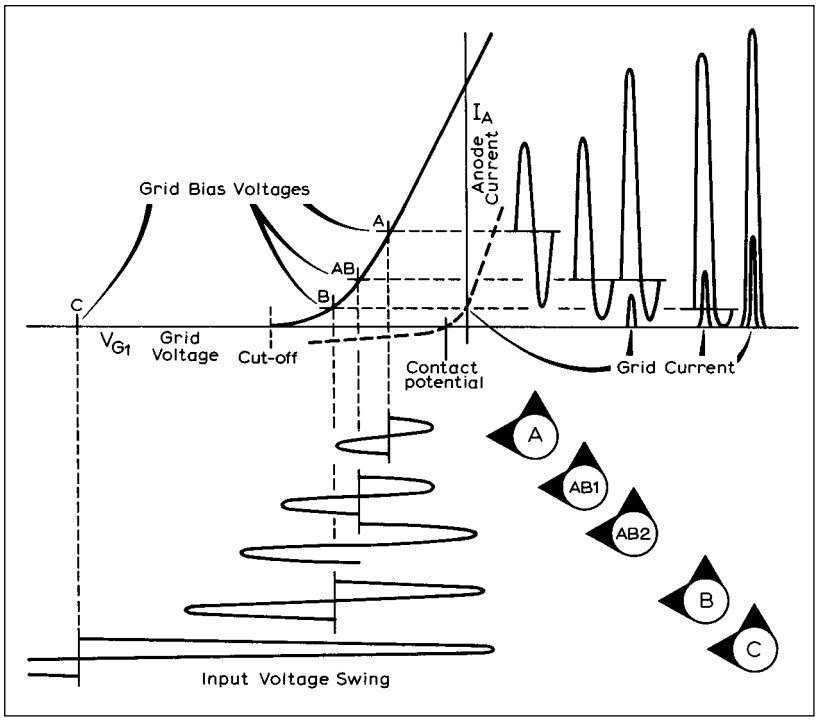


Fig 3.87: The valve as an amplifier: the five classes of operation

no movement with respect to the signal impressed on the grid. The amplifier is said to be linear.

Class AB1: The amount of distortion produced by a non-linear amplifier may be expressed in terms of the harmonics generated by it. When a sine wave is applied to an amplifier the output will contain the fundamental component but, if the valve is allowed to operate on the curved lower portion of its characteristic, ie running into grid current, harmonics will be produced as well. Harmonic components are expressed as a percentage of the fundamental. Cancellation of even harmonics may be secured by connecting valves in push-pull, a method which has the further virtue of providing more power than a single valve can give, and was once widely found in audio amplifiers and modulators.

Class AB2: If the signal input is increased beyond that used in the Class AB1 condition peaks reaching into the positive region will cause appreciable grid current to be drawn and the power output to be further increased. In both the Class AB1 and AB2 conditions the anode current will vary from the zero signal mean level to a higher value determined by the peak input signal.

Class B: This mode, an extension of Class AB2, uses a push-pull pair of valves with bias set near to the cut-off voltage. For zero signal input the anode current of the push-pull pair is low, but rises to high values when the signal is applied. Because grid current is considerable an appreciable input power from the drive source is required. Moreover, the large variations of anode current necessitate the use of a well-regulated power (HT) supply.

Class C: This condition includes RF power amplifiers and frequency multipliers where high efficiency is required without linearity, as in CW, AM and NBFM transmitters. Bias voltage applied to the grid is at least twice, sometimes three times, the cut-off voltage, and is further increased for pulse operation. The input signal must be large compared with the other classes of operation outlined above, and no anode current flows until the drive exceeds the cut-off voltage. This could be for as little as 120° in the full 360° cycle, and is known as the conduction angle. Still smaller conduction angles increase the efficiency further, but more drive power is then needed. Pulse operation is simply 'super Class C'; very high bias is applied to the grid and a very small angle of conduction used.

Grid driving power

An important consideration in the design of Class B or Class C RF power amplifiers is the provision of adequate driving power. The driving power dissipated in the grid-cathode circuit and in the resistance of the bias circuit is normally quoted in valve manufacturers' data.

These figures frequently do not include the power lost in the valveholder and in components and wiring, or the valve losses due to electron transit-time phenomena, internal lead impedances and other factors. Where an overall figure is quoted, it is given as driver power output.

If this overall figure is not quoted, it can be taken that at frequencies up to about 30MHz the figure given should be multiplied by two, but at higher frequencies electron transit-time losses increase so rapidly that it is often necessary to use a driver stage capable of supplying 3-10 times the driving power shown in the published data.

The driving power available for a Class C amplifier or frequency multiplier should be sufficient to permit saturation of the driven valve, ie a substantial increase or decrease in driving power should produce no appreciable change in the output of the driven stage. This is particularly important when the driven stage is anode-modulated.

Passive grid

In linear amplifiers the driver stage must work into an adequate load, and the use of the passive grid arrangement is to be recommended. A relatively low resistance (typically 1kΩ) is applied between grid and cathode with a resonant grid circuit where appropriate. This arrangement helps to secure stable operation but should not be used as a cure for amplifier instability.

Grounded cathode

Most valves are used with the cathode connected to chassis or earth, or where a cathode-bias resistor is employed it is shunted with a capacitor of low reactance at the lowest signal frequency used so that the cathode is effectively earthed. In modulated amplifiers two capacitors, one for RF and the other for AF, must be used.

Grounded grid

Although a triode must be neutralised to avoid instability when it is used as an RF amplifier, this is not always essential if an RF type of tetrode or pentode is employed. However, above about 100MHz a triode gives better performance than a tetrode or pentode, providing that the inherent instability can be overcome. One way of achieving this is to earth the grid instead of the cathode so that the grid acts as an RF screen between cathode and anode, the input being applied to the cathode. The capacitance tending to make the circuit unstable is then that between cathode and anode, which is much smaller than the grid-to-anode capacitance.

The input impedance of a grounded-grid stage is normally low, of the order of 100Ω, and therefore appreciable grid input power is required. Since the input circuit is common to the anode-cathode circuit, much of this power is, however, transferred directly to the output circuit, ie not all of the driving power is lost.

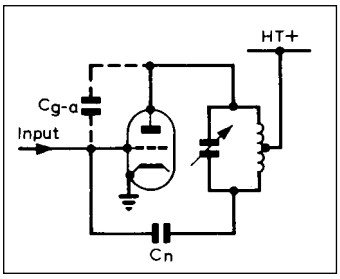


Fig 3.88: Neutralising a grounded-cathode triode amplifier. The circuit is equally suitable for a tetrode or a pentode

Grounded anode

For some purposes it is desirable to apply the input to the grid and to connect the load in the cathode circuit, the anode being decoupled to chassis or earth through a low-reactance capacitor. Such circuits were employed in cathode followers and infinite-impedance detectors.

Neutralising amplifiers

Instability in RF amplifiers results from feedback from the anode to the grid through the grid-to-anode capacitance and is minimised by using a tetrode or pentode. At high frequencies, particularly if the grid and/or anode circuit has high dynamic impedance, this capacitance may still be too large for complete stability. A solution is to employ a circuit in which there is feedback in opposite phase from the anode circuit to the grid so that the effect of this capacitance is balanced out. The circuit is then said to be neutralised.

A typical arrangement is shown in Fig 3.88. Here the anode coil is centre-tapped in order to produce a voltage at the 'free' end which is equal and opposite in phase to that at the anode end. If the free end is connected to the grid by a capacitor (Cn) having a value equal to that of the valve grid-to-anode capacitance (Cg-a) shown dotted, any current flowing through Cg-a will be exactly balanced by that through Cn. This is an idealised case because the anode tuned circuit is loaded with the valve anode impedance at one end but not at the other; also the power factor of Cn will not necessarily be equal to that of Cg-a.

The importance of accurate neutralisation in transmitter power amplifier circuits cannot be overstressed, and will be achieved if the layout avoids multiple earth connections and inductive leads; copper strip is generally preferable to wire for valve socket cathode connections.

Disc Seal Valves

In the disc seal triode (Figs 3.89 and 3.90), characteristically of high mutual conductance, the electrode spacing is minimal. The 'top hat' cathode contains the insulated heater, one side of which is connected to the cathode and the other brought out coaxially through the cathode sleeve connection. The fine-wire grid stretched across a frame emerges through the envelope by an annular connection. Because the clearance between grid and cathode is very small, the cathode surface is shaved during construction to provide a plane surface. The anode also emerges via an external disc for coaxial connection. On larger disc seal valves the anode may form part of the valve envelope.

Disc seal valves are available for power dissipations of a few watts to 100W with forced air cooling and outputs in the frequency range 500-6000MHz. It should be noted that maximum power and frequency are not available simultaneously.

Disc seal valves, although intended for coaxial circuits, may be effectively employed with slab-type circuits. Important points to be observed are (a) only one electrode may be rigidly fixed, in order to avoid fracture of the seals (which is more likely to occur in the glass envelope types); and (b) except in the case of forced-air anode cooling the anode is cooled by conduction into its associated circuit. With a shunt-fed circuit thin mica insulation will function both as a capacitance and a good transmitter of heat.

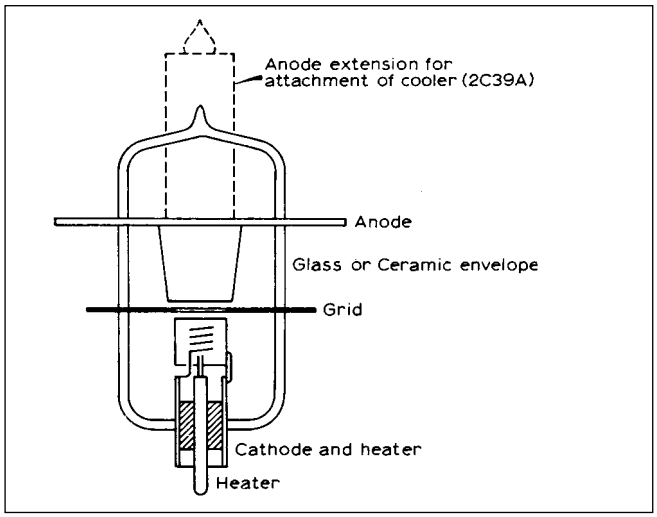


Fig 3.89: General form of a disc seal valve

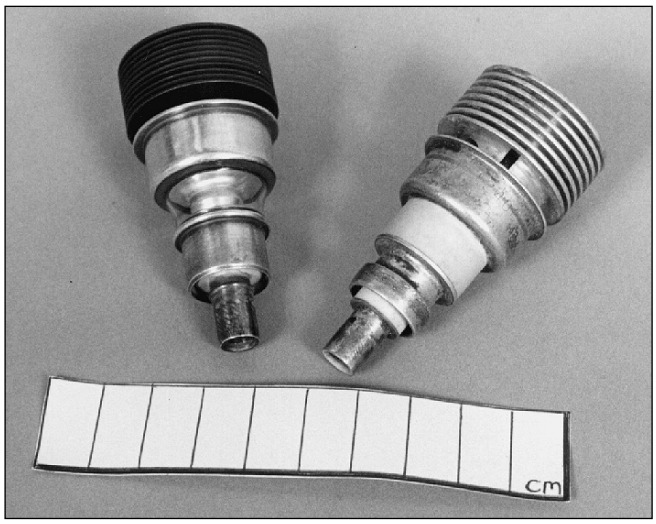


Fig 3.90: The 2C39 disc seal triode, showing glass version (left and ceramic version (right). The anode cooling fins on the ceramic version are removable, clamped in place with Allen bolts

Certain sub-miniature metal ceramic envelope types such as the 7077, although of the generic disc seal form, require special sockets if they are used in conventional circuits. Many of them give significant output at the lower SHF bands.

CATHODE-RAY TUBES

A cathode-ray tube contains an electron gun, a deflection system and a phosphor-coated screen for the display. The electron gun, which is a heated cathode, is followed by a grid consisting of a hole in a plate exerting control on the electron flow according to the potential applied to it, followed in turn by an accelerating anode or anodes.

The simplest form of triode gun is shown at Fig 3.91. The beam is focused by the field between the grid and the first anode. In tubes where fine line spots and good linearity are essential (eg in measurement oscilloscopes), the gun is often extended by the addition of a number of anodes to form a lens system. Beam focusing may be by either electrostatic or electromagnetic means.

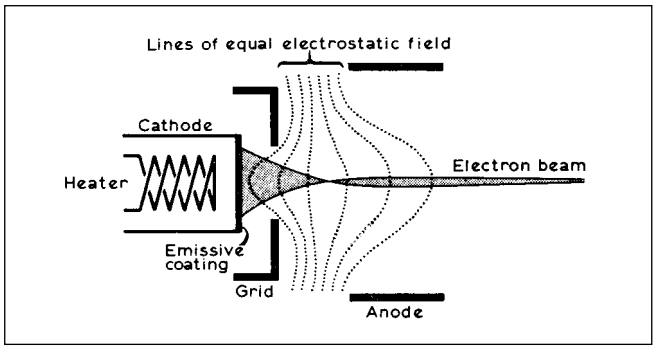


Fig 3.91: Diagram of the electron gun used in cathode-ray tubes, travelling wave tubes and klystrons

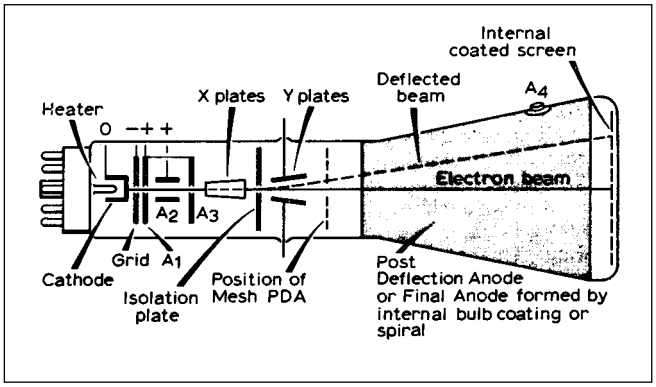


Fig 3.92: Diagrammatic arrangement of a cathode-ray tube with electrostatic focusing and deflection

Oscilloscope Tubes

Electrostatic focusing is used in oscilloscope tubes, the deflection system consisting of pairs of plates to deflect the beam from its natural centre position, depending on the relative potentials applied. Interaction between the two pairs of deflection plates is prevented by placing an isolation plate between them (Fig 3.91).

After its deflection the beam is influenced by a further accelerating electrode known as the post-deflection accelerator (PDA) which may take the form of a wide band of conducting material on the inside of the cone-shaped part of the bulb, or of a close-pitch spiral of conducting material connected to the final anode. For some purposes when it is important to maintain display size constant irrespective of the final anode voltage, a mesh post-deflection accelerator is fitted close to the deflecting plates.

If a double beam is required, the electron flow is split into two and there are two sets of deflection plates. Alternatively, two complete systems are enclosed in one tube. Although a common X-deflection plate may be fitted, the advantage of two complete systems lies in providing complete alignment of the timing (horizontal) deflection. By setting two systems at an angle to one another adequate overlap of each of the displays is provided.

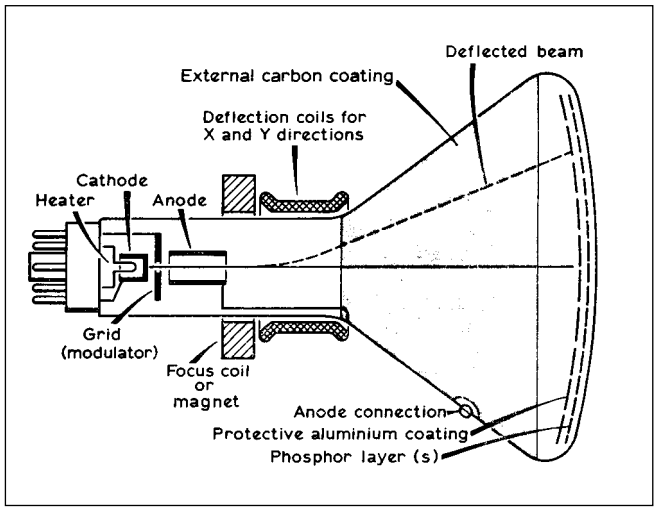


Fig 3.93: Diagrammatic arrangement of a cathode-ray tube having magnetic focusing and deflection

Radar and Picture Tubes

In radar and television tubes magnetic focusing and deflection are common (**Fig 3.93**), the deflection angle being vastly greater than with an oscilloscope tube. Such tubes employ a simple triode or tetrode electron gun with an anode potential of 20kV or more.

Screen Phosphors

Many types of phosphor are used for coating the screens of cathode-ray tubes, their characteristics varying according to the application. In all of them the light output is determined by the final anode voltage used, but where this exceeds about 4kV the phosphor is protected against screen burn by a thin backing layer of evaporated aluminium. Oscilloscope tubes require a phosphor with a wide optical band to give a bright display for direct viewing. This phosphor, yellow-green in colour, extends into the blue region to enable direct photographs to be taken from the display.

CR Tube Power Supplies

Unlike the valve, the cathode ray tube's current requirements are low, the beam current being only a few tens of microamperes. For oscilloscope work the deflector plates need to be at earth potential, cathode and other electrodes consequently being at a high negative potential to earth.

In magnetically focused tubes the cathode may be at earth potential and the final anode many kilovolts above. Power supplies for either type of tube need to be of very high impedance for safety reasons, and the short-circuit current should not exceed about 0.5mA.

Further Details

Klystrons, Magnetrons and travelling wave tubes are covered in earlier versions of this handbook along with technical details of valve circuit design and operation.