REVIEW ARTICLE

# Physics of semiconductor power devices

Rep. Prog. Phys., Vol. 45, 1982. Printed in Great Britain

# Physics of semiconductor power devices

Adolph Blicher†

**Abstract**

Semiconductor power devices are designed to rapidly switch or amplify high currents, to support high voltages, and to control electric power. Because of these requirements, their topographies and structures are different from those of small-signal devices. Specific designs are the result of the understanding of the physics of p–n junction HV breakdown, gain variation at high currents, current instabilities, etc. After introducing elementary semiconductor structures, the article reviews the basic principles of the operation of bipolar and field-effect power transistors and thyristors. This is followed by a discussion of phenomena of special interest for power devices: junction avalanche breakdown, the effect of the current level on the current gain, dynamic and static behaviour at high currents and thermal properties and instabilities. The review includes recent advances in device physics and introduces the reader to new methods of improving device performance.

Power is a relative concept—at microwave frequencies a few watts is a very large quantity; this review therefore includes a section on microwave power devices.

This review was received in September 1981.

†Formerly with RCA Laboratories, Solid State Technology Center, Somerville, NJ, USA.

0034-4885/82/040427+42$08.00  

**Contents**

## 1. Introduction

### *1.1. Power devices versus small-signal devices*

The feature that distinguishes semiconductor power devices from small-signal devices is the magnitude of the requisite currents and voltages. Small-signal devices now encountered almost exclusively in integrated chips, are mostly used as logic switches that are either turned on or off, are reproduced thousands of times on the same crystal chip, and operate with very small currents and low voltages to achieve negligibly small power losses.

Power devices, on the other hand, are meant, in most cases, *to rapidly switch or amplify high currents and to support high voltages*, dissipating as little power as possible. For example, a colour television horizontal deflection transistor is required to turn on a current of 10 A and to support in the off state 2.5 kV. Special techniques such as p–n junction shaping are needed to achieve the requisite high-voltage ratings. The high-current requirement, on the other hand, leads to topographical and structural designs that permit the achievement of high-current gains and minimal internal voltage drops when the device is turned on. The product of this voltage drop and current determines the magnitude of the device's thermal dissipation and its junction temperature, which for silicon is not allowed to exceed 150–170°C.

To achieve a reasonable cost, the area of the semiconductor chip used for the device should be made as small as is compatible with the acceptable internal voltage drop and other requirements such as current gain and speed of operation. A small device area not only reduces the cost of the utilised semiconductor material, but also makes it possible to simultaneously process thousands of devices, thus considerably reducing their cost.

Because of some of these considerations, the physical design of power devices substantially differs from that of small-signal processing devices, both in cross-sectional structures and surface geometry. In some cases, the most recent design of power transistors profitted from the progress made in integrated circuit technology. For example, the power metal-oxide–silicon field-effect transistor (MOSFET) is typically designed with thousands of very small elementary cells integrated on the same chip and all working in parallel, in order to increase device current capability and to decrease its internal resistance and voltage drop.

Semiconductor power devices have found innumerable applications both in industrial and consumer equipment such as power conditioning circuits (inverters, converters, power supplies, etc), motor controls, TV horizontal deflection circuits, light dimmers, audio amplifiers, car ignition systems, temperature control circuits, infrared heating, battery charging, printer solenoids, etc.

### *1.2. Basic components*

Semiconductor devices always include some combination of p–n junction diodes, metal–insulator–semiconductor capacitors, Schottky barrier diodes and non-rectifying ohmic contacts.

The current flowing in an *ideal* bipolar diode upon the application of a potential $V$ is

$$I = I_0\left[\exp\left(\frac{qV}{nkT}\right) - 1\right] \tag{1.1}$$

where $I_0$ is the saturation current flowing when a reverse potential greater than a few $kT/q$ is applied to the diode, $T$ is the junction absolute temperature, $k$ is the Boltzmann constant, $q$ is the electron charge and $n$ equals 1. In non-ideal diodes, the total current contains a component resulting from the generation–recombination of charge carriers in the junction space-charge (depletion) region (figure 1). This component dominates

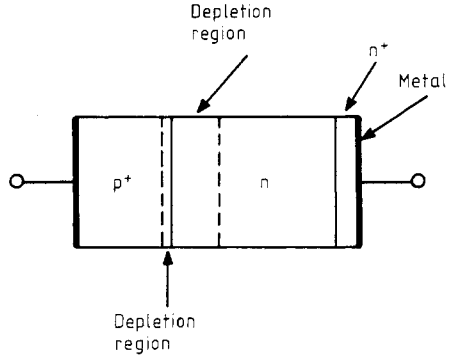


**Figure 1.** $p^+$–n–$n^+$ diode. $n^+$-region added for better ohmic contact. Plus sign indicates highly doped semiconductor.

the diffusion current for reverse voltages or at low forward biases. Expression (1.1) still holds, but with a different magnitude of $I_0$ and with $n$ between 1 and 2. At high current densities, 1 in equation (1.1) can be neglected, $I_0$ acquires a different value, and $n \simeq 2$ because of the appearance of an electric field in the n-base.

An excessively high reverse voltage applied to the diode results in a very rapid increase in the reverse current due to the carriers' avalanche multiplication and junction breakdown.

A Schottky barrier diode, like a p–n junction, has rectifying properties and consists of a metal electrode deposited on a non-degenerate semiconductor surface. The approximate current–voltage relationship of the Schottky diode with uniformly doped silicon is similar in form to expression (1.1) with the constant $n$ slightly higher than unity for low current levels.

The metal-oxide–silicon (MOS) capacitor, the most important of the metal–insulator–semiconductor capacitors, consists of a metal electrode (gate) deposited on an oxidised silicon substrate (figure 2).

Due to the difference between the magnitudes of the metal and silicon work functions, an electrically charged double layer is formed at the metal–silicon interface even in the absence of any external voltage. In the case where the metal work function $q\varphi_m$ is smaller than that of silicon $q\varphi_{Si}$ some electrons are transferred from the gate to the silicon substrate by some very slightly conductive path external to the insulator, and the gate becomes positively charged. The p-silicon substrate becomes negatively charged at the surface because some of the mobile holes are repelled by the positive bias of the gate. The p-type silicon behaves like being less p-doped (depletion) so

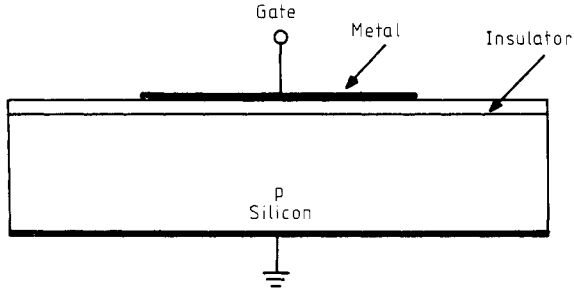


**Figure 2.** Metal-oxide–silicon (MOS) capacitor.

that, at the surface, the silicon band diagram bends as shown in figure 3. The amount of band bending is measured by the surface potential $\varphi_s$.

In order to revert the bands to the flat-band condition, it is necessary to apply a flat-band voltage $V_{FB}$ to the gate equal to the metal-oxide–silicon work function difference.

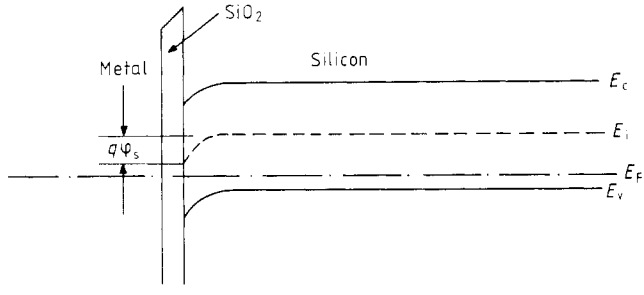


**Figure 3.** Energy band bending at the silicon surface. $\varphi_s$ is the surface potential.

The application of a small positive external gate potential $V_G > V_{FB}$ results in the repulsion of holes from the surface and in the downward band bending. A still higher positive gate voltage attracts electrons to the silicon surface so that it may acquire an excess of electrons and become n-type. This is known as surface inversion and is characterised by the Fermi level (figure 3) crossing the intrinsic level $E_i$. The gate voltage required for strong inversion is known as the threshold voltage $V_T$.

A negative gate potential results in an increase in the number of holes at the silicon surface, i.e. in surface accumulation. In present-day MOS power devices, the gate consists of some refractory metal such as tungsten or, more often, highly-doped polycrystalline silicon.

Fixed and mobile positive oxide charges and interface states have a pronounced effect on the MOS capacitor flat-band and threshold, and also on the silicon surface stability.

### *1.3. Some fabrication techniques*

The most commonly used junction formation technique consists of high-temperature surface impurity deposition with subsequent high-temperature diffusion. Impurity

deposition can also be accomplished by the HV implantation of an appropriate impurity ion.

Commonly used n-type impurities are phosphorus and arsenic. Boron, aluminium and gallium, on the other hand, are frequently used as p-type impurities.

An 'infinite' impurity source, such as phosphorus oxychloride ($POCl_3$) vapour, results in a deposited impurity profile with a complementary error function distribution. On the other hand, diffusion from a *thin* predeposited layer results in a Gaussian impurity distribution.

Ion implantation is widely employed for those cases when an accurately reproducible doping profile is essential. It consists of bombarding the semiconductor surface with the desired impurity ions in an apparatus that is essentially a modified mass spectrometer in which the impurity ions are accelerated to very high velocities by electrical potentials of up to some hundreds of thousands of volts. The implanted impurity distribution is Gaussian but it flattens when followed by a diffusion step so that the concentration peak shifts from under the silicon surface to the surface itself (figure 4). Ion implantation is not useful, however, for very deep diffusions.

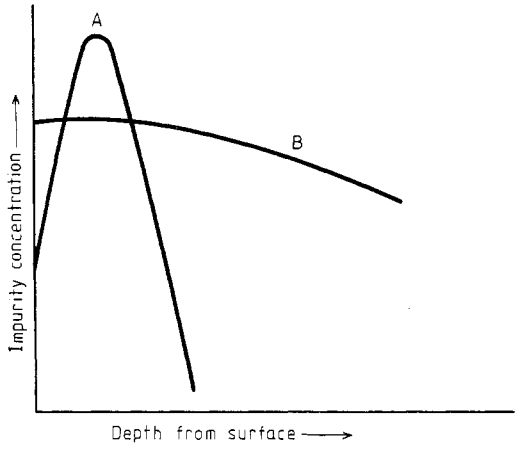


**Figure 4.** Concentration profile of ion implant (A) before and (B) after diffusion.

The high-resistivity silicon substrate used for the fabrication of high-voltage devices may be either bulk or epitaxially grown. For breakdown near or above 1 kV, the best material is the transmutation-doped silicon crystal (Herrman and Herzer 1975). Very uniformally doped silicon, free of resistivity striations, is produced by bombarding silicon crystal with slow thermal neutrons, which transmute the uniformly distributed silicon isotope $^{30}Si$ into phosphorus $^{31}P$. Thus, the silicon crystal becomes lightly and very uniformly doped with an n-type impurity. For voltages below 1 kV, silicon epitaxial layers are often adequate for the fabrication of bipolar and field-effect devices.

## 2. Basics of power device operation

### *2.1. Bipolar power transistors*

A bipolar type transistor consists of the emitter-base and collector-base junctions separated by a base of width $W_B$ smaller than the minority carrier diffusion length $L$.

The collector structure of an n–p–n power device comprises an n -drift region (figure 5) to support the requisite high collector voltage. (A p–n–p power transistor would have a $p^-$-drift region.)

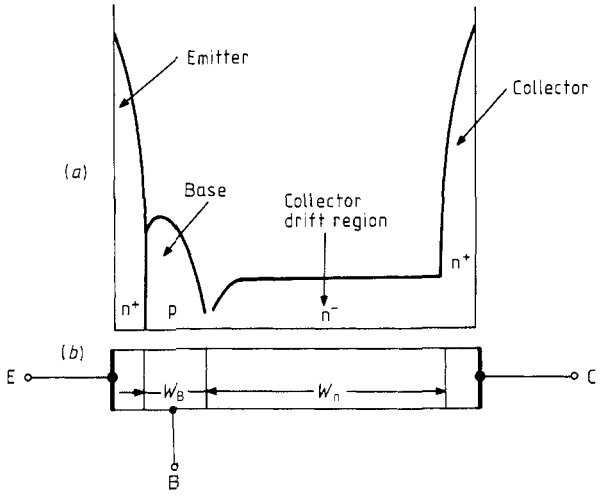


**Figure 5.** (*a*) Diffusion profile and (*b*) structure of the $n^+pn^-n^+$ power transistor.

With the emitter forward-biased and the collector reverse-biased the DC current gain of a transistor in the common-base circuit is

$$\alpha = \gamma\alpha_T \tag{2.1}$$

where $\alpha_T$ is the base transport factor given by

$$\alpha_T = \frac{1}{\cosh(W_B/L)}. \tag{2.2}$$

$\gamma$ is the emitter efficiency defined as the ratio of the current injected by the emitter into the base to the sum of the currents injected into the base and from the base into the emitter.

The collector current $I_C$ is related to the emitter current $I_E$ by

$$I_C = \alpha I_E + I_{CBO} \tag{2.3}$$

where $I_{CBO}$ is the collector leakage current.

At the common-emitter configuration the current gain is

$$h_{FE} = \frac{\alpha}{1-\alpha} \tag{2.4}$$

and the emitter efficiency

$$\gamma_E = \frac{\gamma}{1-\gamma} \tag{2.5}$$

is the ratio of the current injected by the emitter into the base to the current injected by the base into the emitter.

It can be shown (see, for example, Blicher 1981) that for an n–p–n transistor

$$\gamma_E = \frac{Q_E}{D_p}\frac{D_n}{Q_B} \tag{2.6}$$

where $Q_E$ is the total *number* of charges per unit area in the effective emitter and $Q_B$ is the total *number* of charges per unit area in the base. $D_p$ and $D_n$ are the hole and electron diffusivities, respectively. $Q/D$ ratios are known as *Gummel numbers* (Gummel 1961, Gummel and Poon 1970).

The $I$–$V$ characteristics of a high-voltage power transistor in the common-emitter configuration can be represented by the diagram of figure 6 with four distinct parts:

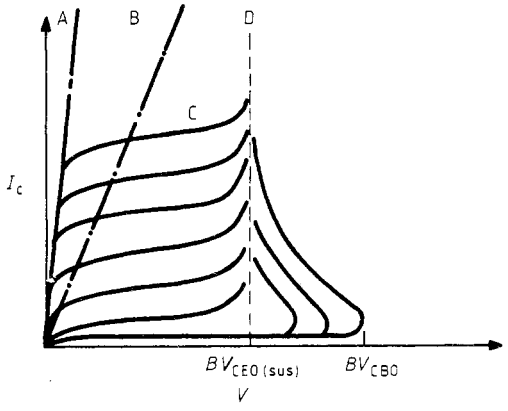


**Figure 6.** Saturation (A), quasi-saturation (B), active (C) and breakdown (D) regions of an $n^+pn^-n^+$ power transistor.

the saturation, quasi-saturation, active and breakdown regions. If the load line intersects the saturation or quasi-saturation region the device is turned on when operating as a switch with the collector forward-biased. The active region is traversed by the load line during turn-off and intersected when the device operates as an amplifier. The collector breakdown voltage $BV_{CBO}$ occurs with the emitter circuit open and is equal to the collector junction breakdown voltage. The $BV_{CEO(sus)}$ is the lowest sustained collector-to-emitter breakdown voltage with the base circuit open. It is substantially lower than $BV_{CBO}$ (Moll *et al* 1970)

$$BV_{CEO(sus)} = fBV_{CBO}(1 - h_{FEO})^{-1/n} \tag{2.7}$$

where $h_{FEO}$ is the common-emitter peak current gain. For step junctions, $f \approx 0.65$ and $n = 4$ for an n–p–n transistor and lower voltages, and $n = 6$ for higher (1000 V) voltages.

### *2.2. Bipolar power thyristors*

The silicon-controlled rectifier (SCR), the Triac (triode AC switch) and the Shockley diode belong to the class of semiconductor devices known as thyristors (see, for example, Blicher 1976). The SCR is a four-layer pnpn structure (figure 7) consisting of an $n^+$ cathode emitter, p-base, $n^-$-base, and p-anode emitter. The control electrode

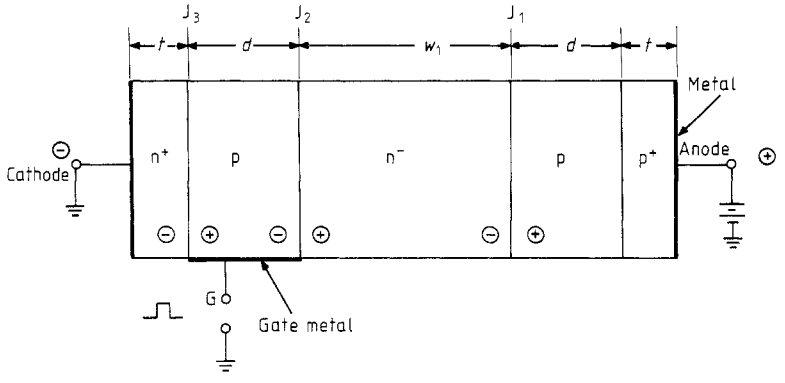


**Figure 7.** Schematic representation of an SCR. Junction polarities shown are for the forward blocking state.

G is known as the gate. Without the gate, the structure is known as the pnpn Shockley diode which can be activated by the application of a high anode voltage.

In the turn-off state the positive (forward) potential applied to the thyristor anode is blocked by the reverse-biased centrally located junction, $J_2$. Junction polarities for this situation are shown in figure 7. A negative potential applied to the anode is supported by the then-reverse-biased junctions $J_1$ and $J_3$. The thyristor is turned-on by the application of a short positive-current pulse to the gate G but is turned off either by opening the anode circuit or by applying a negative (reverse) potential to the anode. When the thyristor is turned on, all three junctions are forward-biased.

An SCR conducts only in one current direction; its $I$–$V$ characteristic is shown on the right of figure 8 (quadrant I). To obtain the bidirectional conduction necessary for AC applications, two pnpn structures with one p-gate are integrated back-to-back on the same silicon chip, forming a Triac, with the $I$–$V$ characteristics of figure 8. The Triac can be turned on in either quadrant I or III.

$I_s$ and $V_s$ are the coordinates of the variously called switching, turn-on or breakover point. $I_h$ and $V_h$ are the coordinates of the holding point beyond which the device is in the on state.

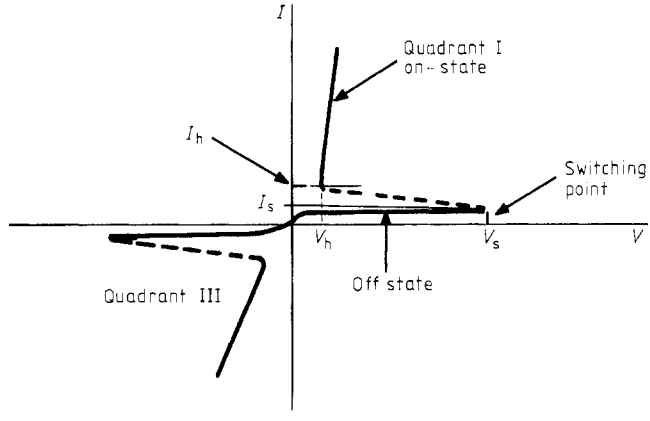


**Figure 8.** $I$–$V$ characteristics of an SCR (quadrant I) and of a Triac (quadrants I and III).

An SCR can be considered as consisting of an npn and a pnp transistor sharing the same junction $J_2$ as their common collector junction. The gate is usually connected to the p-base. Only a small saturation (leakage) current flows in the device when a positive potential is applied to the anode. Upon the application of a positive gate potential, the $n^+$-emitter (cathode) starts emitting electrons and the collector of the npn transistor provides the n-base with electrons which bias the p-emitter (anode) in the forward direction. The pnp transistor becomes activated, supplying a hole current that is collected by $J_2$ and drives the p-base of the npn transistor. The effect becomes regenerative so that both transistors become saturated, i.e. highly conductive with all three junctions forward-biased.

The turn-on at the $I_s$–$V_s$ point takes place when the anode current $I_A$ starts growing extremely fast as a function of the gate current $I_G$, i.e. when

$$\frac{dI_A}{dI_G} = \frac{\alpha_2}{1-(\alpha_1+\alpha_2)} \to \infty. \tag{2.8}$$

$\alpha_1$ and $\alpha_2$, the small-signal common-base current gains of the pnp and the npn transistors, respectively, are very small at low current levels but rapidly increase as the current density increases. The region between the turn-on and the holding points is unstable so that the device switches very fast for sufficiently high gate driving current.

To avoid undesirable switching either by the leakage currents at high temperatures or by capacitive currents, SCR are usually provided with an n-emitter short that consists of, for example, cathode metallisation extended over the p-base as shown by figure 9. A shorted emitter has zero efficiency at low currents but a high efficiency at high

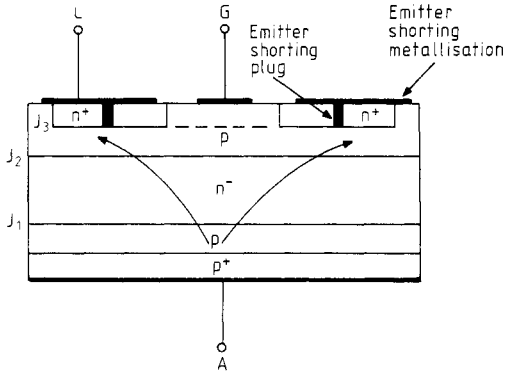


**Figure 9.** SCR cathode with two types of emitter shorts.

currents since some part of the anode current is flowing toward the shorts without reaching the $n^+$-emitter. The capacitive and thermally generated currents flow at least partially through the shorts so that the sum of the $\alpha$ reaches unity much later. This way a shorted emitter provides immunity to the capacitive current $C\,dv/dt$ and allows the device to operate at higher temperatures. Laterally flowing current causes a lateral voltage drop and forward-biases the emitter, making the emitter shorting less effective. This can be avoided by designing emitters consisting of narrow fingers and/or by providing additional shorts, known as shorting holes or plugs as depicted by figure 9.

*2.2.1. Forward and reverse thyristor blocking capability.* In the off state (no conduction) the thyristor centre junction $J_2$ must support the high positive (forward) potential applied to the anode, since the $J_1$ and $J_3$ junctions are forward-biased. In the case of the non-shorted emitter, the $J_2$ junction breakdown $BV$ and the breakover voltage $V_{BRO}$ (the voltage that causes the thyristor to turn-on without gate action) is given very approximately by (Herlet 1968a, b)

$$V_{BRO} = BV(1-\alpha_1-\alpha_2)^{1/n} \tag{2.9}$$

where $n$ is between 4 and 6 for the step junctions. With a shorted cathode (2.9) simplifies still further since at low currents $\alpha_2 \simeq 0$.

To turn off the thyristor, the anode voltage is made negative. This reverse voltage is supported by the $J_1$ junction since $J_2$ becomes forward-biased, and $J_3$ breaks down at very low voltages due to the highly-doped p-base. The device breakdown voltage is, therefore, that of *an open-base pnp transistor* and is about the same as the breakover voltage with shorted n-emitter (Herlet 1968a, b), i.e.

$$V_{CEO(sus)} = BV(1-\alpha_1)^{1/n}. \tag{2.10}$$

In many circuits, such as those used for inverters, the reverse-voltage blocking capability is not utilised and the SCR may be designed to block only the forward potential. This allows more freedom to improve other device parameters such as the forward-blocking capability or the forward voltage drop. Such a device is known as a reverse conducting thyristor (RCT) (Kokosa and Tuft 1970). Since the anode of an RCT does not support any appreciable voltage, it may be shorted similarly to the cathode. This prevents the pnp transistor from magnifying the $J_2$ junction leakage current and makes the forward breakdown voltage equal to the $J_2$ junction breakdown voltage since now both $\alpha_1$ and $\alpha_2$ are negligibly small at low temperatures.

*2.2.2. Light-activated thyristors.* These devices are mainly useful for effective isolation of the triggering device from the main thyristor switch. They are used, for example, to turn on high-voltage transmission lines (Gerlach 1977).

To optically trigger the thyristor, the active region of the pnpn structure is illuminated with light whose energy $h\nu$ exceeds that of the silicon energy gap. The incident light may be introduced into the p-base through an opening in the central portion of the cathode or through a thin $n^+$-emitter (Konishi *et al* 1980). The light generates charge carriers that provide hole and electron currents for the thyristor's p- and n-base, respectively. For very high power devices, smaller auxiliary optically activated thyristors are used to fire the main device. Light-activated thyristors may be easily connected in series to support very high potentials of the order of several kilovolts.

### *2.3. Junction field-effect power transistors (JFET)*

Field-effect (unipolar) transistors operate primarily by the drift of the majority carriers whereas the bipolar devices operate by diffusion and drift of the minority and majority carriers.

Figure 10 represents schematically one-half of an n-channel JFET. The electron current flowing in the device from the negative source to the positive drain is modulated by the device control electrode, the gate, consisting of the $p^+$–n junction. The variable reverse potential applied between the gate and the source changes the width of the

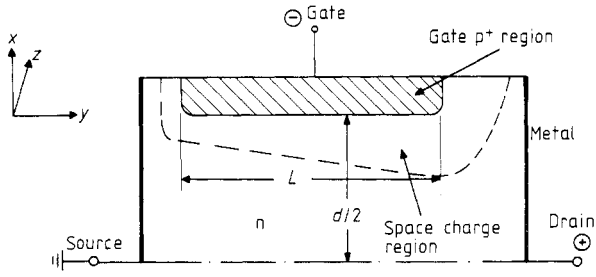


**Figure 10.** Schematic representation of an n-channel junction field-effect transistor (JFET).

gate junction space-charge region and thus controls the magnitude of the electron current.

By computing the ohmic voltage drop along the conducting channel (see, for example, Blicher 1981), it can be shown that at drain voltages much smaller than the gate-to-source voltage the drain current can be expressed by

$$I_D = G_0 V_D f(V_G) \tag{2.11}$$

where

$$G_0 = q\mu_n N_D Z d/L \tag{2.12}$$

and

$$f(V_G) = 1 - \left(\frac{8\epsilon_s(V_i - V_G)}{qN_D d^2}\right)^{1/2} \tag{2.13}$$

where $q$ is the electron charge, $\mu_n$ is the electron mobility in the n-type channel, $\epsilon_S$ is the silicon permittivity, $Z$ is the channel width in the $Z$ direction perpendicular to the page, $L$ is the channel length (figure 10), $d$ is the maximum possible channel width, $V_i$ is the gate-junction built-in voltage, $V_G$ is the applied gate potential and $V_D$ is the drain-to-source voltage. At low drain voltages, the current depends linearly on $V_D$ but varies as the square root of the gate voltage as one would expect because of the assumption that the gate is an abrupt junction.

The output characteristics of a JFET (figure 11($a$)) show that close to the origin the drain current $I_D$ varies linearly with the drain voltage. At larger voltages its growth slows down; eventually, $I_D$ becomes almost constant and independent of $V_D$, i.e. it saturates. This takes place when the gate-junction space-charge regions begin to touch each other at some saturation voltage $V_{Dsat}$. For drain voltages greater than $V_{Dsat}$, the drain current partially flows through the drain space-charge region, where, for electronic fields greater than $10^4$ Vcm$^{-1}$ the electron velocity reaches its maximum value of $v_s \simeq 10^7$ cm s$^{-1}$. As the drain voltage increases beyond $V_{Dsat}$, the space-charge region moves toward the source, only very slightly decreasing the channel length and increasing the drain current somewhat. Consequently, the $I$–$V$ characteristic in the saturation region has a non-zero slope, i.e. the saturation conductance has some finite value. The drain current in the saturation region is proportional to $G_0$ and depends on the first and 3/2 power of the gate voltage.

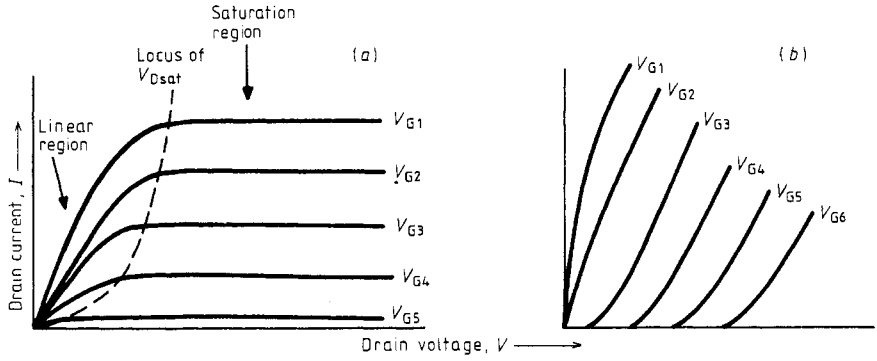


**Figure 11.** (*a*) Pentode-like and (*b*) triode-like $I$–$V$ characteristics of a n-channel JFET or n-channel MOSFET with the gate voltage as a parameter. The JFET drain current increases with the decreasing magnitude of the negative gate potential; the MOSFET current increases with increasing positive gate voltage.

The effectiveness of the gate voltage's control over the drain current is expressed by the transconductance

$$g_m \equiv \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D=\text{constant}}. \tag{2.14}$$

In the saturation region the transconductance reaches its maximum value which is equal to the conductance in the linear region.

*2.3.1. JFET with non-saturating triode-like characteristics.* The $I$–$V$ characteristics of figure 11(*a*) are known as pentode-like; those of figure 11(*b*) are triode-like (Yamaguchi *et al* 1975, Nishizawa *et al* 1975, Yamaguchi and Kodera 1977, Nishizawa and Yamamoto 1978, Gupta 1980, Ohmi 1980). In a very short completely depleted channel a potential barrier at the source prevents the current from flowing at low drain voltages. At sufficiently high drain potential, the barrier is lowered by the drain's electrostatic field so that an electron current starts flowing. The magnitude of this current depends exponentially on the drain voltage $V_D$. In order to obtain triode-like operation, the maximum channel width should be small, of the order of a few $\mu$m, and the gate length should be smaller than one-half of the channel width (figure 10). If the gate is too long, the drain electrostatic field becomes too small to reduce the barrier height and the characteristics remain pentode-like.

An example of a JFET structure known as the V-groove JFET is shown by figure 12 (Mok and Salama 1978). The V-shaped channel region is obtained by anisotropic etching of the silicon crystal. For the high-current handling capability, there are many such devices on the same chip connected in parallel. Therefore, there are special isolation grooves etched as shown in the figure.

Another JFET structure, depicted by figure 13, is known as the static induction transistor (Nishizawa *et al* 1975). The embedded mesh-like $p^+$-grid acts as a field-effect transistor gate. This device was constructed with a very short channel length and minimised source series resistance to obtain triode-like characteristics.

## *2.4. Field-effect thyristor*

The field-effect thyristor (figure 14) (Wessels and Baliga 1978, Baliga 1981) is fabricated by using anisotropic preferential etching to obtain deep wells (grids) filled

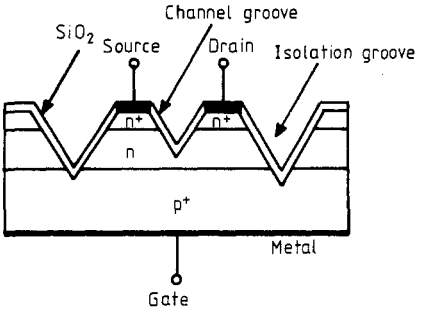


**Figure 12.** Vertical V-grooved JFET (from Mok and Salama 1978).

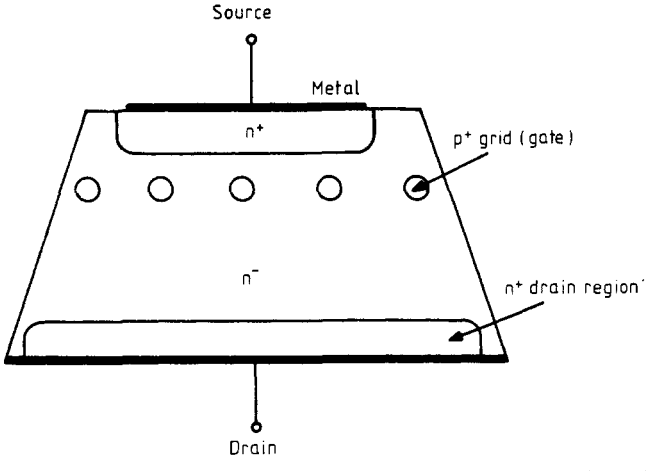


**Figure 13.** Static induction transistor (SIT) (from Nishizawa *et al* 1975).

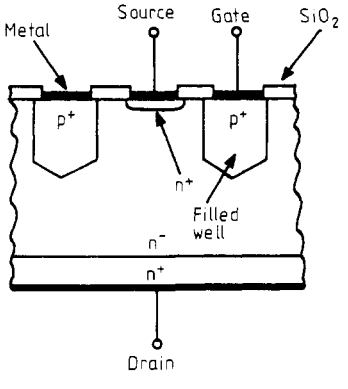


**Figure 14.** Junction field-effect thyristor (from Baliga 1981).

by epitaxial $p^+$ single-crystal silicon. The grid acts as a junction gate, forming a depletion region that modulates the current flowing from the cathode to the anode. These devices are normally in the on state with the anode positively biased. The current flow can be stopped by reverse-biasing the grid. Field-effect thyristors are capable of operation at temperatures up to 200°C, considerably higher than bipolar

thyristors which normally cannot work above 140°C. This is possible because this type of device does not exhibit the regenerative switching mechanism which is very sensitive to the device temperature.

### *2.5. Metal Schottky barrier field-effect transistor (MESFET)*

The $p^+$–n junction gate of the junction field-effect transistor can be replaced by a Schottky barrier consisting of a metal electrode deposited on a semiconductor surface. This type of gate is particularly useful for GaAs devices where good-quality p–n junctions or high-quality surfaces are more difficult to achieve than in silicon. MESFET are very useful as microwave devices at frequencies above 4 GHz. Figure 15 (Liechti

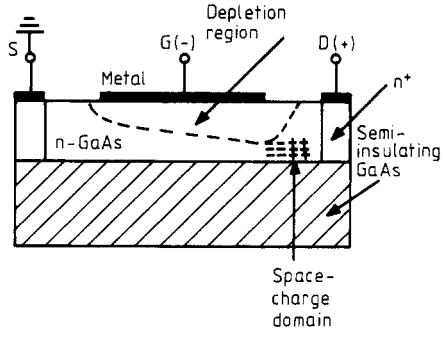


**Figure 15.** Schematic representation of MESFET (from Liechti 1976).

1976) represents a schematic cross section of a device consisting of an n-type GaAs film deposited on a semi-insulating GaAs substrate which also provides a thermal path to a heat sink. Some part of the n-region surface is metallised to form the Schottky barrier. The conducting channel lies below the surface so that the high bulk mobility of GaAs can be fully utilised (the surface mobility is always lower than that of the bulk). Electron mobility in GaAs at room temperature is $8800\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$ whereas that for very lightly doped silicon crystal is only $1350\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$. The dependence of the GaAs electron mobility on the electric field is different from that of silicon (figure 16). The upper curve for GaAs shows the presence of a peak value equal to about $2v_s$ at an electric field of about $3\times10^3\,\mathrm{V\,cm^{-1}}$. For higher fields, the velocity eventually drops down to the saturated velocity $v_s \simeq 10^7\,\mathrm{cm\,s^{-1}}$ at an electric field of about $1.8\times10^4\,\mathrm{V\,cm^{-1}}$. The theory of GaAs MESFET operation is further complicated by the appearance of the Ridley–Gunn stationary space-charge domain

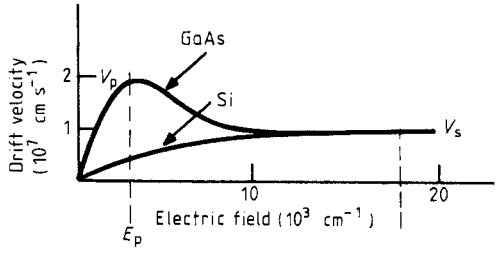


**Figure 16.** Electron mobility of silicon and GaAs as a function of electric field (from Ruch 1972).

formation (figure 15) (Liechti 1976, Shur 1978). The electron drift velocity in a short channel device reaches its peak velocity $v_p$ close to the channel centre but drops down to its $v_s$ value under the drain side of the gate edge, where the channel width is narrowest (figure 15). As a result, electrons accumulate in this region to preserve the current continuity. Beyond this region the channel thickness increases, the field decreases and the electron velocity goes up so that less electrons are needed to carry the same current; therefore, a positive space-charge region is formed.

Figure 17 (Wemple *et al* 1980) represents the $I$–$V$ characteristics of a power GaAs MESFET. At a drain voltage $V_D \simeq 5$ V and $V_G \simeq -8$ V, there is a change in the characteristics' shape and the current beyond this point increases as the magnitude of the reverse gate potential increases. This anomaly, attributed to avalanche breakdown in the drain-gate region, limits the maximum obtainable power output. GaAs MESFET can be fabricated to deliver several watts of power even at frequencies as high as 20 GHz.

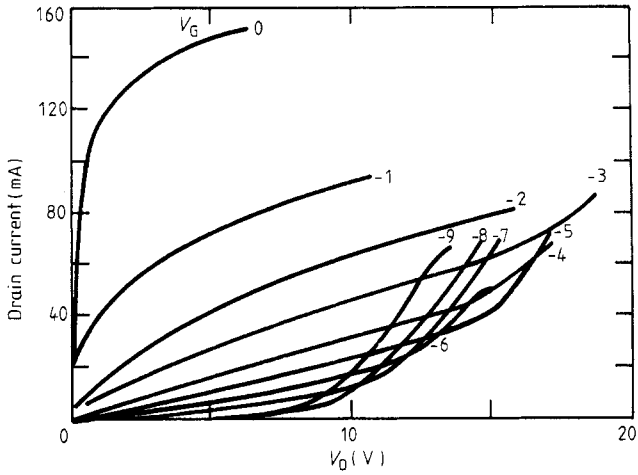


**Figure 17.** MESFET $I$–$V$ characteristics with gate voltage as a parameter (from Wemple *et al* 1980).

### *2.6. Insulated gate field-effect power transistor (IGFET)*

The most common among the insulated gate field-effect transistors is the metal-oxide–silicon field-effect transistor (MOSFET) with silicon dioxide as the gate insulator. The most useful, because of the high electron mobility, is the n-channel enhancement power MOSFET (figure 18). The $n^+$-source region and the p-type bulk (channel region) are connected together by the highly conducting $p^+$-diffusion and are kept at ground potential. The drain junction is positively biased with respect to the source. For a positive gate-to-source voltage greater than the threshold voltage $V_T$, the p-channel becomes induced at the surface and starts conducting electrons from the source to the drain. The drain is a vertical structure for better silicon surface utilisation and easier heat sinking. The device drain contains an $n^-$-drift region to support the requisite high voltage.

The $I$–$V$ characteristics of a MOSFET are very similar to those of a JFET (figure 11($a$)). As the positive gate voltage is increased, the conductance of the channel increases and so does the drain current. At low drain voltages, the drain current is

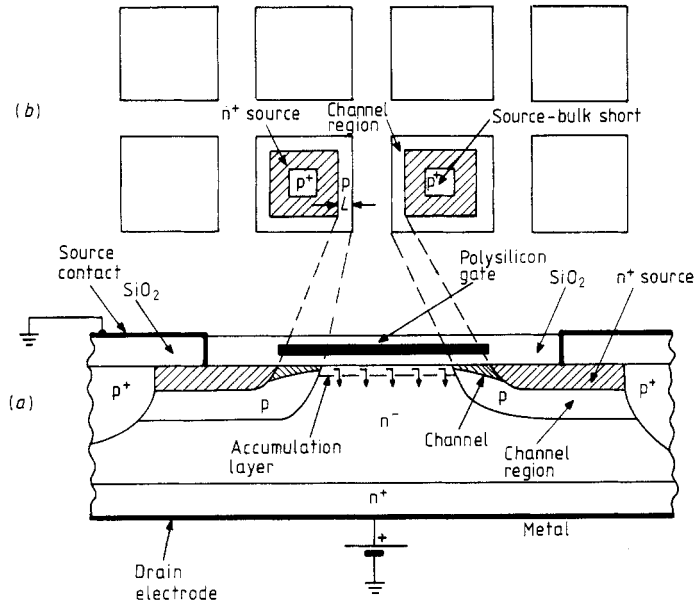


**Figure 18.** (*a*) Schematic cross section of a multichannel MOSFET with vertical $n^-$-drain drift region. (*b*) Schematic representation of the multiple cell geometry.

linearly dependent on the drain and gate voltages and is given by

$$I_D \simeq G_0(V_G - V_T)V_D \tag{2.15}$$

with

$$G_0 = \frac{Z}{L}\mu_n \frac{\epsilon_i}{x_i}. \tag{2.16}$$

$Z$ is the channel width, $L$ is the channel length, $\mu_n$ is the electron mobility at the surface, $x_i$ is the insulator thickness under the gate and $\epsilon_i$ is the insulator permittivity. The threshold voltage $V_T$ is the gate voltage necessary to obtain strong inversion. Its magnitude depends on the nature of the metal used for the gate, channel length, various surface charges, etc. In power devices, the gate conductor is most often highly doped polycrystalline silicon, which lends itself to the fabrication of the self-aligned gates.

The MOSFET $I$–$V$ characteristics (figure 11(*a*)) exhibit a linear portion at low currents and become saturated at the drain voltage $V_{Dsat}$. Beyond this point, the drain current increases only very slightly with the drain voltage. If the drain voltage is made high enough, the drain junction eventually undergoes an electrical breakdown.

A very approximate analysis gives for the saturation current (see, for example, Blicher 1981)

$$I_{Dsat} \simeq \tfrac{1}{2}G_0(V_G - V_T)^2 \tag{2.17}$$

with

$$V_G - V_T \simeq V_{Dsat}. \tag{2.18}$$

$V_{Dsat}$ is constant for a given gate voltage. The channel resistance is only very slightly affected by the extension of the space-charge region into the channel; consequently, the saturation drain current is almost independent of the drain voltage. However, in very short channels—of the order of a few micrometres—the depletion region affects appreciably the channel length. If the channel region (also known as the body or substrate) is very lightly doped, the drain depletion layer is wide even at the small drain voltages. In this situation, current carriers may be injected from the source over a potential barrier into the drain depletion region (Ohmi 1979). The modulation of barrier height by the drain electrostatic field results in the non-saturating triode-like $I$–$V$ characteristics, with the drain current exponentially dependent on the drain voltage (figure 11($b$)).

The MOSFET transconductance $g_m$ is defined the same way as that of the JFET. In the linear region

$$g_m = G_0 V_D \tag{2.19}$$

and in the saturation region

$$g_{msat} \simeq G_0(V_G - V_T). \tag{2.20}$$

The transconductance is larger for smaller channel lengths and larger channel widths.

The MOSFET structure of figure 18 is known as the vertical D-MOS transistor. D stands for the double-diffusion process. First the p-channel region (bulk) is diffused into the $n^-$-drain region; then the $n^+$-source is diffused. This way, it is possible to accurately define the very small lateral spacing between the two diffusions necessary for small channel lengths. The vertical structure allows excellent silicon 'real estate' utilisation and effective device heat sinking. High current capability is achieved by making the channel width $Z$ very large by using a multiplicity of elementary MOSFET (cells) (figure 18).

Another important power transistor structure is the vertical VMOSFET (figure 19) (Salama and Oakes 1978). The channel region is established here by double diffusion of the p-channel region and the $n^+$-source with subsequent preferential etching of a V-groove on ⟨100⟩ oriented surface. The V-groove defines the gate region. A power device contains a multiplicity of such structures on the same silicon chip.

If the groove etching is stopped at the right moment then the groove is shaped like the letter U, whose flat bottom helps to reduce the device series resistance.

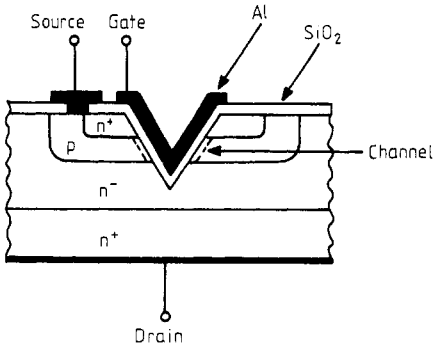


**Figure 19.** V-groove MOSFET (from Salama and Oakes 1978).

## 3. Avalanche breakdown

### *3.1. Avalanche breakdown mechanism*

The electrical breakdown of HV pn junctions takes place by impact ionisation (avalanche breakdown). The free carriers present in the junction depletion region acquire enough energy at the fields higher than $10^5$ V cm$^{-1}$ to break the silicon crystal covalent band, liberating more free carriers, which in turn create more electron–hole pairs and so on. This positive feedback mechanism results in an avalanche of holes and electrons and in the junction electric breakdown. The ratio of the output current to the input current entering the avalanche multiplication region is given by the multiplication factor $M$ which at junction breakdown goes to infinity. If the reverse junction potential is $V_R$ and the junction breakdown voltage is $BV$ (Moll *et al* 1970) then the following empirical relationship exists for the multiplication factor:

$$M = \left[1 - \left(\frac{V_R}{fBV}\right)^n\right]^{-1} \tag{3.1}$$

where $f$ is a parameter depending on the type of the junction impurity profile, the carrier type and the breakdown voltage. At about 1 kV and for a step junction $f \simeq 0.65$ for electrons. The exponent $n$ is equal to about 6 under the same conditions.

The pn junction breakdown voltage depends on its impurity profile. The breakdown voltage of an abrupt, *plane*, one-dimensional silicon junction with one side heavily doped is (Fulop 1967)

$$BV_1 \simeq 5.3 \times 10^{13} N^{-3/4} \text{ V} \tag{3.2}$$

where $N$ is the impurity concentration of the lightly doped side.

The breakdown voltage of a linearly graded, unidimensional junction is given by

$$BV_2 \simeq 8.93 \times 10^9 a^{-2/5} \text{ V} \tag{3.3}$$

where $a$ is the net impurity gradient. The avalanche breakdown voltage of the erfc diffused junction lies somewhere between the values given by (3.2) and (3.3).

Due to the high fields present at the corners, the breakdown voltage of diffused *planar* junctions is always lower than that of plane junctions. The planar junction with cylindrical or spherical corners breaks down at (Sze and Gibbons 1966)

$$BV_3 \simeq 6 \times 10^{13} N^{-3/4} \{[(n+1+\gamma)\gamma^n]^{1/(n+1)} - \gamma\} \tag{3.4}$$

where $\gamma$ is the ratio of the junction depth to the junction depletion layer width at breakdown of the one-sided *plane* abrupt junction. $n = 1$ and 2 for cylindrical and spherical junction, respectively. High breakdown voltages require very deep diffusions unless additional breakdown voltage improvement methods are employed.

### *3.2. Bipolar transistor breakdown voltage*

The collector-to-base breakdown voltage $BV_{CBO}$, with the emitter circuit open, is equal to the collector junction breakdown voltage. The lowest collector-to-emitter breakdown voltage is given by (Moll *et al* 1970)

$$BV_{CEO(sus)} = BV_{CBO} f (1 - \alpha_0)^{1/n} \tag{3.5}$$

where $\alpha_0$ is the peak current gain. For npn transistors $f$ is that for electrons since it is mainly the electrons traversing the collector depletion region that are multiplied. The collector-to-emitter breakdown voltage is considerably lower than $V_{CBO}$.

### 3.3. JFET and MOSFET breakdown voltage

In a JFET, the voltage breakdown normally occurs in the drain-gate depletion region where the field is the highest. The drain-gate breakdown voltage is given by

$$BV_D \simeq BV_j - V_G \tag{3.6}$$

where $BV_j$ is the drain junction avalanche breakdown voltage.

The drain-to-source breakdown voltage of a *long channel* n-MOSFET is essentially that of the drain junction. This is also true for p-channel devices, regardless of the channel length. n-MOSFET with short channels, of the order of a few micrometres, have the drain-to-source breakdown behaviour like that of npn transistors. The drain-to-source breakdown characteristics exhibit differential negative resistance regions similar to those obtained with bipolar transistors operated in the common-emitter configuration, because the main MOSFET is in parallel with an open base, parasitic, bipolar $n^+pn^+$ transistor. The source current undergoes avalanche multiplication in the drain depletion region and creates an excess current flowing laterally from the drain region to the source, creating a voltage drop in the device substrate. This voltage drop forward-biases the source-to-bulk junction which starts acting as the bipolar transistor emitter. The ensuing positive feedback mechanism leads to breakdown characteristics similar to that of open-base transistors. The immunity of the long channel n-MOSFET to the second breakdown (§ 9.2.3) is due to the low current gain of the parasitic device.

The gate oxide breakdown is another important mechanism limiting MOSFET performance. It depends on many factors such as oxide thickness, thermal treatment, temperature, etc. At room temperature, the electric field at breakdown is very approximately equal to 8 MV $cm^{-1}$. Under circumstances that allow metal gate vaporisation, the $SiO_2$ breakdown can become self-healing.

## 4. Improvement methods of breakdown voltage rating

### 4.1. Collector and drain structures

In devices operating above 200 V the requisite breakdown voltage is achieved by adding a high-resistivity $n^-$-drift region to the collector or the drain. Further improvement may be achieved by adding a lightly doped epitaxially grown $p^-$-region sandwiched between the p-base and the $n^-$-collector region. Such an arrangement is known as the $n^+$–p–$\pi$–$\nu$–$n^+$ structure (Blicher and Czorny 1969), where $\pi$ and $\nu$ stand for the lightly doped p and n regions, respectively. In this case, the depletion layer spreads into both the $\nu$ and $\pi$ region and the maximum junction field is lowered.

### 4.2. Junction contouring

The junction breakdown voltage can be considerably increased by the proper contouring of side surfaces. For devices operating below about 500 V, deep moat etching

(figure 20) sufficiently lowers the surface field avoiding surface breakdown, which always occurs at a lower voltage than the breakdown of the bulk silicon. Moat etching for voltages greater than 500 V is, however, much too erratic to be useful. Mechanical bevelling, used mainly for larger devices, guarantees more uniform and better results. By convention (Davies and Gentry 1969) the bevel angle is positive if the junction area decreases when moving from the more heavily doped to the less-doped junction side (figure 21(*a*)). It is negative as the area increases (figure 21(*b*)).

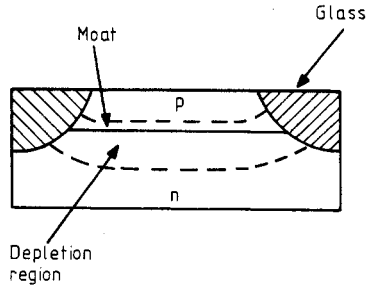


**Figure 20.** Glass-filled moat for breakdown voltage improvement.

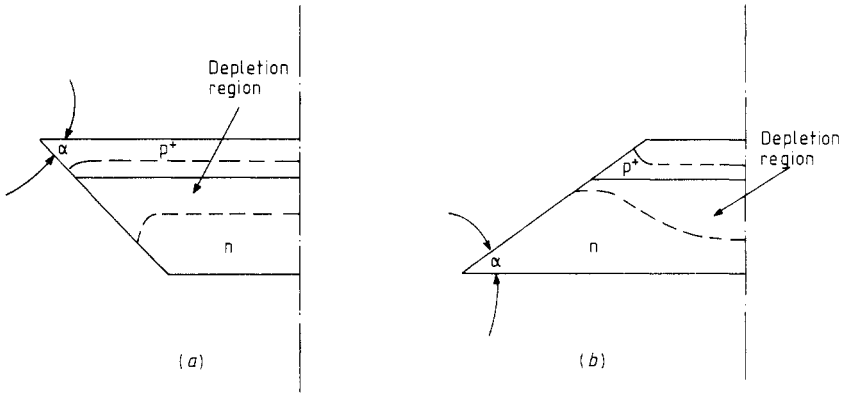


**Figure 21.** (*a*) Positive and (*b*) negative bevel angle convention.

*4.2.1. Positive bevel.* Positively bevelled junctions can achieve maximum possible bulk breakdown voltages. The surface field for the positive bevel is reduced below that in the bulk because at the surface, due to the junction shape, the depletion region expands on the lightly doped junction side and contracts on the heavily doped side. Consequently the breakdown will occur in the bulk, where higher electric fields are acceptable. A device with 45° positive bevel angle will have the peak surface field reduced to about 45% of the maximum bulk field at breakdown (Davies and Gentry 1969).

*4.2.2. Negative bevel.* The peak surface field is also reduced by negative bevelling; however, much smaller bevelling angles are needed than for the positive bevelling. For a negative bevel the minimum distance between the depletion layer boundaries

occurs in the silicon interior (figure 21(*b*)) so that the maximum electric field is shifted from the surface to the device interior (Cornu 1973, Adler and Temple 1976, 1978). The smaller is the negative angle, the higher is the achieved breakdown. Unfortunately, small angles lead to poor silicon area utilisation and high cost. The dielectric constant of the junction protective coating and various oxide charges can substantially modify the device breakdown voltage with any type of bevelling.

### *4.3. Depletion and substrate etching*

Mechanical etching may be expensive; chemical etching is more acceptable. Temple and Adler (1976) increased junction breakdown voltage by chemically etching away (figure 22) some part of the highly doped region beyond the metal contact. This partial etching is applicable to both the plane and planar junctions and results, as in the negative bevelling, in the deeper penetration of the depletion region into the more heavily doped side of the junction. The results surpass those of negative bevelling but the method requires very accurate etch depth control.

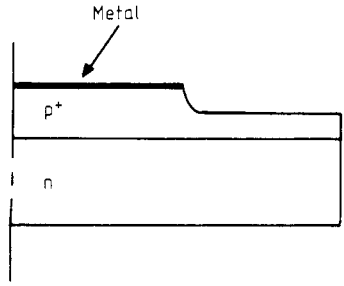


**Figure 22.** Depletion etching of a *plane* junction (from Temple and Adler 1976).

By etching away a portion of the lightly doped junction region (figure 23), it is possible to achieve high breakdown voltages—as high as with positive bevelling—both for the plane and planar junctions. This method is known as substrate etch termination (Temple and Adler 1977) and is less critical than the depletion etch method.

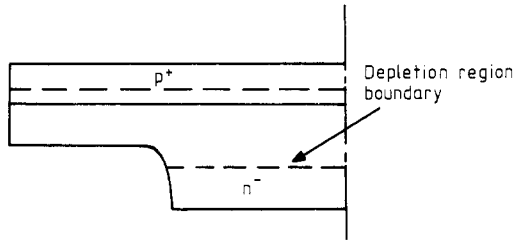


**Figure 23.** Substrate etch termination (from Temple and Adler 1977).

### *4.4. Depletion region control by ion implantation*

Shaping of the depletion region can be achieved by ion implantation (figure 24) (Temple 1977). This technique is applicable to the plane and planar junctions, and permits a much more precise charge control than the chemical etching. The achieved breakdowns have been close to 95% of the maximum possible.

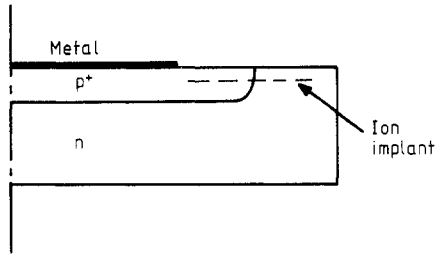


**Figure 24.** Depletion region control by ion implantation (from Temple 1977).

### *4.5. Field plates, guard rings and channel stoppers*

Planar junction breakdown voltage is considerably increased by the use of a metal field plate (figure 25). The plate extends beyond the junction region and remains at the same potential as the $n^+$-region. When reverse potential (positive for the $n^+$–p junction) is applied, the p-region, under the plate edge and the oxide, becomes at least partially depleted of holes for oxides thicker than about 100 nm. With thinner oxides the surface easily inverts and becomes n-type. Surface depletion increases the resistivity of the p-region at the surface and hence the electric breakdown voltage.

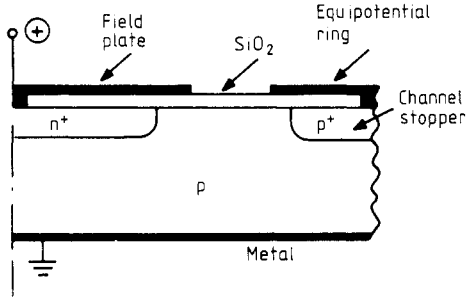


**Figure 25.** pn junction with a metal field plate, metal guard ring and channel stopper.

Rusu and Bulucea (1979) and Rusu *et al* (1980) analysed the two-dimensional behaviour of the field plate and found that the breakdown voltage of a planar junction provided with a field plate may be represented by the same formula as given for the cylindrical planar junction (expression (3.4)) but with

$$\gamma = \gamma' = (\epsilon_S/\epsilon_i)(x_i/x_d) \simeq 3x_i/x_d \tag{4.1}$$

where $\epsilon_S$ and $\epsilon_i$ are the permittivities of silicon and silicon dioxide, respectively, and $x_i$ and $x_d$ are the thicknesses of the oxide and the depletion region, respectively. Making the oxide thicker reduces the field crowding effect at the field plate edge and reduces the field more than an equal increment in the junction diffusion depth.

The field plate is usually used together (figure 25) with an equipotential ring surrounding the plate and electrically connected to the negatively biased substrate. The ring attracts holes nullifying the field plate's inverting effects, and conducts to ground the mobile positive charges accumulated on the oxide surface. The $n^+$-region under the ring improves the quality of the contact and also serves as an inversion layer (channel) stopper.

A very important variation of the field plate technique is the resistive plate formed by bridging the plate and the ring by a semi-insulating film such as semi-insulating oxygen-doped polycrystalline silicon (Sipos) (Matsushita *et al* 1976). The presence of the very slightly conducting film homogenises the electric field at the surface, avoiding field plate edge effects, improves the breakdown, suppresses arcing and stabilises (passivates) the surface. Figure 26 shows the Sipos film applied to a $p^+npp^+$ transistor. The collector reverse bias is also applied to the oxygen-doped Sipos film so that the n-base region surface is able to support high breakdown voltage. The silicon nitride $Si_3N_4$ film deposited on the top acts as a sealing agent against the penetration of moisture and highly mobile alkaline ions. The uppermost $SiO_2$ layer prevents the surface breakdown of the nitride layer.

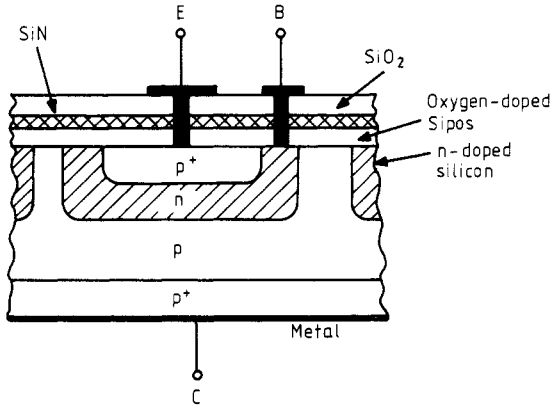


**Figure 26.** Passivating resistive plate (Sipos) applied to a p–n–p transistor (from Matsushita *et al* 1976).

Another very useful technique for avalanche breakdown voltage improvement consists of surrounding the main junction by one or more diffused concentric rings (figure 27) (Kao and Wolley 1967). These field-limiting rings, electrically floating, are of the same semiconductor polarity as the highly doped side of the main junction. The spacing between the first ring and the main junction is small enough to permit the two depletion regions to touch each other before avalanche breakdown takes place. The field limiting rings play a double role: they act as voltage dividers and at the same time increase the radius of curvature of the main junction depletion region, thus improving the breakdown voltage capability.

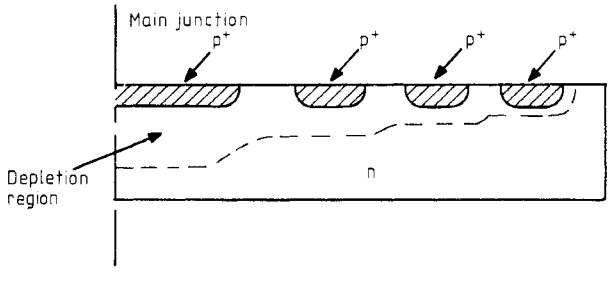


**Figure 27.** $p^+n$ junction with $p^+$ field-limiting rings (from Kao and Wolley 1967).

## 5. Bipolar power transistor current gain

### *5.1. Base widening*

One of the serious limitations of bipolar transistors is the great dependence of their current gain on the current density. In the low-power devices, this dependence is mainly due to current conductivity modulation in the base which lowers the emitter efficiency (Webster 1954). In power devices, particularly those which are provided with $n^-$-collector drift regions, the current gain drops rapidly with the appearance of base widening (Kirk 1962). At high currents, in the $n^+pn^-n^+$ transistor, the electron concentration in transit through the $n^-$-region becomes comparable to or larger than the $n^-$-region doping concentration. Assuming that the electrons in the collector depletion region reach the saturation velocity $v_s$, it is possible to write the Poisson equation in this region as

$$\frac{dE}{dx} = \frac{q}{\epsilon_s}\left(N_D - \frac{J_C}{qv_s}\right) \tag{5.1}$$

where $N_D$ is the $n^-$-region donor concentration and $J_C$ is the collector current density. The solution of this equation is

$$E(x) = E(0) + \frac{q}{\epsilon_s}\left(N_D - \frac{J_C}{qv_s}\right)x. \tag{5.2}$$

$E(0)$ is the field at $x = 0$ at the base edge of the depletion region. The field distribution in the depletion region is, therefore, linearly dependent on the position $x$ with a slope dependent on the current density. For $J_C = qv_sN_D = J_0$ the slope is zero and the field is uniform. For greater currents (figure 28) (Whittier and Tremere 1969) the slope

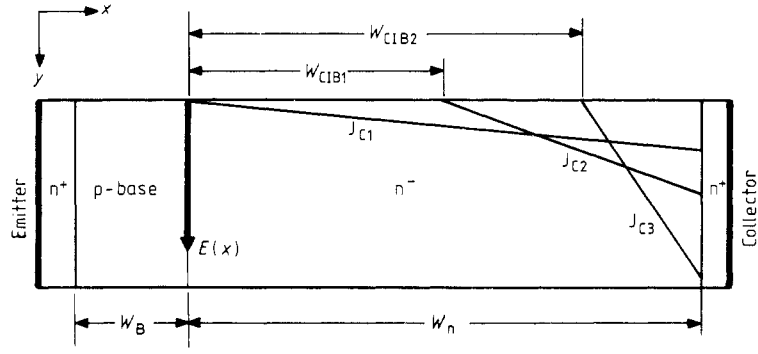


**Figure 28.** Field distribution in the $n^-$-drift region of an $n^+pn^-n^+$ transistor as a function of the collector current density (high-field case). The extended base $W_{CIB}$ is wider for higher current densities (from Whittier and Tremere 1969).

changes its sign and the maximum field shifts away from the p–$n^-$ junction and eventually occurs at the $n^-$–$n^+$ interface. Part or all of the $n^-$-drift region becomes field-free, i.e. neutral, and can therefore be considered to be an extension of the neutral p-base. The current-induced base width $W_{CIB}$ has a maximum value equal to the drift region width $W_n$. The case described above is known as high-field base widening.

When a power transistor operates as a switch, the collector junction voltage in the turn-on state becomes forward-biased, the depletion layer collapses and the device operates in saturation. As a result, the p-base starts injecting holes into the $n^-$ region (figure 29) which are neutralised by electrons coming from the $n^+$-emitter and the $n^+$-collector. Part or all of the $n^-$-region becomes conductivity modulated and neutral, i.e. becomes an extension of the p-base (low-field base widening). The total effective

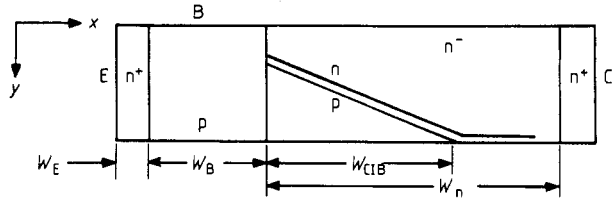


**Figure 29.** Mobile charge distribution in the $n^-$ collector drift region for the low-field case.

base width is therefore

$$W_{\text{eff}} = W_{\text{B}} + W_{\text{CIB}}. \tag{5.3}$$

Van der Ziel and Agourdis (1966) proposed a two-dimensional model for base widening, assuming that the emitter current spreads laterally in the base before reaching the collector region. A more complete base widening analysis would probably show that both effects might take place under some circumstances.

The widening of the base decreases the transistor transport factor $\alpha_{\text{T}}$, the emitter efficiency and lowers the frequency capability.

### *5.2. Emitter crowding*

The base current $I_{\text{B}}$ flowing laterally beneath the emitter (figure 30) causes a lateral voltage drop which makes the emitter centre less forward-biased than its edges so that the current crowds near the periphery with a density sufficient to cause the p-base to widen. Current crowding was analysed by Fletcher (1955) for a semi-infinitely long emitter and by Hauser (1964) for an emitter of finite length. A very simplified analysis

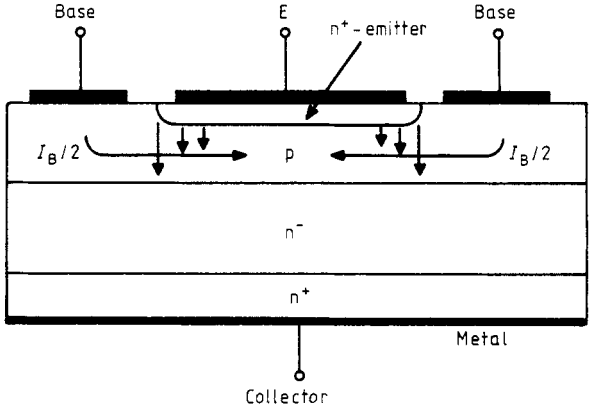


**Figure 30.** Current crowding at the emitter edges.

(Olmstead *et al* 1971) shows that the emitter current density along the emitter decays exponentially from the edge toward the centre with a characteristic length

$$y_0 = KW_B h_{FE}^{1/2} \tag{5.4}$$

where $K$ is a constant. Here, $y_0$ determines that part of the emitter which carries most of the current.

More recent analysis (Hower and Einthoven 1978) makes use of the integral charge control model (Gummel and Poon 1970) to take account of base widening effects in power transistors with $n^-$-drift regions. All the analyses stress the need for narrow emitter fingers interlaced with the base metal contacts (figure 31) to minimise the current crowding which affects severely the current gain and may result in the device second breakdown (§ 9.2).

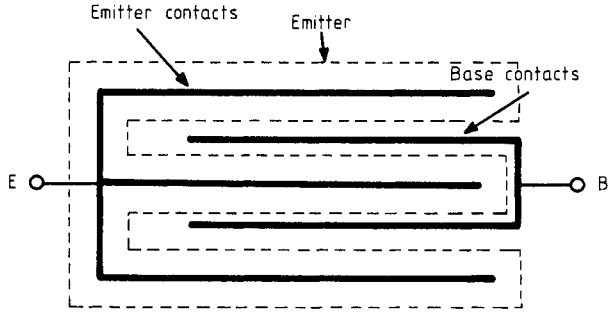


**Figure 31.** Schematic representation of the emitter and base interlacing.

### *5.3. Emitter efficiency*

The emitter efficiency in the common-emitter configuration is given by expression (2.6). At high emitter impurity concentrations ($N_D > 10^{18}$ cm$^{-3}$), the energy gap of silicon narrows (Mahan 1980) due to several effects such as impurity energy-level banding, random impurity distribution, free-carrier interaction, electron–donor interaction, etc. There is a wide disagreement about the most significant band-gap narrowing mechanism and about the amount of the change. A reasonable figure (Possin *et al* 1980) seems to be $\Delta E_g = 100$ meV at the carrier concentration of $5 \times 10^{19}$ cm$^{-3}$. Band-gap narrowing results in an increased intrinsic carrier concentration, which at low concentrations is equal to $n_i$, but at high concentrations becomes

$$n_{ie}^2 = n_i^2 \exp(\Delta E_g / kT). \tag{5.5}$$

For silicon $n_i^2 = 2 \times 10^{20}$ cm$^{-3}$. The increase in the intrinsic carrier concentration changes the emitter charge per unit area to (Mertens *et al* 1973)

$$Q_E = \int_0^{W_E} N_D (n_i / n_{ie})^2 \, dx \tag{5.6}$$

where $W_E$ is the active emitter depth. Thus, the effective emitter charge becomes considerably smaller at higher carrier concentrations and leads to a lower emitter efficiency (expression (2.6)).

At large carrier concentrations the carrier diffusivities and mobilities are substantially reduced due to carrier–carrier scattering effects. At carrier concentrations $n$ greater than about $10^{18}$ cm$^{-3}$ the hole and electron diffusion constants become almost

equal and approximately follow the relationship (Howard and Johnson 1965)

$$D \propto n^{-1/2}. \tag{5.7}$$

This decreases the effective emitter width $W_E$ and lowers the emitter efficiency as well. High carrier concentration also leads to the reduction of the minority carrier lifetime and diffusion length. At concentrations above $10^{17}$ cm$^{-3}$ the carrier recombination takes place both by the Shockley–Read–Hall (SRH) mechanism, involving recombination centres, as well as by Auger recombination. In Auger recombination, when a hole and an electron recombine the liberated energy is imparted to a third charge carrier—either a hole or an electron.

There is not complete agreement as to the relative importance of these various effects in determining the emitter efficiency (Sheng 1975, Adler *et al* 1976, McGrath and Navon 1977, Amantea 1980). There is a consensus, however, that the band-gap narrowing mainly affects shallow emitters 1–3 $\mu$m deep, whose efficiencies drop for surface concentrations exceeding $5 \times 10^{19}$ cm$^{-3}$.

### *5.4. Current gain fall-off*

Collector current gain is plotted as a function of current in figure 32. The gain at first increases, reaches a plateau, then starts falling off with a slope equal to $-1$ since

$$h_{FE} \propto W_{B1}^{-2} I_C^{-1} \tag{5.8}$$

where $W_{B1}$ is the total effective base width. Further increase in the current density results in severe current crowding and causes the gain to decrease more rapidly with a slope of $-2$ (Olmstead *et al* 1971, Hower and Einthoven 1978), since the gain then obeys the relationship

$$h_{FE} \propto W_{B1}^{-2} I_C^{-2}. \tag{5.9}$$

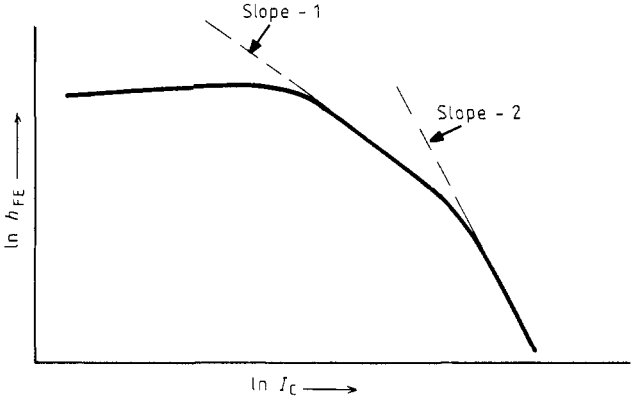


**Figure 32.** ln $h_{FE}$ plotted against the ln of the collector current $I_C$.

## 6. Methods of gain improvement

### *6.1. Bipolar devices*

One approach to improving bipolar device current gain is to use highly interlaced narrow emitter fingers. In microwave transistors, this approach is pushed to the limits

to ameliorate high-frequency current gain. One of the structures used for microwaves, the overlay transistor (Carley *et al* 1965), is depicted in figure 33. It uses a large number of discrete emitter sites paralleled by a thin overlayed metallisation. The $p^+$-base regions carry the base currents. For uniform current distribution, each emitter site is provided with a ballasting resistor, consisting of resistive polycrystalline silicon. Optimum power output is obtained by maximising the ratio of emitter periphery to area, and also the ratio of emitter periphery to base area.

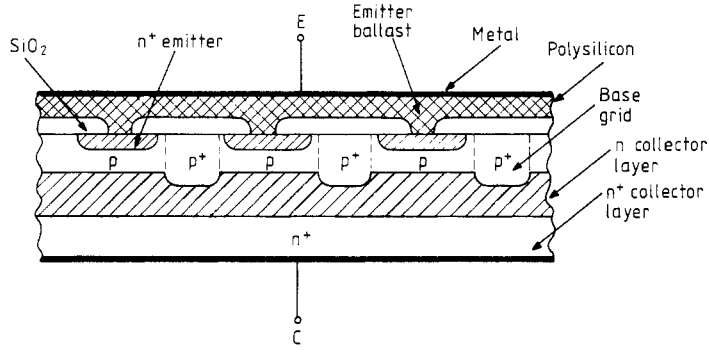


**Figure 33.** Overlay microwave power transistor with multiple emitters and ballasting resistors.

The emitter efficiency may be improved (Yagi *et al* 1974) by using an emitter consisting of a highly doped $n^+$-region and lightly doped $n^-$-region forming an $n^+n^-pn^-n^+$ transistor. The holes injected by the p-base into the $n^-$-emitter cannot reach the emitter contact as they are reflected away from the $n^+$–$n^-$ high–low barrier. This minimises the base-injected hole current and improves the emitter efficiency.

The use of arsenic in place of the more commonly used phosphorus gives steeper diffusion gradients and small band-gap narrowing (Fair 1973, Lillard *et al* 1973) leading to improved emitter efficiencies. Also avoided is so-called emitter push-out, which consists of pushing the base region ahead of the emitter diffusion. This is particularly troublesome for RF devices with narrow bases.

Excellent emitter efficiencies are possible by making the emitter from a semiconductor with a wider energy gap than that of the base. In such heterojunction structures the flow of the base current injected into the emitter is almost completely stopped and the emitter efficiency is unity. The realisation of such an emitter presents serious difficulties due to the presence of a large number of recombination centres at the interface of the two semiconductors. Recently Oh-uchi *et al* (1979) reported successful use of an oxygen- and phosphorus-doped polycrystalline layer (Sipos) as an emitter. According to Oh-uchi *et al*, highly phosphorus-doped Sipos has an energy gap of 1.5 eV, whereas that of silicon is 1.1 eV at room temperature.

A GaAs bipolar transistor with GaAlAs emitter (Bailbey *et al* 1980), developed for UHF applications, is another example of a heterojunction device capable of high current gains at very high current densities ($3 \times 10^3$ A cm$^{-2}$).

### *6.2. Field-effect transistors*

The gain factor $G_0$ largely determines the maximum allowable current, transconductance and power output of a field-effect transistor.

Large channel widths $Z$, small channel lengths $L$ and thin gate oxides make $G_0$ large. Channel widths in MOSFET devices operating at high currents may reach tens of cm. Typically the current per cm of channel width is of the order of 0.5–1.0 A $cm^{-1}$.

In contrast to bipolar transistors, no ballasting resistors are necessary to equalise the elementary cell gate currents because of the extremely high DC input resistances.

Channel lengths of the order of 3–5 $\mu$m are required to obtain adequate transconductances and gain bandwidths. Shorter channel lengths are feasible, although some new effects, such as hot electron injection into the oxide, may, for instance, alter the threshold voltage and cause instabilities.

## 7. Voltage drop in power devices

All power devices should exhibit low power dissipation in order to keep their area and cost to a minimum. This leads to the requirement of low voltage drop in the device during switching and in the on state.

### *7.1. On-state voltage drop of bipolar transistors*

When an $n^+pn^+$ transistor is turned on, both the emitter and the collector are forward-biased, i.e. the device is in saturation (Ebers and Moll 1954) and exhibits minimum resistance to the current flow. In the $n^+pn^-n^+$ transistors the voltage drop is higher because there is a quasi-saturation region, due to base widening. The voltage drop is the algebraic sum of the base–emitter voltage, the drop across the p–$n^-$ junction, the drop across the current-unmodulated region $W_n - W_{CIB}$ (figure 29) and the drop in the extended base (Chudobiak 1970). In most cases the voltage in the unmodulated $n^-$-region dominates over other drops so that the total voltage drop is

$$V_{CE(sat)} \simeq J_c \rho_c (W_n - W_{CIB}) \tag{7.1}$$

where $\rho_c$ is the $n^-$-region resistivity. Using charge control equations, it is possible to more accurately compute the total voltage drop (Hower 1973, 1976) for devices with narrow emitter stripes.

### *7.2. On resistance of MOSFET*

The series device resistance in a MOSFET in the on condition is the sum of the active channel resistance and the drain structure resistances. The channel resistance, inversely proportional to the positive gate voltage, is small compared to the drain resistance. In the structure with a vertical drain the accumulation layer at the drain top surface (figure 18) distributes the current flow more uniformly and improves the drain region voltage drop. The series resistances of various MOSFET structures were computed by Sun and Plummer (1980).

### *7.3. Thyristor on-state voltage drop*

In the forward conducting state all three SCR junctions are forward-biased and the total voltage drop is the algebraic sum of the voltage drops across the $J_1$, $J_2$ and $J_3$ junctions and the voltage drops in all other regions of the $n^+pn^-pn^+$ structure. For

simplicity, we can assume that the only significant voltage drop outside the three junctions is in the n-base which is usually very long in order to support the requisite anode potential. Consequently, the voltage drop across the thyristor is given by

$$V_T = V_1 - V_2 + V_3 + V_n \tag{7.2}$$

where $V_n$ is the drop in the n-base. If we further assume that the device is symmetrical so that the thicknesses and doping levels of the two p-regions are the same, then expression (7.2) reduces to

$$V_T = V_3 + V_n \tag{7.3}$$

with

$$V_3 = \frac{kT}{q} \ln\left(\frac{I_A}{I_s}\right) \tag{7.4}$$

where $I_s$ is the saturation current of the $J_3$ junction. The voltage drop across the n-base region was studied by several authors (Herlet and Raithel 1966, Kokosa 1967, Otsuka 1967, Burtscher *et al* 1975, Adler 1978, Hayashi *et al* 1981). There are rather big discrepancies among the proposed voltage drop expressions. It is possible, however, to generalise the results by expressing the voltage drop in the n-base at high current densities as

$$V_n = f(W_1/L)J^{1/2} \tag{7.5}$$

where $f(W_1/L)$ is a sensitive function of the ratio of the n-base width $W_1$ to the ambipolar carrier diffusion length $L$ at high current densities. Long n-bases and short diffusion lengths rapidly increase the n-base voltage drop.

## 8. Dynamic behaviour of power devices

### *8.1. Gain–bandwidth product ($f_T$)*

The gain–bandwidth product $f_T$, i.e. the frequency at which the common-emitter current gain drops to unity, largely determines the switching speed or the frequency response of a transistor. It is related to the total emitter-to-collector delay time $\tau_{EC}$ by

$$f_T = \frac{1}{2\pi\tau_{EC}}. \tag{8.1}$$

$\tau_{EC}$ is composed of the emitter and collector capacitance charging times, the base transit time and the collector depletion region transit time. At high current levels base widening *causes the base transit time to greatly increase* from the value of $W_B/2D_n$ at low levels to

$$\tau_B = \frac{W_B + W_{CIB}}{4D_n}. \tag{8.2}$$

The factor of 4 in the denominator is due to the base current conductivity modulation effect which doubles the diffusion constant (Webster 1954).

In the MOSFET or JFET, $f_T$ depends on the total electron transit time between the source and the drain. There is no channel widening effect. There exists, however, a

channel shortening effect due to the presence of the drain depletion region. This effect, primarily important for very short channels, decreases the transit time since in the space-charge region, the carriers move with the limiting velocity $v_s \simeq 10^7 \text{cm s}^{-1}$.

### *8.2. Bipolar transistor switching*

In bipolar transistors with resistive load, the switching transient can be described by the *idealised* waveforms of figure 34. The upper waveform depicts the base turn-on and turn-off current pulses. The lower waveform represents the resulting (output) collector current. The idealised waveforms are mainly helpful in understanding the manufacturers' data regarding various collector current time delays.

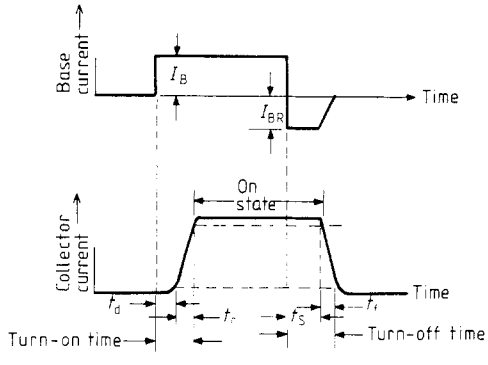


**Figure 34.** Bipolar power transistor switching transient $t_d$, $t_r$, $t_S$ and $t_f$ are the delay time, rise time, storage time and fall time, respectively.

In order to turn the transistor on, a current $I_B$ is applied to the base. The collector current starts appreciably flowing only after a delay time $t_d$. This amount of time is necessary to charge the emitter and the collector junction transition capacitances and to allow the emitted minority carriers to reach the collector. This time interval ends by convention when the collector current reaches 10% of its final value. The rise time $t_r$ is the time necessary for the collector current to reach 90% of the final value. The total turn-on time is the sum of $t_d$ and $t_r$.

In order to turn the transistor off, a reverse current $I_{BR}$ is applied to the base. Because of electric charges stored in various transistor regions, the collector current continues flowing unchanged for a time $t_S$. Upon the removal of most of these charges, the collector current starts falling off and becomes equal to 10% of the maximum collector current after a time interval $t_f$ known as fall-off time. The total turn-off time is the sum of $t_S$ and $t_f$. All of these various time intervals can be calculated comparatively easily (Moll 1954) for small-signal transistors, using the linear differential equations of the charge-control analysis. This type of transistor analysis ties the base and the emitter currents to the base charge. For power transistors, particularly those with $n^-$ high-resistivity regions, the problem becomes much more complex because the transistor current gain, due to the base widening effect, changes much more drastically with current level than in small-signal devices. This makes the charge-control equations non-linear so that the simplicity of the solution of linear equations is lost. To overcome these difficulties, the collector current rise time of a power transistor can be divided into two time intervals (Hower 1976): the initial one with

the device operating in the active region of the $I$–$V$ characteristics, and another one when the transistor reaches deep saturation, i.e. the collector becomes heavily forward-biased, starts emitting electrons (a $n^+pn^-n^+$ transistor is considered here) and the transistor p-base widens. In this approach, the collector rise time is described by *two linear differential equations* which can be easily solved. The analysis shows that, prior to the base widening, when the device is in the active region, the collector current rises with the time constant

$$\tau_1 = h_{FEO}\tau_{B1} \tag{8.3}$$

where $h_{FEO}$ is the peak value of the common-emitter current gain and $\tau_{b1}$ is the minority-carrier transit time in the device's metallurgical base. During the first time interval, the collector current can be expressed by

$$I_{C1}(t) = h_{FEO}I_B[1 - \exp(-t/\tau_1)] \tag{8.4}$$

where $I_B$ is the turn-on base current and $t$ denotes time. As the current further increases beyond some critical value, the base starts widening and the current-induced base width during the rise time varies from zero to its maximum possible value $W_{CIB}$. Neglecting the comparatively small width of the metallurgical base, it is possible to introduce an effective transit time $\tau_{B2}$ for the minority carriers traversing the current-induced base. This effective transit time is obviously shorter than the transit time through the fully current-induced base width $W_{CIB}$.

In the second region, the current rise can be expressed by

$$I_{C2}(t) = h_{FE}I_B\{1 - \exp[-(t + t_a)/\tau_{B2}]\}. \tag{8.5}$$

In this expression, $h_{FE}$ is the DC current gain at high current levels with the base fully widened and $t_a$ is a factor introduced to ensure the continuity of transition from the first to the second time interval.

Equations (8.4) and (8.5) permit the calculation of the total time necessary for the collector current to rise from 10% to 90% of its final value.

The turn-off transient of an $n^+pn^-n^+$ power transistor, because of its complexity, has not been fully studied from the physical point of view and the existing numerical modelling techniques are not satisfactory from the device design point of view. To turn-off an $n^+pn^-n^+$ transistor from the conducting state, i.e. from deep saturation, it is necessary to remove all the charges in excess of the equilibrium charges stored in the various device regions (figure 35). This can be accomplished, for example, by the application of a reverse base current pulse $I_{BR}$. The problem of charge removal

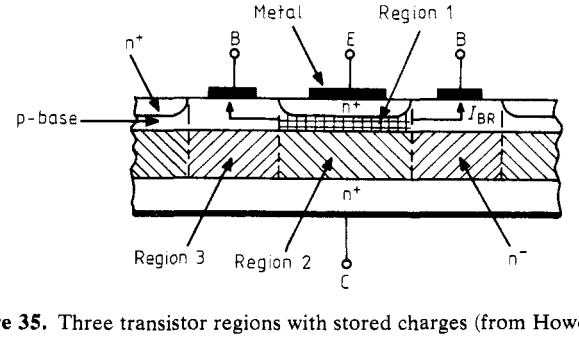


**Figure 35.** Three transistor regions with stored charges (from Hower 1980).

is two-dimensional and is further complicated by the appearance of a lateral voltage drop in the base caused by $I_{BR}$. Due to this effect, the emitter current concentrates in the emitter and base central regions. The charge withdrawal time is thus further increased. During the turn off (Hower 1980), the charge is first removed from region 3 ('remote' base), the closest to the base metal contact. After that, the charge is removed from region 2, i.e. from the current-induced base, and finally from the metallurgical base (region 1). In the most recent, but still rather incomplete model of Hower (1980), the charge moves toward the base metal contact with a lateral velocity $v_L(t)$. Since the base current is the product of the net excess charge its velocity $v_L(t)$ increases as the charge decreases with the lapse of time in order to keep the reverse base current constant.

During the storage time interval, the collector current remains constant and is determined mainly by the resistive collector circuit. The total storage time $t_S$ is the sum (Hower 1980) of the time $t_{S,RB}$ required to withdraw the charge from the 'remote' base and $t_{S,CIB}$ the time required to remove the charge from the current-induced base (CIB). When the transistor comes out of saturation, the collector depletion region is again formed and can block the reverse collector potential.

The explicit expressions of $t_{S,RB}$ and $t_{S,CIB}$ can be determined on the basis of Hower's model. The general approach for this determination is suggested in Hower's paper. Going into these details is beyond the scope of this review. It will suffice to mention, however, that $t_S$ can be reduced by reducing minority carrier lifetime in the base and in the collector high-resistivity region, and by reducing the carrier transit time. Short saturation time intervals are required, for example, for fast switching. Short lifetimes are also beneficial for the reduction of the fall time $t_f$. The requirement of short minority-carrier lifetime unfortunately is *in conflict* with the requirement of reasonably high transistor current gain. In most cases, long minority carrier lifetime is required and bipolar transistor storage and fall times are consequently also long, of the order of several microseconds.

### *8.3. Thyristor switching*

The physics of a thyristor turn-on are still not fully understood. The gate triggering pulse initiates a current flow from the cathode edge periphery close to the gate contact. After a short delay phase, the carriers (electrons) reach the anode junction and bias it in the forward direction so that a hole current starts flowing. When the sum of the two current gains becomes equal to unity the current rises very rapidly (Cornu and Jaecklin 1975, Jaecklin 1976). The fast current rise may also take place when the depletion region of the centre junction $J_2$ widens under the current flow condition so that it reaches the anode (Danhäuser and Voss 1976). The rise time ends when the anode current reaches about 5 A cm$^{-1}$ of the cathode perimenter. Further current increase takes place by the spreading of the hole and electron plasma toward the emitter centre. This lateral turn-on takes place by diffusion according to Longini and Melngailis (1963) and Dodson and Longine (1966) with a velocity proportional to $J_A^{1/n}$ ($J_A$ is the anode current density and $n = 2$–4). On the other hand, Ruhl (1970) attributes the plasma spreading to the lateral electric field in the p-base. In his model, the lateral velocity is given by

$$v_L = A \ln J_A + B \tag{8.6}$$

where $A$ and $B$ are constants.

A more recent two-dimensional analysis (Adler and Temple 1980) shows that plasma spreading is not due to carrier diffusion, but rather starts when the p-base *under the $n^+$-emitter periphery* becomes conductivity-modulated. The size $X$ of the active region obeys

$$\frac{d}{dt}\ln(X) = J_A^{1/m} \tag{8.7}$$

with $m = 1$ or 2 for the non-shorted and shorted emitter, respectively. Typical values of the spreading velocity range from $2.5\times10^3 \text{cm s}^{-1}$ to $9\times10^3 \text{cm s}^{-1}$ for current densities of 40–1000 A $\text{cm}^{-2}$.

If the current rise rate $di/dt$ is too small, the temperature of the initial turn-on region may reach destructive levels. Consequently, the initial turn-on area should be made as large as possible and $v_L$ should be maximised. Interlacing of narrow cathode stripes with the gate metallisation stripes increases the initial injection region (Storm and St Clair 1974). Hard gate triggering also increases the neutral conducting area. Thyristors operating at current levels of hundreds of amperes normally use an additional amplifying gate thyristor integrated on the same silicon chip with the main thyristor (figure 36). The pilot thyristor drives the main device upon the application of a small current pulse to the pilot gate.

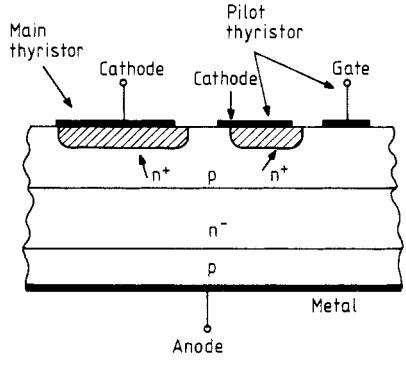


**Figure 36.** Thyristor with an amplifying gate (pilot thyristor).

*8.3.1. Undesirable thyristor switching.* Thermal junction currents or capacitive currents, due to the application of a fast rising anode potential, may provoke an undesirable device turn-on. These effects are significantly minimised by the use of the cathode shorts (figure 9) which collect the capacitive and thermal currents, thus preventing them from reaching the $n^+$-emitter and activating the thyristor.

### *8.4. Thyristor reverse recovery*

A thyristor may be turned off by either opening its anode circuit or by reducing the anode current below the holding level. To remove the charges stored in the device, a negative potential is usually applied to the anode. An idealised turn-off transient with a resisitive load would exhibit two storage time and two fall-off time regions (Sundresh 1967). The positively biased cathode rapidly removes excess electrons from the p-base, so the first storage time is short. Hole removal from the n-base by the

negative anode is much slower due to the long transit time and because holes are constantly being injected into the $n^-$-region by the heavily doped p-base. During the second fall-off period enough carriers are removed (or recombined) from the n-base so that the p–n anode junction recovers and is able to again support a high positive potential. Under more realistic conditions, the anode always contains some inductances and the current cannot be reversed instantaneously. As a result, the two distinct storage times are usually not observed (figure 37). The turn-off time $t_q$ (figure 38) is the time required for the stored charges to decay to a magnitude at which the anode forward voltage with the specified rising rate $dv/dt$ can be reapplied. This time is much longer than just the reverse recovery time and is usually equal to a few minority carrier lifetimes. A short current spike occurs when the anode voltage polarity is again made positive due to the withdrawal of the stored charge balance (Brewster and Schlegel 1974). This current spike may cause premature device turn-on.

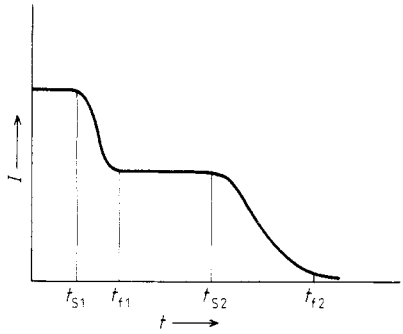


**Figure 37.** Idealised turn-off transient of a thyristor, $t_S$ and $t_f$ indicate the end of the storage and fall times, respectively (from Sundresh 1967).

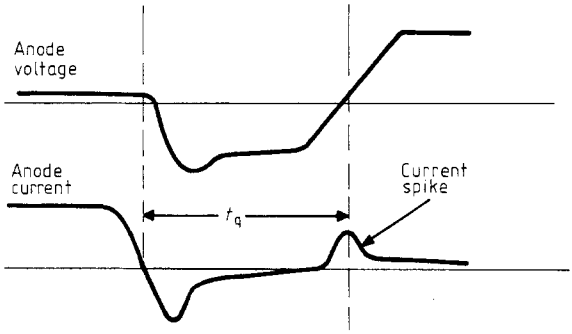


**Figure 38.** Reverse recovery voltage and current waveforms, and forward recovery current spike.

Thyristor switching time can be reduced by reducing (killing) the minority carrier lifetimes and by using gate-assisted turn-off (Schlegel 1976) with the p-gate made negative to remove the excess stored holes, thus reducing the current spike.

*8.4.1. Gate turn-off thyristor (GTO).* This device can be turned off by the application of a negative bias to the gate with the anode potential remaining positive. A negative gate collects holes which prevent their reaching the cathode, and pushes electrons

toward the cathode centre. The device current becomes focused in a narrow filament at the cathode centre (Wolley 1966). In order to avoid excessive current densities, such devices should be highly interlaced with very narrow emitter fingers. Minority carrier lifetimes should also be reduced to speed up turn-off and thus reduce transient thermal dissipation (Becke and Misra 1980).

### *8.5. Power MOSFET switching*

Because of the presence of long, high-resistivity collector drift regions in HV bipolar power transistors, the total minority carrier transit time is long and $f_T$ is small. As a result, the sum of the turn-on and turn-off switching times of these devices is, typically, of the order of several $\mu$s, so that the switching rates are below 100 kHz. At higher temperatures, the switching rate becomes considerably smaller because of the increase in the minority carrier lifetime and increased stored charges. The MOSFET switching transient with resistive load can be described by an idealised waveform similar to that of figure 34. Here, however, total switching times are considerably shorter since they depend on the *majority* carrier transit time. In the MOSFET, $t_s$ is not due to the minority carrier storage but to the charge stored in the input capacitance, roughly equal to $C_G$. MOSFET $t_s$ are typically considerably smaller than those of a bipolar transistor.

The total switching times of the HV MOSFET are of the order of tens of ns and their switching rates are of the order of several hundred kHz. Since there is no minority carrier storage in the MOSFET, switching times are unaffected by a temperature change.

A MOSFET turn-off transient may be seriously affected by the presence of a parasitic $n^+$–p–$n^+$ transistor. If the drain junction is in the avalanche breakdown state, the source-bulk junction may become forward-biased by holes flowing from the avalanche region and the MOSFET turn-off transient becomes that of the activated parasitic transistor.

## 9. Thermal properties and instabilities

### *9.1. Thermal runaway*

A power device's performance is seriously limited by its thermal properties and instabilities. Bipolar transistors usually can operate at junction temperatures of up to 200 °C; thyristors and MOSFET, up to about 150 °C. To obtain maximum power output, devices are mounted on heat-dissipating heat sinks.

The junction temperature rise $\Delta T$ can be determined from the product of the junction-to-ambient thermal impedance $R_{th}(t)$ and the dissipated power $P_T$ as

$$\Delta T = P_T R_{th}(t). \tag{9.1}$$

Junction thermal impedance consists of several components which include all thermal resistances and capacitances along the thermal path. Due to the presence of the thermal capacitances, the maximum steady-state thermal resistance is reached only after a time of the order of a few seconds for larger devices.

The cooling rate of a bipolar device should always be greater than its heating rate to avoid thermal regeneration and runaway. Thermal runaway does not take place in MOSFET devices because, as the temperature increases, the series drain resistance

also increases due to the decreased carrier mobility. Consequently, the drain current for the constant drain voltage is reduced as the temperature goes up.

### *9.2. Current constrictions*

Current constrictions or filaments may appear even in an ideal, defect-free semiconductor; their formation is greatly facilitated, however, when non-uniformities are present. Filamentary current flow occurs to minimise the entropy production rate and to maintain electrical stability (Ridley 1963). Small-diameter (few $\mu$m) current filaments known as microplasmas may grow at higher temperatures to large-sized hot spots (macroplasmas) covering large device areas and causing local heating and runaway.

*9.2.1. Forward second breakdown.* With the emitter forward-biased, current filaments may form, for example, at the emitter edges, where current density is highest and may result in a destructive instability known as forward second breakdown. The onset of this instability in one-dimensional transistors takes place when the stability factor

$$S = R_{\text{th}} V_{\text{CE}} \, \partial I_{\text{C}} / \partial T \tag{9.2}$$

becomes equal to or is greater than unity. $V_{CE}$ is the collector-to-emitter voltage (Scarlett and Shockley 1963, Scarlett *et al* 1963, Hower and Govil 1974).

Sometimes the stability factor exceeds unity but a *stable hot spot* develops because, for instance, the base within the filament drastically widens (Hower *et al* 1976). The stable hot-spot temperature is usually high enough to degrade the device and often leads to a destructive second breakdown.

*9.2.2. Reverse second breakdown.* During bipolar transistor turn-off the emitter-base junction is purposely reverse-biased and the base current flows laterally to the base contact (figure 35). Due to the resulting lateral voltage drop, the emitter centre becomes more forward-biased than its edges. Most of the emitter current is, therefore, concentrated at the centre. If the collector load is even slightly inductive, the collector junction may reach its avalanche breakdown, generating electron–hole pairs. These holes flow to the negative base contact, further magnifying the current focusing effect. Due to the high current density, the base widens and the maximum electric field shifts to the $n^-$–$n^+$ interface and the onset of avalanche breakdown becomes mainly dependent on the density of mobile charges. This phenomenon is known as the avalanche injection (Hower and Reddi 1970, Beatty *et al* 1976). The collector-to-emitter breakdown voltage drops to a lower value and continues decreasing with increasing current. Thus, the collector exhibits negative differential resistance and current is carried by both electrons and holes (double injection). The initiation of instability is, therefore, purely electrical. The unstable current filament that is formed is known as the reverse second breakdown. To avoid device degradation or destruction, power transistor manufacturers supply information about the safe operation areas (SOA) where forward and reverse second breakdown can be avoided.

*9.2.3. MOSFET second breakdown.* MOSFET are not immune to second breakdown because of the presence of a parasitic n–p–n transistor (Krishna 1977, Nagata 1977, Nakagiri and Ida 1977). During an inductive turn-off the device supports very high drain voltages in the presence of very high drain currents which lead to avalanche injection. The avalanching drain junction provides an excess p-substrate (bulk) current

that develops a forward voltage drop across the source (emitter)-bulk junction and establishes a positive feedback mechanism. The probability of second breakdown occurrence is greater for short n-channel devices. Second breakdown is not observed in p-channel devices, because the current gain of a p–n–p parasitic transistor is much smaller than that of an n–p–n transistor with the same geometry.

*9.2.4. Thyristor instabilities.* The transition region between the off state and the on state (figure 8) is characterised by a negative differential resistance. If the device remains there too long, a current filament may form, leading to device destruction. Hard thyristor driving with large gate currents is effective in preventing this phenomenon. In a GTO, the turn-off mechanism results in current filament formation and leads to an instability analogous to second breakdown.

*9.2.5. Improvement of second breakdown rating.* Multiple emitter sites with ballasting resistors help greatly in the improvement of the device forward second breakdown capability. Ballasting resistors in the base circuit are usually much less effective (Hower and Govil 1974).

An increase in the energy required to initiate reverse second breakdown may be achieved by incorporating an n-type collector buffer layer, thus forming an $n^+pn^-n^+n^+$ structure. The buffer doping concentration and thickness are adjusted to minimise the electric field at the $n\text{–}n^+$ interface and to improve the breakdown.

An emitter structure (Owyang and Shafer 1978) with a thin and lightly doped emitter central region avoids excessive current focusing during the device turn-off.

Second breakdown in a MOSFET is considerably improved by connecting together (shorting (figure 18)) the source and the substrate (bulk), thus substantially lowering the parasitic transistor current gain.

## 10. Surface stabilisation and passivation

To assure the stability of device parameters, the device surface must be perfectly clean after all processing is done, and then properly passivated to block the penetration of any external contaminants. Chlorine and chlorine compounds, such as HCl and trichloroethane, are often used for furnace quartz-tube cleaning and for clean silicon surface oxidation. To protect the device from moisture and external contaminants, devices are often packaged in hermetic packages. Many power devices are, however, first sealed in glass and are electrically tested before the final enclosure. Oxygen-doped Sipos layers covered with silicon nitride and glass provide excellent protection against contamination.

## 11. Conclusion

The performance of a semiconductor power device depends, in most cases, on its capability to support high voltage and to carry high current with minimal power loss. The achievement of breakdown voltages of up to a few thousand volts became a reality due to the inclusion of high-resistivity drift regions into bipolar and field-effect power transistor designs, accompanied by the use of such breakdown voltage improvement methods as field plates and depletion region control by ion implantation, etching

or mechanical shaping. The preservation of the device breakdown and its surface properties has been made possible by the application of novel and very effective passivation techniques.

In bipolar transistors, the requirement of high-voltage capability is in conflict with the requirement of a reasonable current gain and low voltage drop at high currents.

In thyristors and field-effect transistors, the high-voltage requirement opposes the requirement for small voltage drop in the highly conducting state.

Despite these constraints power transistors with 2 kV collector junctions can now be fabricated with very reasonable current gains at 10–20 A collector currents. HV bipolar thyristors that can switch 3000–4000 A and support current surges greater than 15 000 A are now realisable. Power MOSFET with drain junction ratings of 500–700 V and currents of up to 100 A are now feasible.

The presence of long, high-resistivity collector drift regions increases the total minority-carrier transit time and the total HV bipolar transistor switching time. Total switching time of field-effect transistors, on the other hand, is little affected by the presence of the lightly doped drain drift region. Consequently, HV field-effect transistors can be made faster than bipolar devices for the same current level.

At the low voltages (10–30 V) used for bipolar microwave transistors, the high-resistivity collector drift region is very short so that high-frequency response is possible even at high current densities. Above 4 GHz, however, the GaAs field-effect transistor (MESFET) becomes superior in HF performance to bipolar transistors. Power outputs of about one watt at 20 GHz are now obtainable. It appears that for GaAs MESFET, the maximum achievable microwave power output is inversely proportional to the square of the frequency (DiLorenzo and Wisseman 1979, Higashisaka *et al* 1980).

Thermal runaway and current instabilities substantially limit power device performance. Field-effect transistors are much less susceptible to, but not entirely free from, instabilities such as second breakdown, and are not affected by thermal runaway.

Good progress has been made in the last decade in the understanding and improvement of many power device properties such as breakdown voltage, current gain and surface stability. Much less progress has been made, however, in such areas as charge dynamics modelling. For most power devices, therefore, the semi-empirical approach still remains an important design technique.

## Acknowledgments

I wish to thank Dr H Kressel of RCA, David Sarnoff Research Center, Princeton, NJ for very useful suggestions, and my son A Peter Blicher (University of California, Berkeley) for critical reading of the manuscript.

Special thanks are due to Mr A A Rago for the skilful preparation of the illustrations and to Mrs M E Dixon for the expert typing of the manuscript.

## References

Adler M S 1978 *IEEE Trans. Electron Devices* **ED-25** 16–22

Adler M S, Beatty B A, Krishna S, Temple V A K and Torreno M L 1976 *IEEE Trans. Electron Devices* **ED-23** 858–63

Adler M S and Temple V A K 1976 *IEEE Trans. Electron Devices* **ED-23** 956–60

—— 1978 *IEEE Trans. Electron Devices* **ED-25** 1266–70

—— 1980 *IEEE Trans. Electron Devices* **ED-27** 483–502

Amantea R 1980 *IEEE Trans. Electron Devices* **ED-27** 1231–8
Bailbey J P, Marty A, Hiep P H and Rey G E 1980 *IEEE Trans. Electron Devices* **ED-27** 1160–4
Baliga B J 1981 *IEEE Trans. Electron Devices* **ED-28** 257–64
Beatty B A, Krishna S and Adler M S 1976 *IEEE Trans. Electron Devices* **ED-23** 851–63
Becke H W and Misra R P 1980 *IEDM Tech. Digest* 649–53
Blicher A 1976 *Thyristor Physics* (Berlin: Springer-Verlag)
—— 1981 *Field-Effect and Bipolar Power Transistor Physics* (New York: Academic)
Blicher A and Czorny B R 1969 *US Patent* 3427515
Brewster B J and Schlegel S E 1974 *IEEE Conf. Record, Industry Applications Society Meeting* 663–72
Burtscher J, Danhäuser F and Krausse J 1975 *Solid St. Electron.* **18** 35–63
Carley D R, McGeough P L and O'Brien J F 1965 *Electron.* **70** 38
Chudobiak W J 1970 *IEEE Trans. Electron Devices* **ED-17** 843–52
Cornu J 1973 *IEEE Trans. Electron Devices* **ED-20** 347–52
Cornu J and Jaecklin A 1975 *Solid St. Electron.* **18** 683–9
Danhäuser D F and Voss P 1976 *IEEE Trans. Electron Devices* **ED-23** 928–39
Davies R L and Gentry F E 1964 *IEEE Trans. Electron Devices* **ED-11** 313–23
DiLorenzo J V and Wisseman W R 1979 *IEEE Trans. Microwave Theory and Techniques* **MTT-27** 367–78
Dodson W H and Longini R L 1966 *IEEE Trans. Electron Devices* **ED-13** 478–84
Ebers J J and Moll J L 1954 *Proc. IRE* **42** 1761–72
Fair R B 1973 *IEEE Trans. Electron Devices* **ED-20** 642–7
Fletcher N H 1955 *Proc. IRE* **43** 551–9
Fulop W 1967 *Solid St. Electron.* **10** 39–43
Gerlach 1977 *Solid State Devices. Inst. Phys. Conf. Ser. No 32* ed R Muller and E Lange (Bristol: Institute of Physics) pp 111–33
Gummel H K 1961 *Proc. IRE* **49** 837
Gummel H K and Poon H C 1970 *Bell System Tech. J.* **49** 115–20, 827–52
Gupta R K 1980 *Solid St. Electron.* **23** 1011–4
Hauser R J 1964 *IEEE Trans. Electron Devices* **ED-11** 238–42
Hayashi H, Mamine T and Matsushita T 1981 *IEEE Trans. Electron Devices* **ED-28** 246–51
Herlet A 1968a *Solid St. Electron.* **11** 655–71
—— 1968b *Solid St. Electron.* **11** 717–42
Herlet A and Raithel K 1966 *Solid St. Electron.* **9** 1089–105
Herrman H A and Herzer H 1975 *J. Electrochem. Soc.* **122** 1568–9
Higashisaka A, Takayama Y and Hasegawa A F 1980 *IEEE Trans. Electron Devices* **ED-27** 1025–9
Howard N R and Johnson G N 1965 *Solid St. Electron.* **8** 275–84
Hower P L 1973 *IEEE Trans. Electron Devices* **ED-20** 426–35
—— 1976 *IEEE Trans. Electron Devices* **ED-23** 863–70
—— 1980 *IEDM Tech. Digest* 289–92
Hower P L, Blackburn D L, Detinger F E and Rubin S 1976 *Power Electronics Specialists Conf., PESC Record* 234
Hower P L and Einthoven W G 1978 *IEEE Trans. Electron Devices* **ED-25** 465–71
Hower P L and Govil P K 1974 *IEEE Trans. Electron Devices* **ED-21** 617–23
Hower P L and Reddi V G K 1970 *IEEE Trans. Electron Devices* **ED-17** 320–55
Jaecklin A 1976 *IEEE Trans. Electron Devices* **ED-23** 940–7
Kao Y C and Wolley E D 1967 *Proc. IEEE* **55** 1409–14
Kirk C T Jr 1962 *IEEE Trans. Electron Devices* **ED-9** 164–74
Kokosa R A 1967 *Proc. IEEE* **55** 1389–400
Kokosa R A and Tuft B R 1970 *IEEE Trans. Electron Devices* **ED-17** 667–72
Konishi N, Mori M, Naito M, Tanaka T, Miyata K and Okamura M 1980 *IEDM Tech. Digest* 642–5
Krishna S 1977 *Solid St. Electron.* **20** 875–8
Liechti C A 1976 *IEEE Trans. Microwave Theory and Techniques* **MTT-24** 279–95
Lillard R D, Saltich S L and Volk C E 1973 *J. Electrochem. Soc.* **123** 1116
Longini R L and Melngailis J 1963 *IEEE Trans. Electron Devices* **ED-10** 178–85
McGrath E J and Navon D H 1977 *IEEE Trans. Electron Devices* **ED-24** 1255–9
Mahan G D 1980 *J. Appl. Phys.* **51** 2634–46
Matsushita T, Aoki T, Ohtsu T, Yamoto H, Mayashi H, Okayama M and Kawana Y 1976 *IEEE Trans. Electron Devices* **ED-23** 826–30
Mertens R P, DeMan H J and Van Overstraeten R J 1973 *IEEE Trans. Electron Devices* **ED-20** 772–8
Mok T D and Salama C A T 1978 *IEEE Trans. Electron Devices* **ED-25** 384–6

Moll J L 1954 *Proc. IRE* **42** 1773–84
Moll J L, Su J L and Wang A C M 1970 *IEEE Trans. Electron Devices* **ED-11** 420–3
Nagakiri and Ida K 1977 *Jap. J. Appl. Phys.* **16** 1187
Nagata N 1977 *Jap. J. Appl. Phys.* **16** 212
Nishizawa J I, Terasaki T and Shibata J 1975 *IEEE Trans. Electron Devices* **ED-22** 185–97
Nishizawa J I and Yamamoto K 1978 *IEEE Trans. Electron Devices* **ED-25** 314–22
Ohmi T 1979 *Research Institute of Technical Communication, Tohoku University, Sendai, Japan, Tech. Rep. No* 43
—— 1980 *IEEE Trans. Electron Devices* **ED-27** 536–45
Oh-uchi N, Hayashi H, Yamamoto H and Matsushita T 1979 *IEDM Tech. Digest* 522–4
Olmstead J, Einthoven W, Ponczak S and Kannam P J 1971 *RCA Rev.* **32** 221–46
Otsuka M 1967 *Proc. IEEE* **55** 1400–8
Owyang K and Shafer P 1978 *IEDM Tech. Digest* 667
Possin G E, Adler M S and Baliga B J 1980 *IEDM Tech. Digest* 270–5
Ridley B K 1963 *Proc. Phys. Soc.* **82** 954
Ruch J 1972 *IEEE Trans. Electron Devices* **ED-19** 652
Ruhl H J Jr 1970 *IEEE Trans. Electron Devices* **ED-17** 672–80
Rusu A and Bulucea C 1979 *IEEE Trans. Electron Devices* **ED-26** 201–5
Rusu A, Pietreanu O and Bulucea C 1980 *Solid St. Electron.* **23** 473–80
Salama C A T and Oakes J G 1978 *IEEE Trans. Electron Devices* **ED-25** 1222–8
Scarlett R M and Shockley W 1963 *IEEE Int. Conf. Record* Part 3, p 3
Scarlett R M, Shockley W and Haitz R F 1963 *Physics of Failures in Electronics* (Baltimore, Maryland: Spartan Books) p 194
Schlegel E S 1976 *IEEE Trans. Electron Devices* **ED-23** 888–92
Sheng W W 1975 *IEEE Trans. Electron Devices* **ED-22** 25–7
Shur M C 1978 *IEEE Trans. Electron Devices* **ED-25** 612–8
Storm H F and St Clair J G 1974 *IEEE Trans. Electron Devices* **ED-21** 520–2
Sun S C and Plummer J D 1980 *IEEE Trans. Electron Devices* **ED-27** 356–67
Sundresh T S 1967 *IEEE Trans. Electron Devices* **ED-14** 400–2
Sze S M and Gibbons G 1966 *Solid St. Electron.* **9** 831
Temple V A K 1977 *IEDM Tech. Digest* 423–4
Temple V A K and Adler M S 1976 *IEEE Trans. Electron Devices* **ED-23** 950–5
—— 1977 *IEEE Trans. Electron Devices* **ED-24** 1077–81
Van der Ziel A and Agourdis D 1966 *Proc. IEEE* **54** 411–2
Webster W M 1954 *Proc. IRE* **42** 914–20
Wemple H S, Niehaus W C, Cox M H, DiLorenzo J V and Schlosser W O 1980 *IEEE Trans. Electron Devices* **ED-27** 1013–8
Wessels B and Baliga B J 1978 *IEEE Trans. Electron Devices* **ED-25** 1261–5
Whittier R J and Tremere D A 1969 *IEEE Trans. Electron Devices* **ED-16** 39–57
Wolley E D 1966 *IEEE Trans. Electron Devices* **ED-13** 590–7
Yagi H, Tsuyuki T, Koma K, and Miyazawa 1974 *IEDM Tech. Digest* 282
Yamaguchi K and Kodera H 1977 *IEEE Trans. Electron Devices* **ED-24** 1061–9
Yamaguchi K, Toyabe T and Kodera H 1975 *IEEE Trans. Electron Devices* **ED-22** 1047–9