2 Electronic devices

2.1 Executive summary

This chapter introduces the active devices commonly used in high-speed electronics. It starts with a discussion of the metal–semiconductor field effect transistor, or MESFET – historically the oldest FET concept, which for decades was the most prominent device in microwave electronics. Its pitfalls led to the development of an advanced transistor structure, the high electron mobility transistor (HEMT). It incorporates heterostructures to gain additional freedom in device design. HEMTs mostly replaced MESFETs in micro- and millimetre-wave applications.

Metal-oxide-semiconductor field effect transistors (MOSFETs), which dominate digital electronics, are rapidly making inroads at microwave and even millimetre-wave frequencies. They will be discussed as well, and we will recognise similarities between HEMTs and MOSFETs in the physics of the intrinsic transistor.

Finally, bipolar junction transistors (BJTs) will be introduced, showing how a dilemma in the optimum design of the base layer led to the invention of the heterojunction bipolar transistor (HBT) – again, heterostructures come to the rescue.

For all these components, the chapter will discuss their fundamental physical operation, non-ideal and parasitic effects, and linear and non-linear models, as well as examples in several material systems.

2.2 MESFET

2.2.1 Introduction and current control mechanism

The **m**etal–**s**emiconductor **f**ield **e**ffect **t**ransistor (MESFET) is conceptually the simplest of the commonly used transistor structures and shall therefore be discussed here first. The fundamental idea is quite straightforward: the current flowing through a slab of semiconductor material (from now on called the *channel*) depends on three fundamental parameters for a given externally applied voltage:

(i) velocity of charge carriers,
(ii) density of charge carriers,
(iii) the geometric cross-section the carriers flow through.

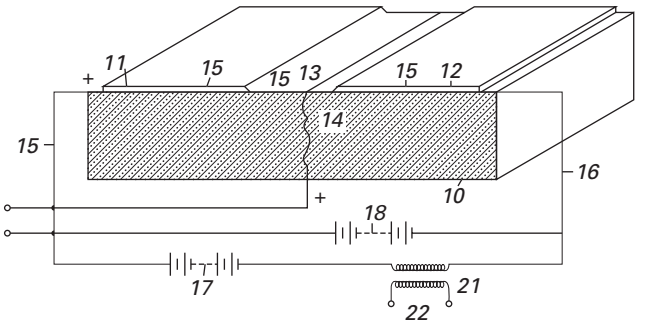

Fig. 2.1 Lilienfeld's FET concept, from his US patent application in 1926.

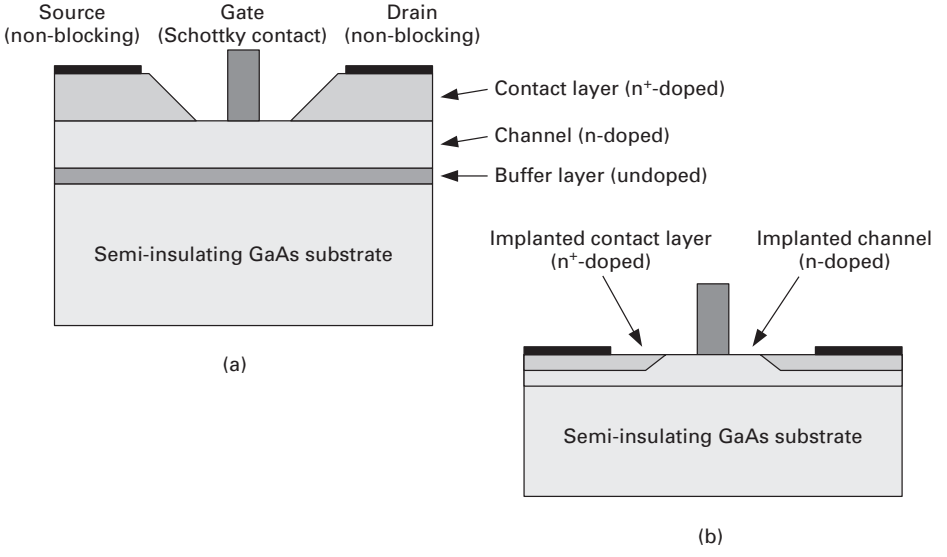

Fig. 2.2 Simplified cross-section of a MESFET with (a) an epitaxially grown channel, (b) fabricated using ion implantation.

While the carrier velocity will depend on the local electric field, in the simplest case the density of charge carriers is given by the doping concentration. The channel cross-section can be influenced externally, if we constrict the current flow using the depletion region of a diode. This method was recognised very early and is the object of a patent filed in 1926 by Julius Edgar Lilienfeld [35]. Lilienfeld's concept (see Figure 2.1), already used a metal–semiconductor junction to control the current flow, but was never realised. The practical realisation of the MESFET is predated by the silicon junction field effect transistor (JFET), which uses a p–n diode as the controlling element and was first described by Shockley [57].

Figure 2.2 shows two somewhat simplified cross-sections of what a MESFET looks like. The layer structure in Figure 2.2(a) is defined by epitaxial growth. Above a semi-insulating substrate, a thin undoped buffer layer is grown to improve the interface

quality, then the channel layer follows whose doping concentration and thickness are very important design parameters, as we will shortly see. Above it, we find a highly doped contact layer, intended to improve the formation of non-blocking contacts and to reduce series resistances between the source and drain contacts and the channel region, but whose exact thickness and doping concentration have no bearing on the fundamental properties of the transistor. Below the gate contact, the contact layer is etched away to allow the blocking Schottky contact to contact the channel layer directly.

Figure 2.2(b) shows a very similar structure; only now the differently doped semiconductor regions are formed by ion implantation. This results in lower cost, however; the lattice damage caused by the ion bombardment will negatively impact carrier velocity and also lead to an increase in low-frequency noise. This will not be discussed in detail here.

In both cases, it is assumed that the carrier species in the channel are electrons (n-channel), as this is the more common variant; however, p-channel devices can be fabricated with equal ease.

While a MESFET can be structured on many different semiconductor materials, only devices fabricated in GaAs and in SiC are commercially relevant. The GaAs MESFET was, for many years, the mainstay of microwave solid-state electronics and shall be discussed here, while the SiC MESFET with its excellent thermal properties and high breakdown voltages is used predominantly in power amplifiers for mobile phone base stations.

For the benefit of clarity and to obtain analytic expressions, we will simplify the structure even further. Figure 2.3 shows the three-dimensional view of the simplified structure. First of all, note the coordinate system which will be used similarly throughout. The x axis is parallel to the 'long' extension of the gate stripe. The y axis is

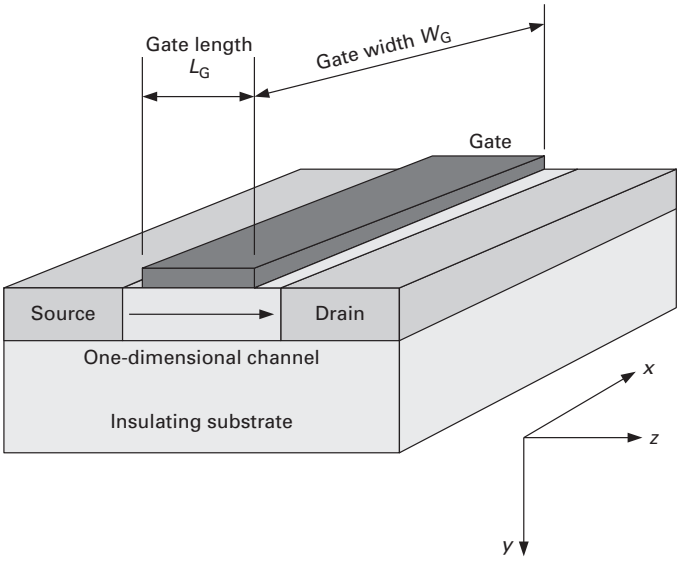

Fig. 2.3 Simplified MESFET structure used in the analytic calculations.

perpendicular to the semiconductor surface, while the z axis is parallel to the surface in the direction of the 'short' extension of the gate. The 'long' gate dimension in x-direction is called the *gate width* W_G, while the *gate length* L_G is the extension in the z direction.

We assume now that the channel is one-dimensional – the electric field in the channel has only a z component. To neglect the electric field in the x direction is generally justified as $W_G \gg L_G$, but to neglect the electric field in the channel in the y direction is a simplification.

Another important simplification in the channel is the *gradual channel approximation*. In general, current flow in semiconductor devices can be driven by the electric field (this is the drift current) or by concentration gradients – this is the diffusion current or a combination of both. Here, we assume that the drift current entirely dominates and the diffusion current can be neglected.

The gate electrode forms a blocking contact with the semiconductor layer under the channel, a *Schottky diode*, which was discussed already in Chapter 1.

Figure 2.4 shows only the cross-section of the device. The source electrode shall be the reference potential, hence $V_S = 0$.

The gate electrode potential with respect to the source is the gate-source voltage V_{GS}. In an n-channel device, where the channel layer is n-doped, it will generally be negative to maintain the gate-channel diode in a blocking state. The drain-source voltage V_{DS} in an n-channel device will be positive.[1]

The extension of the space charge region, shown schematically in Figure 2.4, depends on the local gate-channel voltage V_G. We find for $V_G(z)$:

$$V_G(z) = V_{GS} - V(z), \tag{2.1}$$

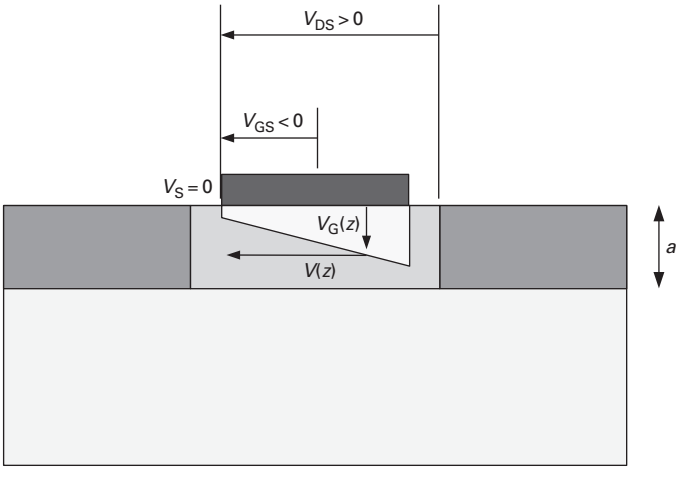

Fig. 2.4 MESFET channel with space charge region for small drain-source voltages.

[1] Because our structure is symmetric with respect to the non-blocking contacts, this defines the drain – in an n-channel device, the drain is the contact with the higher potential.

where $V(z)$ is the voltage drop in the channel between point z and source. As the drain is at a higher potential than the source, $V(z) > 0$, the gate-channel voltage becomes more negative as z increases.

At point z, the extension of the space charge region is

$$h(z) = \sqrt{\frac{2\epsilon_s[V_{bi} - V_G(z)]}{qN_D}} = \sqrt{\frac{2\epsilon_s[V_{bi} - V_{GS} - V(z)]}{qN_D}}. \tag{2.2}$$

with N_D the channel doping concentration, assumed to be constant throughout the channel.

2.2.2 Drain current using a constant-mobility assumption

Let us first consider small V_{DS}, such that $h(z) < a$, with a the channel thickness for all $0 < z < L_G$ – there is always an undepleted part of the channel remaining. We will now calculate the channel current.

The channel current is always calculated in the same fashion: by multiplying the moved charge density (here, qN_D), the cross-section through which it is moved (here $W_G[a - h(z)]$) and the charge velocity. For low fields, the charge velocity can be calculated from the electron mobility μ_n and the local electric field, here $dV(z)/dz$. Therefore, we find an expression for the channel current as a function of the z coordinate:

$$I(z) = qN_D W_G[a - h(z)]\mu_n \frac{dV(z)}{dz}. \tag{2.3}$$

As a consequence of Kirchhoff's law, we know that the charge current entering at the source will be equal to that leaving at the drain – this is called *current continuity*. It allows us to eliminate the z dependence of the current through a simple mathematical trick.

As $I(z) = \text{const} = I_D$, obviously $\int_0^{L_G} I(z)dz = I_D L_G$.

Consider further that

$$h^2(z) = \frac{2\epsilon_s}{qN_D}[V_{bi} - V_{GS} + V(z)].$$

Differentiating both sides with respect to z leads to

$$2h(z)\frac{dh(z)}{dz} = \frac{2\epsilon_s}{qN_D}\frac{dV(z)}{dz},$$

and finally

$$\frac{dV(z)}{dz} = \frac{qN_D}{\epsilon_s}h(z)\frac{dh(z)}{dz}.$$

Through parameter substitution, we find

$$I_D = \frac{1}{L_G}\int_{z=0}^{z=L_G} I(z)dz = \frac{q^2 N_D^2 W_G \mu_n}{\epsilon_s L_G}\int_{h(0)}^{h(L_G)} h(z)[a - h(z)]dh.$$

As $V(0) = 0$, $h(0) = \sqrt{\frac{2\epsilon_s}{qN_D}(V_{bi} - V_{GS})}$. Incidentally, the necessary gate-source voltage to fully close the channel at the source end is the *pinch-off voltage* V_P:

$$V_P = V_{bi} - a^2\frac{qN_D}{2\epsilon_s}. \tag{2.4}$$

Using V_P, we can write Equation (2.2) in the following form:

$$h(z) = a\sqrt{\frac{V(z) - V_{GS} - V_{bi}}{V_{bi} - V_P}}. \tag{2.5}$$

The required value $h(L_G)$ is now found very easily – we know that $V(z = L_G) = V_{DS}$ and therefore

$$h(z = L_G) = a\sqrt{\frac{V_{DS} - V_{GS} - V_{bi}}{V_{bi} - V_P}}. \tag{2.6}$$

We can now finally solve the current integral using the constant-mobility assumption, and find for the drain current:

$$I_D(V_{GS}, V_{DS}) = \frac{q^2 N_D^2 \mu_n a^3 W_G}{6\epsilon_s L_G} \left\{ \frac{3V_{DS}}{V_{bi} - V_P} - 2\frac{(V_{DS} - V_{GS} + V_{bi})^{3/2} - (V_{bi} - V_{GS})^{3/2}}{(V_{bi} - V_P)^{3/2}} \right\}. \tag{2.7}$$

We had so far assumed that the channel would remain at least partially open. This requires that $h(L_G) \leq a$. From Equation (2.6) we find that this translates into

$$V_{DS} \leq V_{GS} - V_P \equiv V_k, \tag{2.8}$$

where V_k is the *knee voltage*.

For

- $V_{DS} \leq V_k$ the MESFET is in the *linear regime*, while for
- $V_{DS} > V_k$ it is in the *saturated regime*.

Figure 2.5 shows simulated output current–voltage characteristics for a hypothetical MESFET with a pinch-off voltage $V_P = -2\,V$ and a built-in voltage of $V_{bi} = 0.7\,V$, calculated using Equation (2.7). The drain current has been normalised to $q^2 N_D^2 \mu_n a^3/(6\epsilon_s L_G)$.

Note that for very small V_{DS}, the dependence of I_D on V_{DS} is almost linear. MESFETs can be used as electronically controllable resistors, e.g. in variable microwave attenuators or in transmit/receive switches.

For $V_{DS} \to V_k$, the drain current saturates. A common assumption in simple FET models is that for $V_{DS} > V_k$, $I_D(V_{DS} > V_k) = I_D(V_{DS} = V_k) = \text{const}$.

Conceptually, current continuity in the constant-mobility model requires that for increasing z, the electric field increases to compensate for the decrease in undepleted channel height. Near $V_{DS} = V_k$, $a - h(z) \to 0$ would imply $dV(z)/dz \to \infty$ at the drain end, which is a fundamental flaw of this model. It is still valuable to investigate MESFET behaviour at low V_{DS}.

2.2.3 Constant-velocity approximation

In all semiconductor materials, the assumption that the carrier velocity increases linearly with increasing electric field, i.e. that mobility is a constant, only holds for small electric fields. For large electric fields, the carrier velocity becomes independent of the electric

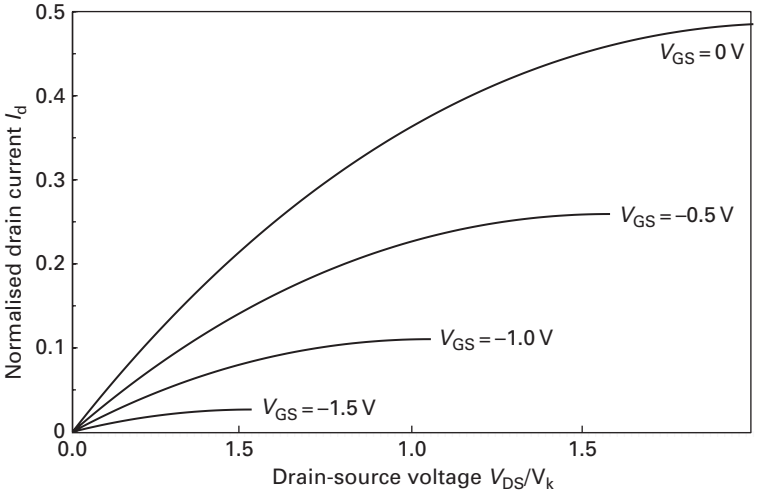

Fig. 2.5 Simulated output I–V characteristics of a MESFET using the constant-mobility assumption ($V_P = -2$ V, $V_{bi} = 0.7$ V).

field (in good approximation); its value is then called the *drift saturation velocity*, $v_{sat,n}$ for electrons or $v_{sat,p}$ for holes, respectively.

Let us now boldly assume that the charge carriers in the channel reach their drift saturation velocity immediately after entering the channel at the source side.

The channel current in our n-channel MESFET now becomes

$$I(z) = qN_D v_{sat,n} W_G[a - h(z)] = \text{const} = I_D,$$

due to the current continuity requirement. As the carrier velocity is now constant, this implies that the channel height must also be constant: $h(z) = \text{const} = h$. Hence,

$$I_D = qN_D v_{sat,n} W_G(a - h). \tag{2.9}$$

The extension of the space charge region can be easily calculated at $z = 0$ for a homogeneous channel doping profile:

$$h = h(0) = a\sqrt{\frac{V_{bi} - V_{GS}}{V_{bi} - V_P}}. \tag{2.10}$$

Inserting this into Equation (2.9), we find

$$I_D = qN_D v_{sat,n} W_G a \left(1 - \sqrt{\frac{V_{bi} - V_{GS}}{V_{bi} - V_P}}\right). \tag{2.11}$$

The value for $V_{GS} = 0$ is referred to as I_{DSS}:

$$I_{DSS} = qN_D v_{sat,n} W_G a \left(1 - \sqrt{\frac{1}{1 - \frac{V_P}{V_{bi}}}}\right). \tag{2.12}$$

The constant-velocity model is a priori only valid for high electric fields in the channel, in the saturation region of MESFET operation ($V_{DS} > V_k$). For very small

V_{DS}, Equation (2.7) still holds, and the electric field will not 'jump' close to source, in every case there will be a region close to source where the constant mobility approximation is more appropriate. More realistic models of MESFET operation will therefore have to combine the constant-velocity and constant-mobility approaches, as was first pointed out by Pucel, Haus and Statz [46]. This is, however, beyond the scope of this introduction.

In a technical MESFET with short gate length and in saturation region, the constant-mobility regime is restricted to an area very close to source, and the majority of the channel is velocity saturated. Equation (2.7) is, therefore, a good approximation for $I_D(V_{GS} > V_P)$ in saturation.

In practical cases, the onset of saturation is not due to $h(L_G) \rightarrow a$, the channel pinching off at the drain end, but due to the onset of velocity saturation in the channel. This occurs much earlier, and hence the FET will pass from the linear to the saturated regime at significantly lower V_{DS} than predicted by the constant-mobility model.

The discussion so far was restricted to MESFETs with constant doping concentration in the channel. Often, however, N_D varies in the channel in the y direction. Two important examples are as follows:

(i) The *ion-implanted MESFET* (see Figure 2.2(b)). Here, the doping concentration varies according to

$$N_D(y) = \frac{Q}{\sqrt{2\pi}\sigma} \exp\left[-\left(\frac{y - R_P}{\sqrt{2\pi}\sigma}\right)^2\right],$$

where Q is the implanted dose, σ the standard deviation and R_P the projected range.

(ii) The *pulse-doped MESFET*, where only a fraction of the channel is highly doped (see Figure 2.6). The discussion of the pulse-doped MESFET is interesting because it has a distribution of mobile carriers similar to that of the HEMT which will be discussed further down.

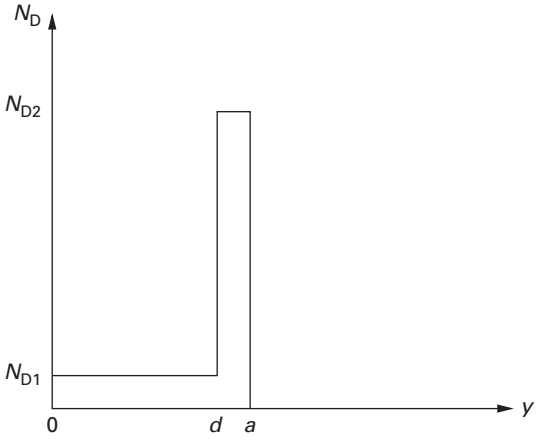

Fig. 2.6 Channel doping profile of a pulse-doped MESFET.

The general procedure to handle these non-uniform doping profiles is as follows:

Poisson's equation is used to obtain a relationship between the potential in the y direction and the space charge distribution $N_D(y)$:

$$\frac{d^2V(y)}{dy^2} = -\frac{q}{\epsilon_s}N_D(y).$$

Integrating Poisson's equation twice yields the required relationship between the gate-channel voltage and the extension of the space charge region. The current is now found by integrating the free carrier concentration over the undepleted channel cross-section:

$$I_D = q\, v_{sat,n} W_G \int_{h(z)}^{a} N_D(y)dy.$$

In the case of a pulse-doped MESFET [40] and $N_{D2} \gg N_{D1}$, the calculation yields

$$I_D = I_{DSS}\left[1 - \frac{\sqrt{1 + \left(\frac{a^2}{d^2} - 1\right)\frac{(V_{bi}-V_{GS})}{(V_{bi}-V_P)}} - 1}{\frac{a}{d} - 1}\right]. \tag{2.13}$$

Figure 2.7 compares the transfer characteristics $I_D = f(V_{GS})$ for a homogeneously doped MESFET and a pulse-doped MESFET with $a/d = 1.1$, i.e. where only 10% of the channel is highly doped. The drain current is normalised to the respective I_{DSS}, which will be different in both cases. Figure 2.7 should not be read suggesting that the pulse-doped MESFET has a lower transconductance than the homogeneously doped MESFET.

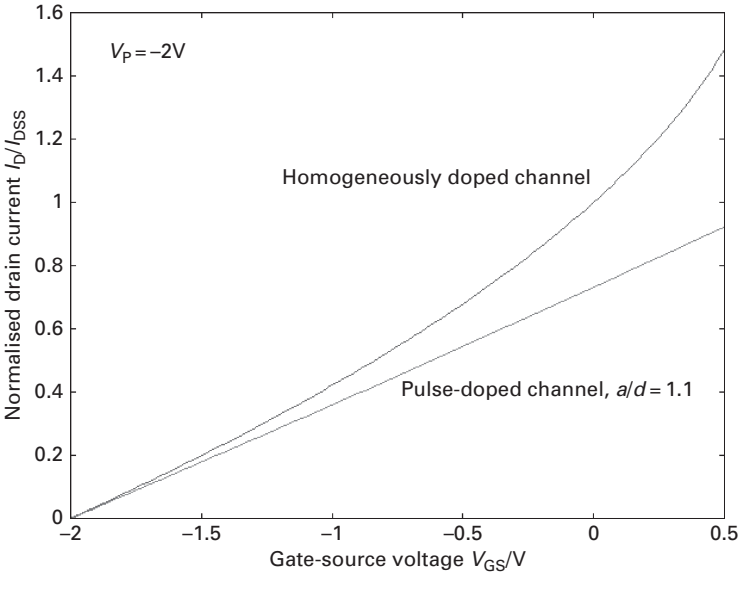

Fig. 2.7 Transfer characteristics $I_D = f(V_{GS})$ for homogeneously and pulse-doped MESFETs.

A striking difference is that $I_D = f(V_{GS})$ is very linear for the pulse-doped MESFET, which is an important advantage from the circuit designer's point of view.

All of the above models consider the drain current in the saturation regime to be independent of the drain-source voltage. In reality, however, I_D depends weakly on V_{DS} there.

The predominant reason for the $I_D = f(V_{DS})$ behaviour in the saturated regime is the scattering of charge carriers into the buffer/substrate layer under the channel. For a semi-insulating substrate, where the Fermi level is near mid-gap, the potential barrier between the channel and the substrate is approximately $E_G/2$ and may be overcome by electrons with sufficient kinetic energy. These electrons may produce two different effects, both of which lead to an $I_D = f(V_{DS})$ dependence:

(i) They may lead to a parasitic conduction current through the buffer or the substrate, adding to the channel current.
(ii) They may be captured by crystal faults or impurities in the buffer or the substrate, which act as charge traps. The resulting modification of charge below the channel influences the channel cross-section, just as a gate electrode would ('backgating').

The latter effect leads to a pronounced frequency dependence of the $I_D = f(V_{DS})$ behaviour.

2.2.4 Large-signal CAD model

For circuit design applications, the physical models considered so far are not convenient. For once, it would be useful to have a model which describes the full range of operation in one closed formula. More importantly, the physical design parameters such as channel thickness and doping concentration are often not accessible to the circuit designer.

Large-signal CAD models, therefore, are empirical in nature and have extractable parameters which can be determined from measurements on the final device.

An early empirical model which may be used for MESFETs in the saturation regime is the 'square-law' JFET model implemented in SPICE:

$$I_D(V_{GS}) = \beta(V_{GS} - V_P)^2, \tag{2.14}$$

where

$$\beta = \frac{I_{DSS}}{V_P^2}.$$

It fits the transfer characteristics of the constant-mobility model quite well at $V_{DS} = V_k$.

The model can be made to fit to non-parabolic transfer characteristics through a modification introduced by Statz *et al.* [58]:

$$I_D(V_{GS}) = \frac{\beta(V_{GS} - V_P)^2}{1 + \alpha(V_{GS} - V_P)}. \tag{2.15}$$

The linear region can be elegantly included in a closed form through the use of a hyperbolic tangent function, as was first pointed out by Curtice [10]:

$$I_D(V_{GS}, V_{DS}) = I_S(V_{GS}) \tanh(\gamma V_{DS}), \tag{2.16}$$

where $I_S(V_{GS})$ is the drain current according to Equation (2.15).

The dependence of the drain current on the drain-source voltage in the saturation regime can be introduced through an additional $1 + \lambda V_{DS}$ term. We arrive finally at the common CAD model expression for the MESFET:

$$I_D(V_{GS}, V_{DS}) = \frac{\beta(V_{GS} - V_P)^2}{1 + \alpha(V_{GS} - V_P)} \tanh(\gamma V_{DS})(1 + \lambda V_{DS}). \tag{2.17}$$

Capacitance model

We will now leave the quasi-static realm and introduce capacitances. The discussion will be restricted first to the intrinsic FET structure, while parasitic capacitances will be introduced in context with the small-signal equivalent circuit.

Assume a MESFET with a homogeneously doped channel region with $V_{DS} = 0$, which implies a constant extension of the gate space charge region, h. The charge on the gate electrode counter-balances the charge in the channel. In this case (n-channel) the gate charge is positive:

$$Q_{G0} = qN_D W_G L_G(a - h) = -qN_D W_G L_G a \left(1 - \sqrt{\frac{V_{bi} - V_{GS}}{V_{bi} - V_P}}\right), \tag{2.18}$$

using Equation (2.5) and $V(z) = 0$ due to $V_{DS} = 0$.

The gate-channel capacitance for $V_{DS} = 0$ can now be calculated as the first derivative of the gate charge with respect to the gate-channel voltage (which is identical to V_{GS} as $V_{DS} = 0$):

$$C_{GC} = \frac{\delta Q_{G0}}{\delta V_{GS}} = C_0 \sqrt{\frac{V_{bi} - V_P}{V_{bi} - V_{GS}}}, \tag{2.19}$$

where

$$C_0 = q\frac{N_D W_G L_G a}{2(V_{bi} - V_P)}.$$

For $V_{DS} > 0$, the Meyer capacitance approach originally developed for MOSFETs [38] is often used, which distinguishes between the linear ($V_{DS} < V_k$) and the saturated ($V_{DS} > V_k$) regimes:

- For $V_{DS} < V_k$,

$$C_{GS} = \frac{2}{3} C_{GC} \left[1 - \left(\frac{V_k - V_{DS}}{2V_k - V_{DS}}\right)^2\right]$$

$$C_{GD} = \frac{2}{3} C_{GC} \left[1 - \left(\frac{V_k}{2V_k - V_{DS}} \right)^2 \right].$$

- For $V_{DS} > V_k$,

$$C_{GS} = \frac{2}{3} C_{GC}$$

$$C_{GD} = 0.$$

The intrinsic $C_{GD} = 0$ in the saturated regime means that the gate-drain voltage has no influence on the channel charge.

Parasitic circuit elements

Our discussion so far was restricted to the intrinsic transistor, more precisely to the channel region. A realistic transistor model will also have to take extrinsic circuit elements into account (see Figure 2.8). The most important ones are:

(i) The source resistance R_S and the drain resistance R_D. They contain contributions from the semiconductor–metal contact at the source, and the semiconductor regions between the channel and the source and drain contacts, respectively. The source contact is most important, because it has a direct impact on the controlling gate-source voltage. Because the gate current can be generally neglected, the intrinsic gate-source voltage V_{GS} is related to the externally applied $V_{GS,e}$ as follows:

$$V_{GS} = V_{GS,e} - R_S I_D.$$

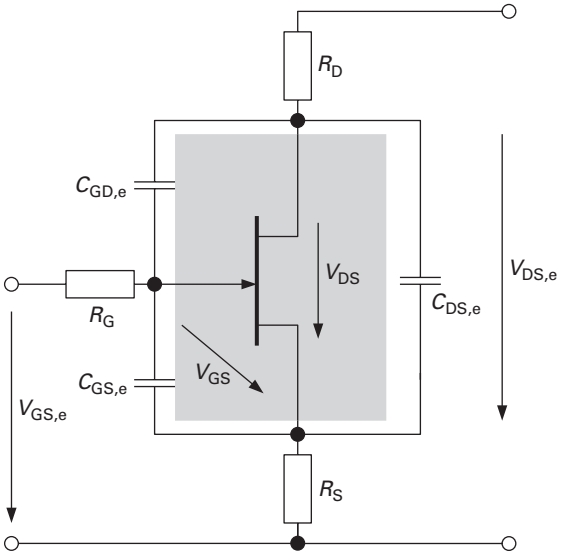

Fig. 2.8 Extrinsic circuit elements in the MESFET. The transistor symbol in the shaded box is the intrinsic transistor discussed so far.

(ii) The gate resistance, which is due to the series resistance of the gate electrode (in x direction in Figure 2.3). This can be a problem especially in modern devices with very small gate length L_G, typically $\leq 0.25\,\mu$m.

(iii) The parasitic capacitances are $C_{GS,e}$, $C_{GD,e}$ and $C_{DS,e}$. They are mostly due to the contact and interconnect metallisations within the transistor structure. $C_{GD,e}$ is of particular importance because it is in a feedback path in the frequently used common-source transistor configuration where it will give rise to the so-called *Miller capacitance*, and also may lead to amplifier instability.

2.2.5 Small-signal equivalent circuit

Introduction: small-signal versus large-signal model

The physical behaviour of electronic devices is generally non-linear, as has been seen above. However, in many cases, we only deal with very small perturbations around a given bias point, so that the non-linear functions can be approximated by linear ones, dramatically simplifying the calculation effort.

For example, the non-linear dependence of the drain current on the gate-source and drain-source voltages, $I_D(V_{GS}, V_{DS})$, can be approximated for small perturbations around a bias point $(I_{D,0}, V_{GS,0}, V_{DS,0})$ by a two-dimensional Taylor series, which is aborted after the linear term:

$$i_d = \frac{\delta I_D}{\delta V_{GS}} v_{gs} + \frac{\delta I_D}{\delta V_{DS}} v_{ds} + \cdots . \tag{2.20}$$

The lower-case symbols i_d, v_{gs} and v_{ds} denote small deviations from the bias point:

$$i_d = (I_D - I_{D,0});\ v_{gs} = (V_{GS} - V_{GS,0});\ v_{ds} = (V_{DS} - V_{DS,0}).$$

MESFET small-signal equivalent circuit

Refer again to Equation (2.20).

The first partial derivative is the *transconductance* g_m:

$$\frac{\delta I_D}{\delta V_{GS}} \Big|_{V_{GS,0}, V_{DS,0}} \equiv g_m.$$

In saturation, we can use Equation (2.17) to calculate its bias-dependent value:

$$\begin{aligned} g_m &= \frac{\delta}{\delta V_{GS}} \left[\frac{\beta (V_{GS} - V_P)^2}{1 + \alpha (V_{GS} - V_P)} \tanh(\gamma V_{DS})(1 + \lambda V_{DS}) \right] \\ &\approx \beta \frac{2(V_{GS,0} - V_P)[1 + \alpha (V_{GS,0} - V_P)] - \alpha (V_{GS,0} - V_P)^2}{\left[1 + \alpha (V_{GS,0} - V_P)\right]^2} \\ &= \beta \frac{\alpha (V_{GS,0} - V_P)^2 + 2(V_{GS,0} - V_P)}{\left[1 + \alpha (V_{GS,0} - V_P)\right]^2}, \end{aligned} \tag{2.21}$$

assuming that $\lambda V_{DS} \ll 1$, and that when sufficiently in saturation, $\tanh(\gamma V_{DS}) \to 1$.

The second partial derivative in Equation (2.20) is the *output conductance* g_{ds}:

$$\frac{\delta I_D}{\delta V_{DS}} |_{V_{GS,0}, V_{DS,0}} \equiv g_{ds}.$$

In saturation, assuming again that $\tanh(\gamma V_{DS}) \to 1$:

$$\begin{aligned} g_{ds} &= \lambda \frac{\beta (V_{GS,0} - V_P)^2}{1 + \alpha (V_{GS,0} - V_P)} \\ &\approx \lambda I_{D,0}, \end{aligned} \tag{2.22}$$

if we assume also $\lambda V_{DS,0} \ll 1$.

In saturation, the bias-dependent intrinsic gate-source capacitance is (see p. 57):

$$C_{GS,i} = \frac{2}{3} C_{GC} = q \frac{N_D W_G L_G a}{3 \cdot \sqrt{(V_{bi} - V_P)(V_{bi} - V_{GS,0})}}. \tag{2.23}$$

To this, we have to add the extrinsic gate-source capacitance, so that

$$C_{GS} = C_{GS,i} + C_{GS,e}.$$

The gate-drain capacitance has only an extrinsic component (refer again to p. 57):

$$C_{GD} = C_{GD,e}.$$

Likewise, C_{DS} is purely extrinsic:

$$C_{DS} = C_{DS,e}.$$

Adding the series resistances R_G, R_S and R_D, we arrive at a first small-signal equivalent circuit for the MESFET (see Figure 2.9).

A more complete small-signal equivalent circuit will add two more elements:

(i) the resistance R_i which improves the modelling of the non-velocity-saturated part of the channel near the source;
(ii) the domain capacitance C_{DC}.

The domain capacitance accounts for a charge dipole forming at the drain end of the channel. Provided that $R_i \ll 1/(\omega C_{GS})$, it can safely be omitted as it is absorbed in C_{DS}.

It is important to note that in Figure 2.10, V_{GS} drops over C_{GS} only.

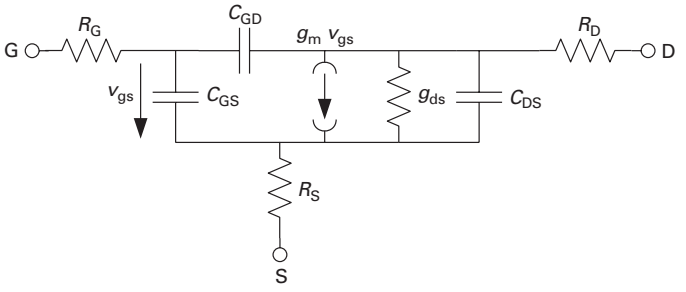

Fig. 2.9 Simple small-signal equivalent circuit of a MESFET.

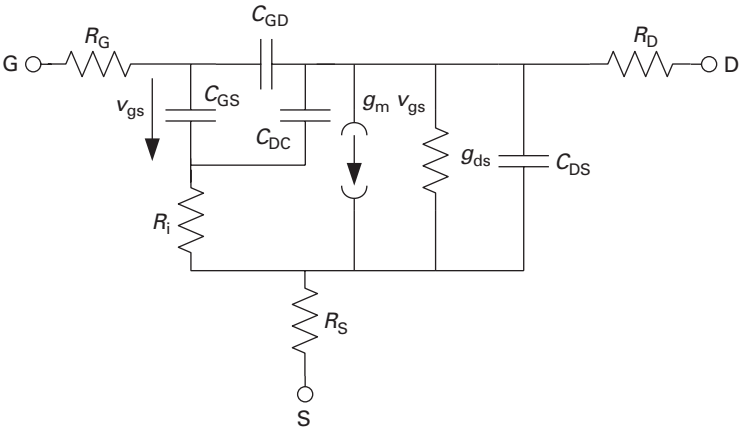

Fig. 2.10 Small-signal equivalent circuit of a MESFET including R_i and C_{DC}.

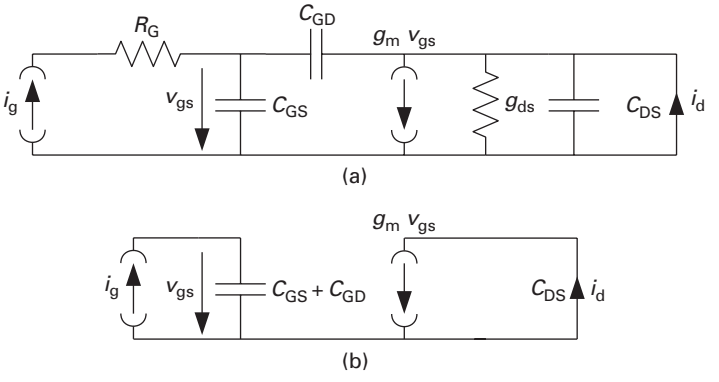

Fig. 2.11 (a) Simplified MESFET small-signal equivalent circuit connected for measuring f_T.
(b) Collapsed equivalent circuit due to current source at input and short circuit at output.

Transit frequency

The transit frequency of a two-port is defined as the frequency where the magnitude of the *short-circuit current gain* h_{21} becomes one:

$$|h_{21}(f = f_T)| = \frac{i_d}{i_g}\,|_{v_{ds}=0} = 1. \tag{2.24}$$

To calculate f_T from the small-signal parameters, we refer back to the simple MESFET equivalent circuit (Figure 2.9), and further simplify it by omitting the series resistances R_S and R_G, which in technical MESFETs are quite small.

Equation (2.24) can be interpreted as forcing a current i_g into the gate terminal, while measuring a short-circuit current i_d between the source and the drain terminals.

The appropriate connections are indicated in Figure 2.11(a). Because R_G is in series with an ideal current source, it has no effect here and can be omitted. Elements g_{ds} and C_{DS} are short-circuited and can be omitted also. C_{GD} is in parallel to C_{GS}.

Figure 2.11(b) shows the extremely simple equivalent circuit after taking these findings into consideration. The current transfer function is now:

$$i_d = g_m v_{gs} = i_g \frac{g_m}{j\omega(C_{GS} + C_{GD})},$$

which means that

$$h_{21}(\omega) = \frac{g_m}{j\omega(C_{GS} + C_{GD})}.$$

The magnitude of h_{21} becomes unity at

$$\omega_T = \frac{g_m}{C_{GS} + C_{GD}},$$

or

$$f_T = \frac{g_m}{2\pi(C_{GS} + C_{GD})}. \tag{2.25}$$

We will now relate the transit frequency to physical parameters. Let us go back to the simple velocity-saturated MESFET model (Section 2.2.3). In the simple model, we do not use the Meyer capacitance approach, but attribute the full gate-channel capacitance Equation (2.19) to the gate-source capacitance. Parasitic capacitances are neglected:

$$C_{GS} = \frac{q N_D W_G L_G a}{2\sqrt{(V_{bi} - V_P)(V_{bi} - V_{GS})}}.$$

For the transconductance, we derive Equation (2.11) with respect to V_{GS} and find

$$g_m = \frac{q N_D v_{sat} W_G a}{2\sqrt{(V_{bi} - V_P)(V_{bi} - V_{GS})}}.$$

Therefore,

$$f_T = \frac{g_m}{2\pi C_{GS}} = \frac{1}{2\pi}\frac{v_{sat}}{L_G}. \tag{2.26}$$

The transit frequency can be directly deduced from the carrier transit time through the channel.

Maximum frequency of oscillation

The maximum frequency of oscillation f_{max} is a measure of the power gain of a two-port (see Section 5.2.4). A common formulation [32] quoted in [36] for f_{max} from the small-signal parameters is

$$f_{max} = \frac{f_T}{2\sqrt{(R_G + R_i + R_S)g_{ds} + 2\pi f_T R_G C_{GD}}}. \tag{2.27}$$

The expression refers to the equivalent circuit in Figure 2.10, but neglecting C_{DC}.

Note the importance of the series resistances, which did not factor into the calculation of f_T at all. f_{max} is much more useful to benchmark FETs for power amplification at microwave frequencies.

2.2.6 Noise performance

When discussing the noise performance of semiconductor devices, we have to distinguish between microwave noise, where the spectral power density of the contributing noise sources is frequency-independent (white noise), and low-frequency noise phenomena, where the spectral power density of the contributing noise sources increases with decreasing frequency.

Microwave noise

To assess the microwave noise performance of a FET, in principle three different noise sources need to be included, each due to the stochastic movement of charge carriers in different parts of the device. The simplified equivalent circuit in Figure 2.12 contains the MESFET's main noise sources:

(i) Areas where mobility is constant, i.e. the region behaves like an ohmic resistor, give rise to *thermal* or *Johnson* noise. In a realistic MESFET, we need to include Johnson noise for the gate resistance R_G and the source resistance R_S. The mean-squared value of a Johnson noise source can be expressed as: $\langle |e|^2 \rangle = 8kTR\Delta f$, where R is the resistance and Δf is the measurement bandwidth. Hence,

$$\left\langle |e_G|^2 \right\rangle = 8kTR_G\Delta f$$

$$\left\langle |e_S|^2 \right\rangle = 8kTR_S\Delta f.$$

(ii) Current flowing across an energy barrier gives rise to *shot noise*, which is proportional to the current. Here, a potential gate leakage current flows across the gate-channel Schottky diode, resulting in a shot noise component of

$$\left\langle |i_{glc}|^2 \right\rangle = 8qI_{GLC},$$

where I_{GLC} is the DC value of the gate leakage current.

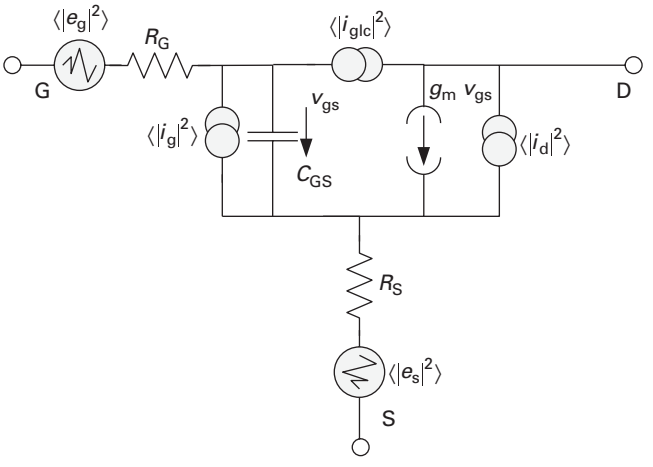

Fig. 2.12 Simplified noise equivalent circuit of a MESFET.

(iii) In the channel, the carrier velocity experiences fluctuations due to phonon and impurity scattering – this kind of noise is commonly called *channel noise*, first predicted by van der Ziel [62]. According to van der Ziel,

$$\left\langle |i_\mathrm{d}|^2 \right\rangle = 8kTg_\mathrm{m}P\Delta f,$$

where P is a fitting parameter equal to 1...3.[2]

Note that van der Ziel did not yet include velocity saturation effects. In fact, analytic FET noise models are strictly valid only below the onset of saturation. However, deviating behaviour in the saturated region and in the presence of velocity saturation can be accommodated by a bias dependence in the parameter P [20].

(iv) Another effect must be taken into account. Due to the close proximity of the gate electrode to the channel, any charge fluctuation in the channel will lead to a phase fluctuation with the opposite sign on the gate electrode. This effect is the *induced gate noise* and was again pointed out by van der Ziel [63]:

$$\left\langle |i_\mathrm{g}|^2 \right\rangle = 8kT\Delta f(\omega C_\mathrm{GS})^2 R/g_\mathrm{m},$$

where R is a fitting parameter, which accommodates different geometries and bias points.

Because of their linked physical origin, $\langle |i_\mathrm{d}|^2 \rangle$ and $\langle |i_\mathrm{g}|^2 \rangle$ are not statistically independent, but show a strong correlation. The correlation coefficient is imaginary (due to the capacitive coupling) and strongly bias-dependent.

The gate leakage noise contribution $\langle |i_\mathrm{glc}|^2 \rangle$ is commonly neglected, because the gate diode is reverse biased. Using this assumption, Cappy [6] expressed the minimum noise figure as

$$F_\mathrm{min} = 1 + 2\sqrt{P + R - 2C\sqrt{PR}}\frac{f}{f_\mathrm{T}} \sqrt{g_\mathrm{m}(R_\mathrm{S} + R_\mathrm{G}) + \frac{PR(1-C^2)}{R + P - 2C\sqrt{RP}}}, \tag{2.28}$$

where C is the magnitude of the correlation coefficient. For $C = 1$, Equation (2.28) is equivalent to the famous *Fukui equation* [21] for the minimum noise figure of FETs:

$$F_\mathrm{min} = 1 + k_\mathrm{F}\frac{f}{f_\mathrm{T}}\sqrt{g_\mathrm{m}(R_\mathrm{G} + R_\mathrm{S})}, \tag{2.29}$$

where k_F is a fitting factor.

Both Equations (2.28) and (2.29) calculate f_T using the approximation in Equation (2.25).

The noise equations contain an implicit bias dependence, which cannot be discussed in detail. Delagebeaudeuf *et al.* [12] showed for the bias dependence of parameter P,

$$P = \frac{I_\mathrm{D}}{\mathcal{E}_\mathrm{crit}L_\mathrm{G}g_\mathrm{m}}, \tag{2.30}$$

[2] van der Ziel discusses this in terms of the channel conductance g_d0, which is identical to the transconductance g_m at the very low V_DS.

which points towards a $\sqrt{I_D}$ dependence for F_{min}, at least where $g_m \approx$ const. At very low I_D, however, g_m also decreases and F_{min} increases again. Optimum drain currents for low-noise operation are typically at 0.15–0.25I_{DSS}. In Equation (2.30), $\mathcal{E}_{crit}$ is the critical electric field for velocity saturation.

Low-frequency noise

Low-frequency noise is only discussed briefly here; however, it will be shown that it has significant impact on circuit performance, especially in oscillators.

While there are quantum mechanical reasons for low-frequency noise occurring in any conducting or semi-conducting material, practical devices exhibit low-frequency noise levels significantly above the quantum limit. This excess noise is due to interaction with impurities or dislocations which create energy levels inside of the forbidden gap of semiconductor materials. These traps may

- locally lead to enhanced scattering of charge carriers – *mobility fluctuation noise*; or
- modify the number of charge carriers through trapping and release, with a characteristic time constant τ – *number fluctuation noise*.

Even though it was derived at first only for mobility fluctuation noise in bulk semiconductors, the empirical Hooge equation [24] is often applied to low-frequency noise parameters. Applied to the drain current I_D, the Hooge relationship finds for the spectral power density of the drain current fluctuations:

$$S_{ID} = I_D^2 \frac{\alpha_H}{N \cdot f}, \tag{2.31}$$

where N is the number of carriers in a given volume and f is the frequency. Due to the observed frequency relationship, low-frequency noise is often coined *1/f noise*.

This ideal $1/f$ noise spectrum is frequently superimposed by generation-recombination spectra through a trap with distinct capture and re-emission time constant τ. Such traps lead to noise with a low-pass limited spectral noise power density:

$$S_N(f) \sim \frac{1}{1 + (2\pi f)^2 \tau^2}. \tag{2.32}$$

Figure 2.13 shows qualitatively a low-frequency noise spectrum of the drain current in the presence of a distinct trap with generation-recombination noise, a $1/f$ noise component and white noise at higher frequencies.

The absolute spectral density depends very strongly on the technology. A high spectral density at a given frequency is indicative for a high number of defects or deep impurities. So it is not surprising that low-frequency noise is much more pronounced for ion-implanted MESFETs (with significant radiation-induced defects) than for epitaxially grown structures.

2.2.7 MESFETs in the third millennium

Commercially, MESFETs were the transistors of choice for microwave circuits, including monolithic microwave ICs (MMICs), from the 1970s well into the 1990s. Initially,

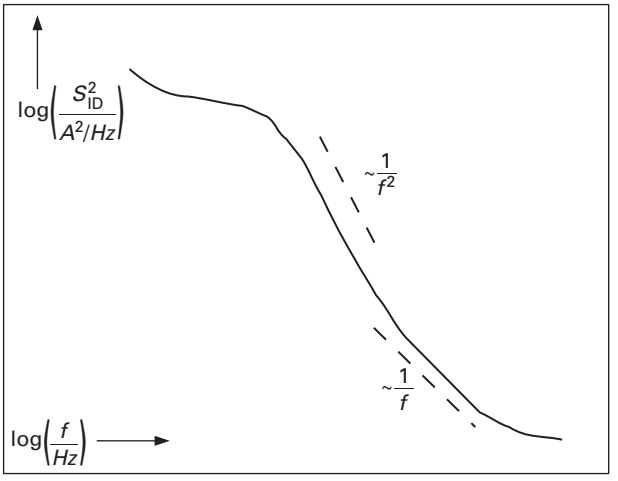

Fig. 2.13 Qualitative low-frequency noise spectrum of the drain current in the presence of $1/f$ noise, generation-recombination noise and white noise.

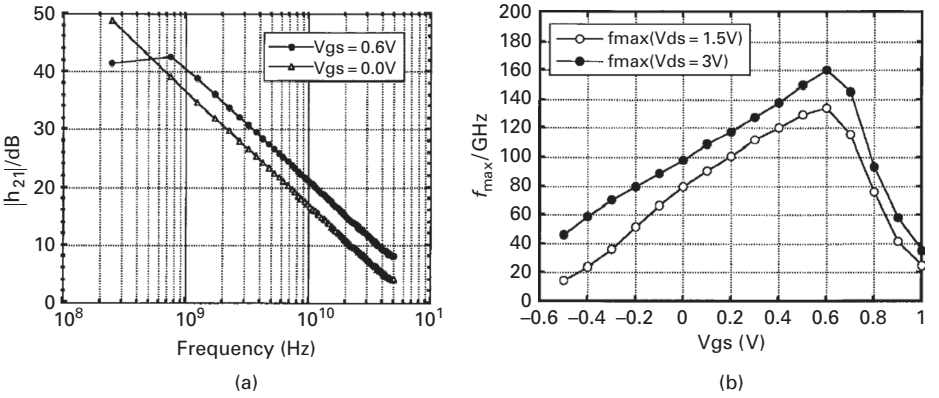

Fig. 2.14 (a) $|h_{21}|$ of a $2 \cdot 0.12\,\mu\text{m} \cdot 75\,\mu\text{m}$ ion-implanted MESFET ($V_{DS} = 1.2$ V). (b) f_{max} of a $2 \cdot 0.12\,\mu\text{m} \cdot 25\,\mu\text{m}$ device in the same technology (H. Hsia, Z. Tang, D. Caruth, D. Becher and M. Feng, *IEEE Electron Device Letters*, Vol. EDL-20, No. 5, pp. 245–247, May 1999. ©1999 IEEE).

they only gained importance on GaAs substrates – MESFET structures on other materials, including Si, were but an academic curiosity. GaAs MESFETs have been almost exclusively replaced in contemporary designs – either by HBTs or HFETs. MOSFETs are also making inroads into the former realm of GaAs MESFETs.

It should be noted that MESFETs did reach cutoff frequencies in excess of 100 GHz, even for devices fabricated by ion implantation. Hsia and co-workers [25] reported a device with $L_G = 0.12\,\mu\text{m}$, which exhibited $f_T = 121$ GHz and $f_{max} = 160$ GHz, albeit not at the same bias point or the same device size. Figure 2.14 shows h_{21} versus frequency and f_{max} versus V_{GS} for this technology.

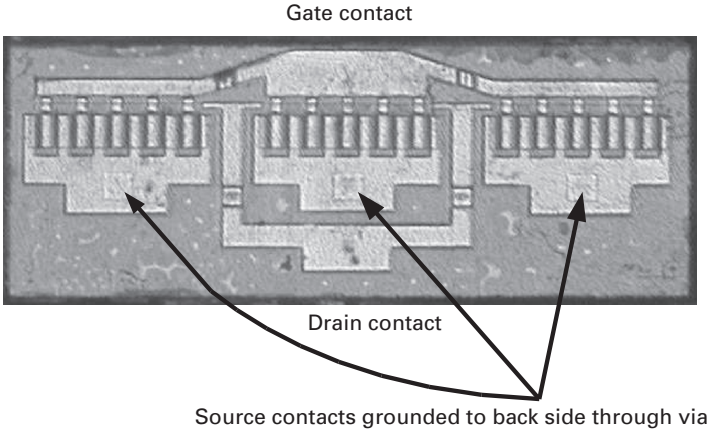

Fig. 2.15 Example of a SiC power MESFET (Chip photo adapted from M. Südow, K. Andersson, N. Billström, J. Gran, H. Hjelmgren, J. Nilsson, P.-A. Nilsson, J. Stahl, H. Zirath and N. Rorsman, *IEEE Transactions on Microwave Theory and Techniques*, Vol. MTT-54, No. 12, pp. 4072–4078, December 2006. ©2006 IEEE).

Note that the maximum f_T is measured at $V_{GS} = 0.6$ V, i.e. the gate electrode starts to be forward-biased. This is also visible from the lower $|h_{21}|$ below 1 GHz. For a more practical $V_{GS} = 0$ V, $f_T = 70$ GHz. f_{max} peaks also at $V_{GS} = 0.6$ V, but is improved by a larger V_{DS}, because the latter will further reduce C_{DG}.

The record f_T and f_{max} values are measured for different device geometries. This is a common trick – f_T is measured for a larger gate finger width (here, 75 μm), because R_G does not matter, and the wider finger leads to a better ratio of intrinsic and parasitic C_{GS}. For f_{max}, R_G does matter, and hence a smaller gate finger width is chosen (here 25 μm).

The MESFET structure makes a strong comeback on SiC, with important applications in power amplifiers, e.g. for mobile radio base stations in the lower GHz range. Figure 2.15 shows an example of such a structure [60].

Note the multi-finger layout which is very common in power FETs. Due to the limited current-carrying ability per unit width (in this case 350 mA mm^{-1}), the total device periphery needs to be extended. As the series resistance per unit length of the gate stripe is quite high in case of submicron gate length (here, $L_G = 0.4$ μm), R_B can be kept small by choosing a short individual gate finger length and connecting transistor cells in parallel, in this example for a total gate width $W_G = 0.4$ mm.

The major advantage of a semiconductor material with a large band gap is the very high electric field at breakdown. In this case, the gate-drain breakdown voltage is 180 V. The transistor shown produces a saturated output power of 3.1 W at 3 GHz, or 7.8 W/mm gate width, when biased at a drain-source voltage $V_{DS} = 65$ V. The power added efficiency in this mode of operation is 70%.

The device has a small-signal $f_T = 8$ GHz and a maximum frequency of oscillation $f_{max} = 20$ GHz. Record transit frequencies were reported at 28 GHz, and record maximum frequencies of oscillation at 50 GHz.

2.3 High electron mobility transistor

While the MESFET is conceptually a very simple device, yielding sufficient performance well into the millimetre wave range, it does not unleash the full potential of group III–V semiconductor materials. The fact that free charge carriers and ionised dopants share the same space in MESFETs leads to a reduction of low-field mobility through electrostatic fields, a major effect which we will consider first.

2.3.1 The importance of Coulomb scattering

Figure 2.16 shows the electron mobility for nominally undoped GaAs as a function of the absolute temperature, along with the two dominant scattering mechanisms. Other scattering mechanisms have been omitted for clarity.

We notice that at room temperature (300 K) the scattering of electrons is mostly due to lattice vibrations – longitudinal optical phonons. As we lower the temperature and lattice vibrations are increasingly suppressed, another mechanism becomes dominant – *Coulomb* scattering. Coulomb scattering is due to the electrostatic force between the mobile charge carriers and the fixed ionised atoms. In doped semiconductors, the main source of fixed charge are the ionised doping atoms. Therefore, the main electrostatic effect we need to consider in an n-channel MESFET is between the negatively charged electrons and the positively charged ionised donors. This is clearly shown in Figure 2.16 through the strong doping dependence of the mobility.

From electrostatic theory we know that the force created between two objects with a charge of magnitude q – the elementary charge – and the opposite sign of charge is

$$F = \frac{q^2}{4\pi\epsilon_s d^2} \propto \frac{1}{d^2}. \tag{2.33}$$

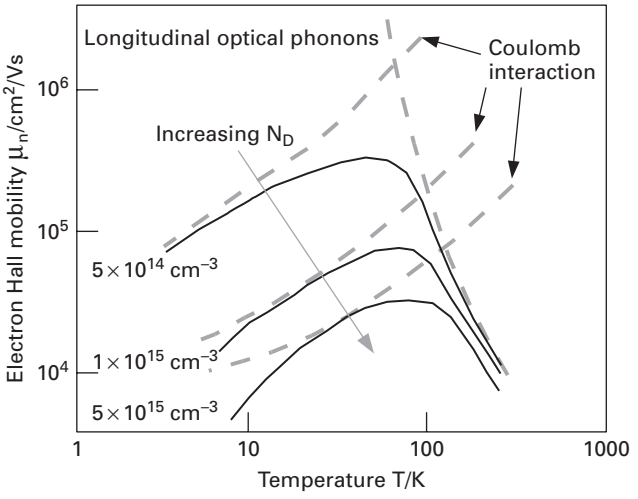

Fig. 2.16 Electron mobility versus absolute temperature for nominally undoped GaAs, and the underlying dominant scattering mechanisms. Data adapted from [59].

With increasing doping concentration, the mobility limiting effect of Coulomb scattering will become more pronounced. This is also shown in Figure 2.16.

Coulomb scattering becomes an increasing problem as we reduce the gate length in MESFETs:

- As we reduce the gate length L_G, we also have to reduce the channel thickness a to keep the *aspect ratio* L_G/a constant.[3]
- To compensate for the reduction in a, we need to increase the *channel doping concentration* N_D.
- Then, however, the significance of Coulomb scattering will increase and reduce the mobility!

If the physical co-location of free and fixed charge is the reason for the increased dominance of Coulomb scattering, then the following idea is immediately apparent: why not physically separate free and fixed charge, i.e. the electrons and the ionised donors in an n-channel device?

To find out how this may be done, let us investigate a *heterojunction* in the n^+ AlGaAs/p^- GaAs material system. The AlGaAs/GaAs material system has the advantage that the lattice constant is almost independent of the material composition.

The band gap in an $Al_xGa_{1-x}As$/GaAs heterojunction adjusts as follows:

E_g(GaAs)	1.42 eV	
ΔE_C(AlGaAs – GaAs)	0.62 ΔE_g	for $x_{Al} < 0.37$
ΔE_g(AlGaAs – GaAs)	1.255 eV x_{Al}	as above

In this example, the Al concentration, doping types and concentrations are:

$Al_{0.25}Ga_{0.75}As$	n-type	$N_D = 10^{18}\,cm^{-3}$
$GaAs$	p-type	$N_A = 10^{15}\,cm^{-3}$

Further, a thin (~5–10 nm) layer of undoped AlGaAs is inserted at the heterojunction – this is the *spacer layer*. Figure 2.17 shows this material combination schematically. When discussing heterostructures, the doping type of large-gap materials will be denoted with capital letters, while the doping type of narrow-gap materials is shown in lower case letters.

The conduction band diagram of this heterostructure is shown in Figure 2.18. The discontinuity at the AlGaAs/GaAs interface, $\Delta E_C = 0.2$ eV, and the potential barrier towards the p^--GaAs form a triangular quantum well structure, which is the most important feature – note how it dips below the Fermi level. Close to the hetero-interface, the potential in the GaAs layer can be approximated by a linear function.

[3] Otherwise, the assumption that the electric field is directed predominantly in parallel to the surface will break down. Among other things, this would significantly increase the output conductance in the saturated regime.

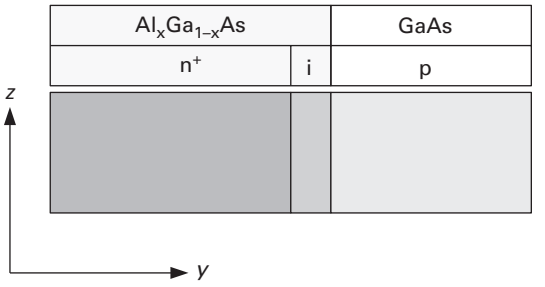

Fig. 2.17 AlGaAs/GaAs n^+–i–p heterostructure.

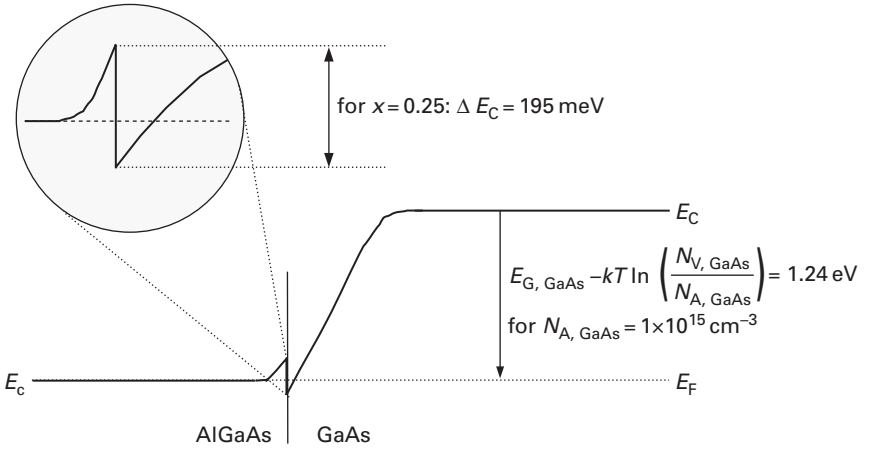

Fig. 2.18 Conduction band diagram of the AlGaAs/GaAs heterostructure.

2.3.2 Charge control

If we force the free electrons out of the n-AlGaAs layer, they will congregate in the potential well where they are separated from the ionised donor atoms by the undoped AlGaAs spacer layer – the sought-after *reduction in Coulomb scattering* can be achieved this way.

We can 'force' the free electrons to leave the AlGaAs when we *deplete* the doped AlGaAs layer (the *supply layer*) by means of a Schottky contact. The electrons can then either tunnel through the spacer layer or overcome the conduction band spike at the heterostructure by thermionic emission.

Note that from now on, it will suffice to draw just the conduction band diagram, because we consider electrons only.

Figure 2.19 represents the band diagram, without applied external voltage, of a high electron mobility transistor, or HEMT. By applying a positive gate voltage, the density of free electrons in the potential well increases; a negative gate voltage will decrease it.

An important difference between the MESFET and the HEMT is the current control mechanism: in the MESFET, we controlled the thickness of the channel, while the

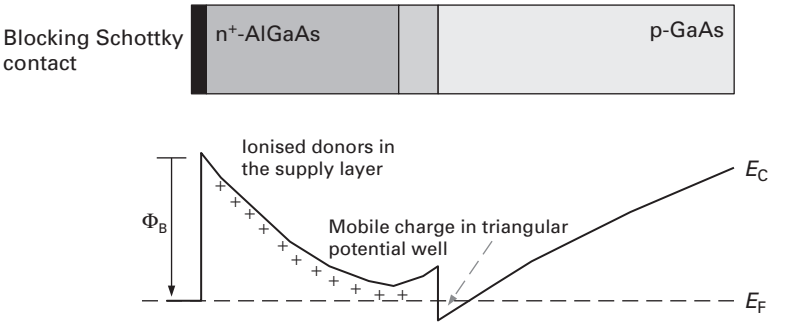

Fig. 2.19 Conduction band diagram of HEMT in thermodynamic equilibrium.

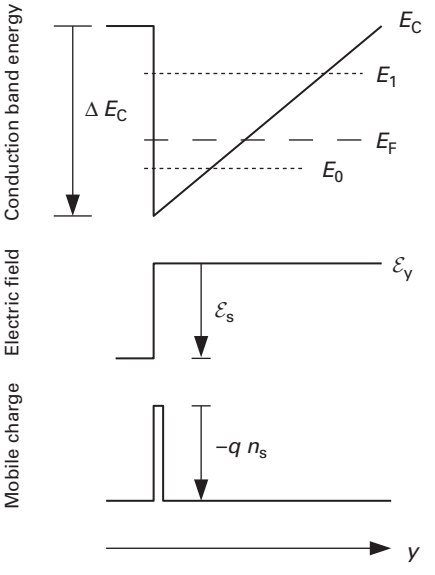

Fig. 2.20 Approximation of the HEMT channel region as a triangular quantum well.

density of charge carriers in the channel remained constant. Here we change the density of carriers in the channel, while the thickness of the channel, given by the triangular well, remains approximately constant.

The free carrier ensemble in the channel is called a *two-dimensional electron gas* (2DEG).

Let us investigate the triangular potential well in more detail. First, we have to be aware that the triangular potential well is narrow enough to introduce *quantisation of energy levels*. Consider Figure 2.20. Note that for convenience, $y = 0$ at the hetero-interface.

We initially assume that the potential walls are infinitely high. In the potential well, electrons may only occupy the discrete energy levels E_l, with $l \in [0, 1, 2, \ldots]$.

Solution of Schrödinger's equation yields these energies:

$$E_l = \left(\frac{h^2}{8\pi^2 m_n^*}\right)^{\frac{1}{3}} \left[\frac{3}{2} q\, \mathcal{E}_y\, \pi \left(l + \frac{3}{4}\right)\right]^{\frac{2}{3}}, \tag{2.34}$$

where $\mathcal{E}_y$ is the y component of the electric field in the well.

The potential increases linearly beyond the heterostructure. The electric field as the gradient of the potential is therefore constant:

$$\mathcal{E}_y = \mathcal{E}_S = -dV(y)/dy.$$

The discontinuity of the electric field at $y = 0$ necessitates a sheet charge in this plane, whose charge density is

$$q\, n_S = \epsilon_1 \mathcal{E}_S = -\epsilon_1 \frac{dV(y)}{dy}, \tag{2.35}$$

where ϵ_1 is the dielectric constant of the semiconductor in the channel region – here GaAs. This sheet charge is the 2DEG.

n_S is the sum over the sheet charge densities in the discrete energy levels: $n_S = \sum_{l=0}^{\infty} n_l$, where only the first two ($l = 0, 1$) typically need to be evaluated, because in practice the walls of the quantum well have a finite height, set by the conduction band discontinuity ΔE_C.

Using $\mathcal{E}_S = q\, n_S/\epsilon_1$ we find

$$E_l = \left(\frac{h^2}{8\pi^2 m_n^*}\right)^{\frac{1}{3}} \left[\frac{3}{2}\frac{q^2}{\epsilon_1}\pi \left(l + \frac{3}{4}\right)\right]^{\frac{2}{3}} n_S^{\frac{2}{3}}. \tag{2.36}$$

The first two terms are material-dependent and shall be combined into a constant γ_l:

$$\left(\frac{h^2}{8\pi^2 m_n^*}\right)^{\frac{1}{3}} \left[\frac{3}{2}\frac{q^2}{\epsilon_1}\pi \left(l + \frac{3}{4}\right)\right]^{\frac{2}{3}} \equiv \gamma_l,$$

and therefore,

$$E_l = \gamma_l\, n_S^{\frac{2}{3}}. \tag{2.37}$$

For GaAs,

$$\gamma_0 = 2.5 \times 10^{-12}\, e\mathrm{V\,m}^{\frac{4}{3}}$$
$$\gamma_1 = 3.2 \times 10^{-12}\, e\mathrm{V\,m}^{\frac{4}{3}}.$$

Next, we need to calculate the sheet charge density of the 2DEG as a function of the Fermi energy (note that the Fermi energy is referenced to the conduction band minimum here).

Consider the density of states for the two-dimensional electron gas in Figure 2.21. The constant $D = q\, m_n^*/(2\pi^2 h^2)$ is $D = 3.24 \times 10^{17}\,\mathrm{m^{-2}\,V^{-1}}$ for GaAs.

The density of the *occupied* states can be calculated from

$$n_S = \textit{density of states} \cdot \textit{occupation probability.}$$

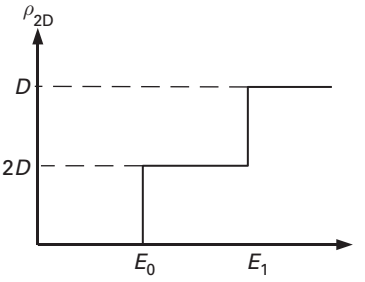

Fig. 2.21 Density of states in the triangular quantum well.

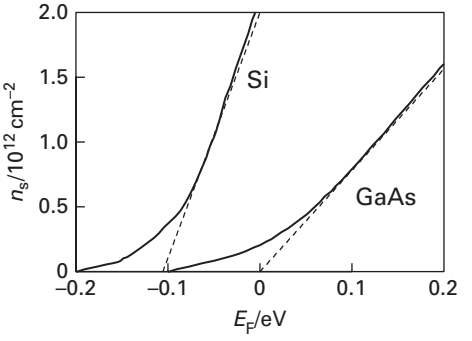

Fig. 2.22 Density of the 2DEG in a HEMT structure as a function of the Fermi energy.

The probability that an allowed state is occupied must be calculated using *Fermi–Dirac statistics* here, because the Fermi energy is inside the conduction band. Only two discrete energy levels are considered:

$$n_S = D \int_{E_0}^{E_1} \frac{dE}{1 + \exp\left(\frac{E-E_F}{kT}\right)} + 2D \int_{E_1}^{\infty} \frac{dE}{1 + \exp\left(\frac{E-E_F}{kT}\right)}. \tag{2.38}$$

For the integral, we find

$$\int \frac{dx}{1 + \exp(a\,x)} = -\frac{1}{a} \ln(1 + e^{-ax}),$$

and therefore, the sheet charge density is

$$n_S = D\,kT \sum_{l=0}^{1} \left[(l+1) \cdot \ln\left(1 + e^{\frac{E_F - E_l}{kT}}\right) \right]. \tag{2.39}$$

Because on the other hand $E_l = \gamma_l\, n_S^{2/3}$, this transcendental equation has to be solved iteratively.

Figure 2.22 [64] shows its solution for the case of Si and for the case of GaAs. For larger charge carrier densities, $n_S(E_F)$ can be approximated by a linear relationship:

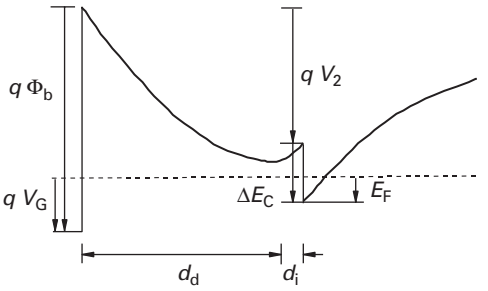

Fig. 2.23 Conduction band diagram of a HEMT structure under gate bias control ($V_G \neq 0$).

$$n_S \approx \frac{E_F - \Delta E_{F0}}{q\,a}. \tag{2.40}$$

For GaAs,

$$\Delta E_{F0}(300\,\mathrm{K}) = 0\,\mathrm{eV}$$
$$\Delta E_{F0}(77\,\mathrm{K}) = 25\,\mathrm{meV}$$
$$a = 0.125 \times 10^{-12}\,\mathrm{V\,cm^2}.$$

It is this linear relationship which we will use in our future calculations.

Finally, we need the relationship between n_S and the gate-channel voltage V_G. This means the potential across supply layer and spacer has to be included in the calculation.

We consider a structure where the supply layer is homogeneously doped and the spacer undoped. Integrating Poisson's equation twice, we find a parabolic potential in the supply layer, and a linear potential in the spacer (see Figure 2.23).

The built-in voltage drop over the AlGaAs layer V_2 is

$$V_2 = \frac{q\,N_D}{2\epsilon_2} d_d^2 - \mathcal{E}_S(d_d + d_i),$$

with $\mathcal{E}_S = q\,n_S/\epsilon_1$. ϵ_1 is the dielectric constant in the small-bandgap material (here, GaAs) and ϵ_2 is the corresponding value in the large-bandgap region (here, AlGaAs).

For the relationship between E_F and n_S, we use the linear approximation, Equation (2.40). Solving for n_S,

$$n_S = \frac{\epsilon_1}{q\left(d_d + d_i + \frac{\epsilon_2\,a}{q}\right)} \left(\frac{q\,N_D}{2\,\epsilon_2} d_d^2 + V_G - \Phi_b - \frac{\Delta E_{F0} - \Delta E_C}{q} \right). \tag{2.41}$$

We introduce a *threshold voltage* V_{off} as the gate-channel voltage where the interface carrier density disappears:

$$V_{off} = \Phi_b + \frac{\Delta E_{F0} - \Delta E_C}{q} - \frac{q\,N_D}{2\,\epsilon_2} d_d^2. \tag{2.42}$$

For simplification, we define a *virtual increase of the supply layer thickness*:

$$\Delta d = \frac{\epsilon_2\,a}{q}. \tag{2.43}$$

We can now write Equation (2.41) in a more compact form:

$$n_S = \frac{\epsilon_1}{q} \frac{V_G - V_{off}}{d_d + d_i + \Delta d}. \tag{2.44}$$

The threshold voltage can be controlled via the thickness of the doped layer. Calculate the supply layer thickness where $V_{off} = 0$:

$$d_{d0} = d_d(V_{off} = 0) = \sqrt{\frac{2\,\epsilon_2}{N_D\,q}\left(\Phi_b + \frac{\Delta E_{F0} - \Delta E_C}{q}\right)}. \tag{2.45}$$

If now:

$d_d > d_{d0}$: The HEMT is normally on or operating in 'depletion-mode' – it will pass drain current for $V_{GS} = 0$.

$d_d < d_{d0}$: The HEMT is normally off or operating in 'enhancement-mode' – it will not pass drain current for $V_{GS} = 0$.

The threshold voltage in practical HEMTs is often tailored for the maximum transconductance to occur for $V_{GS} = 0$, which implies $V_{off} < 0$, as we will see further down.

Gate-channel capacitance. The gate-channel capacitance can be easily calculated by differentiating the charge in the 2DEG with respect to V_G:

$$C_0 = q\,W_G\,L_G \frac{dn_S}{dV_G} = \epsilon_2 \frac{W_G L_G}{(d_d + d_i + \Delta d)}, \tag{2.46}$$

for $V_G > V_{off}$.

For $V_G \leq V_{off}$, the 2DEG will be depleted, and in first-order approximation, the gate-channel capacitance disappears.

A practical HEMT example

Before we continue to consider the channel current as a function of gate-source and drain-source voltages, let us briefly look at a practical transistor structure.

The structure shown in Figure 2.24 is the classic cross-section of a HEMT. The source and drain contacts are non-rectifying ('Ohmic') contacts. To facilitate a low contact resistance, they sit on highly n-doped GaAs. AlGaAs habitually forms aluminium

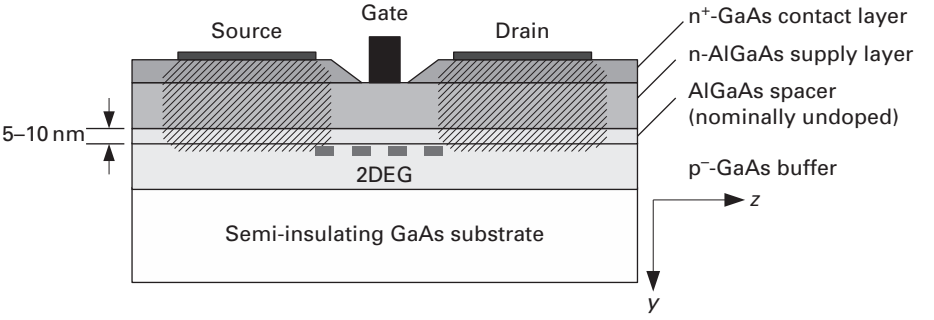

Fig. 2.24 Classic AlGaAs/GaAs HEMT structure.

oxide at the surface which would seriously increase the contact resistance. The alloying process will drive the contacts down through the heterostructure to make contact with the 2DEG. This is indicated by the shaded regions in Figure 2.24.

The gate contact must be a Schottky contact as shown. It sits in a recess through the GaAs contact layer. The recess depth controls the gate-channel separation ($d_d + d_i$) and is an important technological parameter.

The n-doped AlGaAs layer is called *supply layer* because its doping atoms supply the free carriers in the channel.

The thickness of the spacer layer (typically 5–10 nm) controls not only the reduction of Coulomb scattering, but also the transfer of electrons from the supply layer into the channel. It must be carefully optimised. The AlGaAs in the spacer is not actually *intrinsic* – it is just not intentionally doped.

The GaAs layer should be low doped, but it must be p-type. The free carriers in the channel must come from the supply layers and not from the GaAs buffer, otherwise the HEMT cannot be shut off under gate control – it exhibits *parallel conduction*. According to the mass action law,

$$n_p = \frac{n_i^2}{N_A}.$$

As in GaAs the intrinsic carrier concentration is[4] $n_i = 2.1 \cdot 10^6\,\mathrm{cm}^{-3}$, even a very low acceptor doping concentration, e.g. $N_A = 10^{15}\,\mathrm{cm}^{-3}$, will virtually eliminate free electrons in the p-buffer.

The AlGaAs/GaAs heterostructure was first grown and analysed by R. Dingle at Bell Laboratories in 1974. For a review of early work on AlGaAs/GaAs heterostructures, refer to [13]. Mimura [39] was the first to practically realise a HEMT.

Channel current – constant mobility

So far, we only considered the case of $V_{DS} = 0$. Now, we will allow $V_{DS} > 0$, i.e. a current will flow between source and drain. This current is

$$I_D = q\, n_S(z)\, v_n(\mathcal{E}_z)\, W_G = \text{const},$$

due to current continuity in the channel. Because of the voltage drop along the channel between a point z and a source $V(z)$, the density of the 2DEG now becomes z-dependent:

$$n_S(z) = \frac{\epsilon_1}{(d_d + d_i + \Delta d)\, q}[V_{GS} - V_{off} - V(z)].$$

Figure 2.25 shows the immediate channel region and the appropriate voltages affecting the channel.

As in the MESFET, we assume that the channel is

- *one-dimensional*, i.e. the electric field $\mathcal{E}$ has only a component in z direction ($\mathcal{E}_z$);
- *gradual*, i.e. the carrier densities change so slowly that diffusion currents can be neglected.

[4] www.ioffe.rssi.ru/SVA/NSM/Semicond/GaAs/bandstr.html

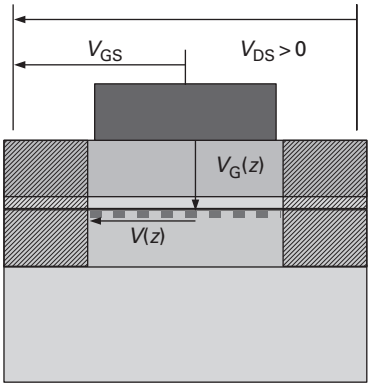

Fig. 2.25 HEMT channel region.

As in the MESFET, we will first consider the low-field case where $\mu_n = \text{const}$.

$$\begin{aligned} I_D &= q\, n_S(z)\, W_G\, \mu_n\, \mathcal{E}_z(z) \\ &= \frac{\epsilon_1\, W_G}{d_d + d_i + \Delta d} \left[V_{GS} - V_{off} - V(z)\right] \mu_n \frac{dV(z)}{dz}. \end{aligned} \tag{2.47}$$

Obeying current continuity, we find

$$I_D = \frac{\epsilon_1\, \mu_n\, W_G}{(d_d + d_i + \Delta d)\, L_G} \int_0^{L_G} \left[V_{GS} - V_{off} - V(z)\right] \frac{dV(z)}{dz} dz.$$

Let β be the transconductance parameter:

$$\beta = \frac{\epsilon_1\, \mu_n\, W_G}{(d_d + d_i + \Delta d)\, L_G}. \tag{2.48}$$

Then, using parameter substitution to integrate over V instead of z:

$$I_D = \beta \int_{V(0)}^{V(L_G)} \left[V_{GS} - V_{off} - V(z)\right] dV.$$

Note that $V(L_G) = V_{DS},\ V(0) = 0$.

Hence, we obtain the current–voltage characteristics in the linear regime (small V_{DS}):

$$I_D = \beta \left[(V_{GS} - V_{off})\, V_{DS} - \frac{V_{DS}^2}{2} \right]. \tag{2.49}$$

For very small $V_{DS} \ll 2(V_{GS} - V_{off})$, we note $I_D \approx \beta\, V_{DS}\, (V_{GS} - V_{off})$. In this regime, the HEMT acts as a 'voltage-controlled resistor'. The parameters β and V_{off} can be easily extracted if $V_{DS} = \text{const}$. This is shown in Figure 2.26. The drain current is measured for two V_{GS} while keeping $V_{DS} = \text{const} \ll 2(V_{GS} - V_{off})$. Linear extrapolation towards small V_{GS} provides V_{off} at the intersection with the V_{GS} axis. Once V_{off} is known, the transconductance parameter can be calculated as

$$\beta = -\frac{I_D(V_{GS} = 0)}{V_{DS}\, V_{off}}.$$

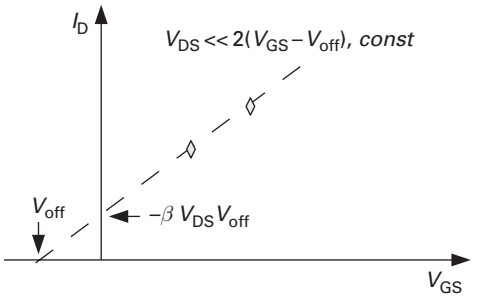

Fig. 2.26 Extraction of V_{off} and β at small V_{DS}.

Channel current – constant velocity

The *ansatz* in Equation (2.47) assumes that $n_S(z) > 0$ for all $0 < z < L_G$. If V_{DS} is sufficiently large, the gate-channel voltage may drop below V_{off} in the channel and the channel will become fully depleted. As $V(z) \leq V_{DS}$, this happens first at the drain end: $V_{GS} - V_{DS} = V_{off}$ or

$$V_{DS} \equiv V_k = V_{GS} - V_{off}. \tag{2.50}$$

This corresponds to what we saw in the MESFET, where the undepleted channel height disappeared at the drain end.

As in the MESFET, we can argue that velocity saturation prevents channel closure – $n_S(z) \to 0$ implies $\mathcal{E}_z(z) \to \infty$ due to the current continuity requirement, so that the constant-mobility assumption breaks down much earlier and $v_n(z) \to v_{sat}$.

Using a two-region model for the electron velocity, where mobility is constant for $|\mathcal{E}_z| < \mathcal{E}_{crit}$ and velocity is constant for $|\mathcal{E}_z| > \mathcal{E}_{crit}$, we can calculate the drain-source voltage V_{DSS} for which velocity saturation happens at the drain end of the channel ($z = L_G$). At this point the local electric field is $\mathcal{E}_z(z = L_G) = \mathcal{E}_{crit}$. Using the constant-mobility current Equation (2.49), we find

$$I_D = \frac{\epsilon_1 \mu_n W_G}{L_G(d_d + d_i + \Delta d)} \left[(V_{GS} - V_{off}) V_{DSS} - \frac{V_{DSS}^2}{2} \right].$$

On the other hand, I_D can be calculated using a constant-velocity *ansatz*:

$$I_D = q\, n_S\, W_G\, v_{sat}.$$

In the two-region model, $\mu_n \mathcal{E}_{crit} \equiv v_{sat}$. Further, n_S can be calculated from Equation (2.44) using $V_G = V_{GS} - V_{DSS}$, so that

$$I_D = \frac{\epsilon_1\, W_G\, \mu_n}{(d_d + d_i + \Delta d)} (V_{GS} - V_{off} - V_{DSS})\, \mathcal{E}_{crit}\, L_G.$$

The drain current expressions for constant mobility and constant velocity must be equal for $V_{DS} = V_{DSS}$, because we are transitioning from the constant mobility to

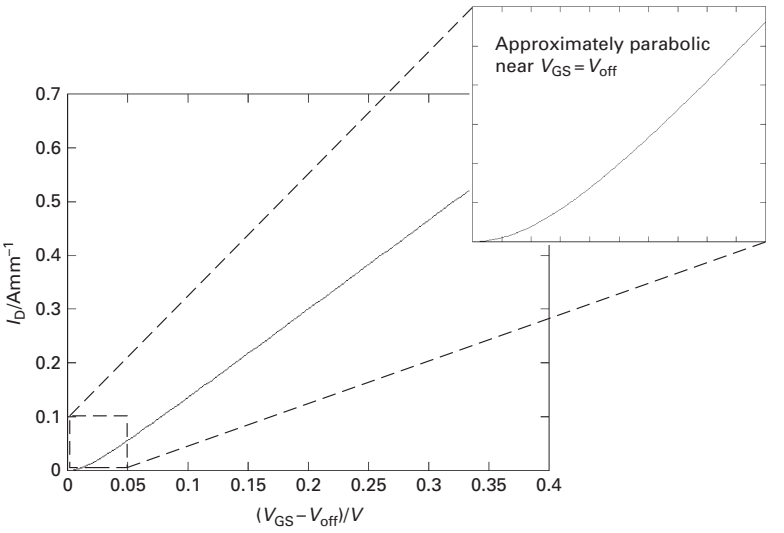

Fig. 2.27 Sample calculation of I_D following Equation (2.52). Parameters are $\beta = 83\,\mathrm{S\,(Vm)^{-1}}$, $\mathcal{E}_{crit} = 1\,kV\,\mathrm{cm}^{-1}$ and $L_G = 0.2\,\mu\mathrm{m}$.

the constant-velocity regime. This leads to a quadratic equation in V_{DSS}, which we can solve to find the necessary drain-source current for velocity saturation to set in:

$$V_{DSS} = V_0 \left[1 + \frac{V_{GS} - V_{off}}{V_0} - \sqrt{1 + \left(\frac{V_{GS} - V_{off}}{V_0} \right)^2} \right], \tag{2.51}$$

where $V_0 = \mathcal{E}_{crit}\, L_G$.

For the corresponding drain current in the velocity-saturated case, we finally find

$$I_{DSS} = \beta\, V_0^2 \left[\sqrt{1 + \left(\frac{V_{GS} - V_{off}}{V_0} \right)^2} - 1 \right]. \tag{2.52}$$

In Figure 2.27, a sample calculation has been performed using Equation (2.52). Note that for the most part, I_D is a strictly linear function of V_{GS}. Very close to V_{off}, a more parabolic behaviour dominates.

2.3.3 Small-signal parameters

Transconductance in the saturated regime is calculated by differentiating Equation (2.52) with respect to V_{GS}:

$$g_m = \frac{d\,I_{DSS}}{d\,V_{GS}} = \frac{\beta\,(V_{GS} - V_{off})}{\sqrt{1 + \left(\frac{V_{GS} - V_{off}}{V_0} \right)^2}}. \tag{2.53}$$

For large V_{GS}, transconductance is predicted to be constant, while it is approximately linearly dependent on V_{GS} for small V_{GS}. We will see later that this ideal behaviour is superseded by parasitic effects, however.

The gate-source capacitance C_{GS} can be found by differentiating the total channel charge Q_T with respect to V_{GS}:

$$C_{GS} = \frac{dQ_T}{dV_{GS}} = \frac{d}{dV_{GS}} W_G q \int_0^{L_G} n_S(z) dz.$$

Particularly simple – and practically important – is the case of velocity saturation in the whole channel. As

$$I_D = q\, n_S(z)\, v_{sat}\, W_G = \text{const} \Rightarrow n_S(z) = \text{const} = n_{SS},$$

where

$$n_{SS} = \frac{\beta\, V_0^2}{q\, v_{sat}\, W_G} \left[\sqrt{1 + \left(\frac{V_{GS} - V_{off}}{V_0} \right)^2} - 1 \right]$$

and therefore

$$Q_T = q\, n_{SS}\, W_G\, L_G = \frac{\beta\, V_0^2\, L_G}{v_{sat}} \left[\sqrt{1 + \left(\frac{V_{GS} - V_{off}}{V_0} \right)^2} - 1 \right].$$

The gate-source capacitance becomes in this case

$$\begin{aligned} C_{GS} &= \frac{dQ_T}{dV_{GS}} \\ &= \frac{L_G}{v_{sat}} \frac{\beta\, V_0\, (V_{GS} - V_{off})}{\sqrt{(V_{GS} - V_{off})^2 + V_0^2}} \\ &= \frac{L_G}{v_{sat}} g_m. \end{aligned} \tag{2.54}$$

As in case of the MESFET, we find for the transit time of carriers under the gate:

$$\tau_T = \frac{L_G}{v_{sat}} = \frac{C_{GS}}{g_m}.$$

The gate-drain capacitance can be calculated similarly:

$$C_{GD} = \frac{dQ_T}{dV_{GD}} = 0$$

in this simple model because $Q_T \neq f(V_{GD})$. In reality, C_{GD} is non-zero because of the geometric capacitance between the metal contacts and other parasitic effects. As in the MESFET, $C_{GD} \ll C_{GS}$ in saturation.

For the small-signal equivalent circuit, we can use the same topology as for the MESFET. Accordingly, the transit frequency can be approximated by

$$f_T = \frac{g_m}{2\pi C_{GS}}. \tag{2.55}$$

2.3.4 'High electron mobility'?

The common name 'high electron mobility transistor' deserves some critical reflection. Remember that while Coulomb scattering is the dominant mobility-limiting mechanism at cryogenic temperatures, phonon scattering is dominant at room temperature (Figure 2.16).

Realistic enhancement factors for the electron mobility in HEMTs are

- a factor of two at room temperature;
- up to a factor of 100 at cryogenic temperatures (e.g. 77 K – liquid nitrogen).

Furthermore, we found that in short-channel FETs velocity saturation dominates in the channel – hence the enhancement in mobility has two major advantages:

(i) reduction of series resistances, most importantly R_S;
(ii) lowering of the critical field for velocity saturation, i.e. saturated velocity will be reached sooner.

However, there are other advantages of the HEMT structure which are also significant:

- The carrier distribution is similar to that of a pulse-doped MESFET – we expect a constant transconductance g_m and therefore a high linearity. There are parasitic effects which prevent this from happening – more about this later.
- In active operation, the supply layer is fully depleted and hence the gate-source capacitance C_{GS} should be constant. Again, this is not quite true in reality (see p. 81).
- The potential barrier towards the substrate reduces carrier injection into the substrate and increases output resistance.

Note that, compared to a MESFET with an epitaxially grown channel, the HEMT structure is technologically not much more complicated.

As an aside, the HEMT has many other names and has jokingly been called a *multi-acronym device* (MAD). To name but a few:

HFET	heterostructure field effect transistor
MODFET	modulation-doped field effect transistor
TEGFET	two-dimensional electron gas FET
SDFET	selectively doped field effect transistor

2.3.5 Non-ideal behaviour

In the following pages, we will discuss how in practical HEMTs the experimentally observed behaviour deviates from the theory developed so far. The explanation of these non-ideal features will have important implications for the design of optimised HEMT devices.

Non-ideal HEMT behaviour for large V_{GS}

From simple HEMT theory as outlined above, we expect that for sufficiently large $V_{GS} - V_{off}$, the drain current increases linearly with V_{GS} and hence the transconductance is constant. Also, we would expect that the gate-source capacitance is constant in the same region.

Experimentally, however, transconductance and gate-source capacitance show the behaviour in Figure 2.28 [1]: after a sharp increase above the threshold voltage, the transconductance goes through a maximum, then decreases again for higher V_{GS}. The gate-source capacitance initially tracks the transconductance, as predicted by Equation (2.54) (save for a constant parasitic contribution), but then increases further for higher V_{GS}.

This compression of the transconductance is due to a 'parasitic MESFET' effect [33]. To understand its origin, please consider Figure 2.29.

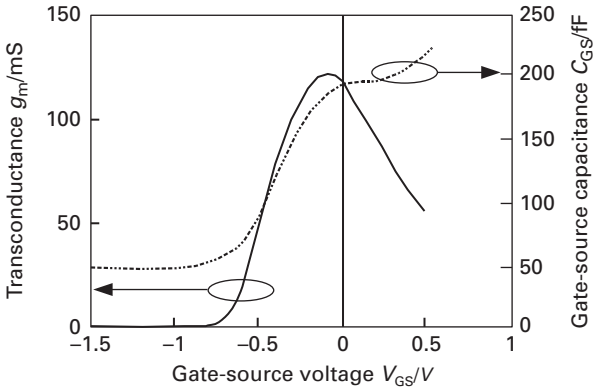

Fig. 2.28 Experimental transconductance and gate-source capacitance versus gate-source voltage.

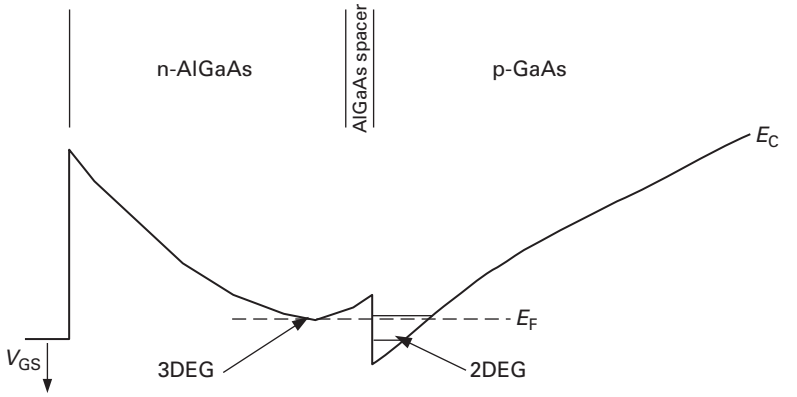

Fig. 2.29 Conduction band diagram of a HEMT under high V_{GS}. The arrows indicate the locations of the 2DEG and three-dimensional electron gas (3DEG).

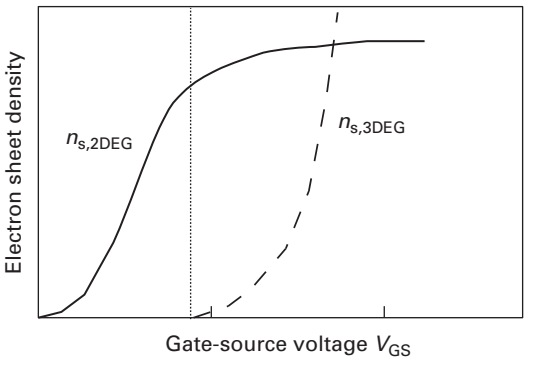

Fig. 2.30 Dependence of the electron sheet densities in the 2DEG and 3DEG as a function of the gate-source voltage.

If V_{GS} is sufficiently high, the conduction band minimum in the AlGaAs supply layer dips below the Fermi level. At this point, the free electron density in the supply layer will rapidly increase, the supply layer is no longer depleted. Because the free electron population in the AlGaAs conduction band minimum has very little confinement, it is referred to as the *three-dimensional electron gas* or *3DEG*. In its low confinement, this channel is very similar to a MESFET's, hence the term *parasitic MESFET*.

Once the 3DEG builds up, it electrostatically shields the 2DEG from the gate electrode – the 2DEG density $n_{S,2D}$ saturates; any further increase in charge density due to a further increase in V_{GS} will benefit only $n_{S,3D}$. This is schematically shown in Figure 2.30.

The rise of the 3DEG has two substantial effects:

- Because the mobility is much lower in the ternary AlGaAs supply layer than in the GaAs channel region, the resulting transconductance due to the 3DEG channel is lower, causing the overall transconductance to decrease.
- The additional charge under the gate leads to a strong increase in the gate-source capacitance.

Recall Equation (2.55) – the simultaneous decrease in transconductance and increase in gate-source capacitance will have a very negative impact on the transit frequency f_T. Using the data from Figure 2.28, this is exemplified in Figure 2.31.

The transit frequency, which in our simple theory was predicted to be independent of frequency, now shows a pronounced maximum, which occurs for gate-source voltages slightly lower than the transconductance maximum. For the design of high-speed circuits, this is an important observation.

The ungated FET structure can be used as a model for the situation at the onset of the parasitic MESFET effect, where $n_{S,3D} = N_D$. Figure 2.32 shows the conduction band diagram.

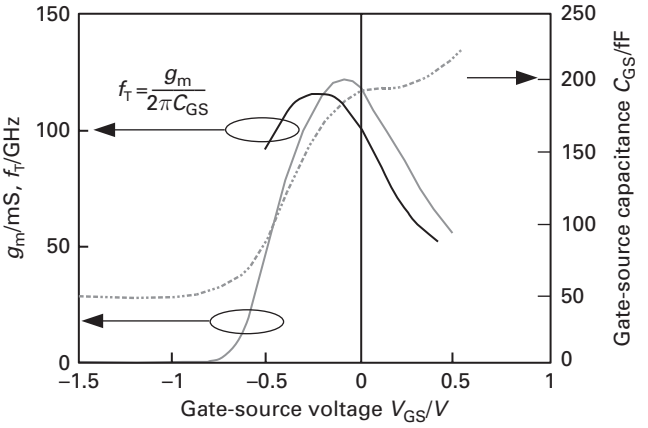

Fig. 2.31 Calculated transit frequency of the HEMT in Figure 2.28 versus gate-source voltage.

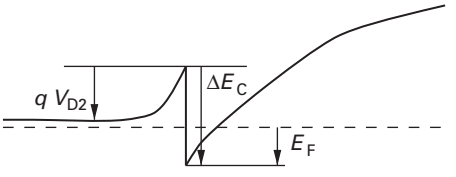

Fig. 2.32 Conduction band diagram of an ungated HEMT structure.

We find that

$$E_F = \Delta E_C - q V_{D2} - kT \ln \frac{N_{C,AlGaAs}}{N_D} \approx \Delta E_C - q V_{D2},$$

for large N_D. V_{D2} is the built-in potential in the large-band-gap part of the heterostructure. Using Equation (2.40),

$$n_{S,max} \approx \frac{\Delta E_C - \Delta E_{F0} - q V_{D2}}{q\,a}.$$

To maximise $n_{S,max}$, we must therefore choose a material combination with large ΔE_C. In a conventional HEMT structure, $n_{S,max} \approx 1 \cdot 10^{12}\,\mathrm{cm}^{-2}$.

Trapping effects

In an $Al_xGa_{1-x}As/GaAs$ HEMT, we may obtain a larger ΔE_C by increasing the Al mole fraction x. However, consider the following detrimental effects.

For $x_{Al} > 0.3$, the *effective energy depth of the donor level* increases – the number of free carriers provided by a given doping density N_D will decrease.

Earlier, for $x_{Al} > 0.25$, the *density of deep traps* (*DX centres*) will increase. These traps are energy states in the forbidden gap which can interact with the valence or conduction band. In this case, they are closer to the conduction band, at an energetic depth E_T – they are 'donor-like'. The 'X' denotes that their physical origin was long unknown. A trap will capture a free electron from the conduction band and eventually re-emit it.

The characteristic time constant for re-emission is strongly temperature-dependent:

$$\tau_{\mathrm{RE}}(T) = \tau_0 \exp\left(\frac{E_{\mathrm{T}} + E_{\mathrm{B}}}{kT}\right), \tag{2.56}$$

where E_{B} is an additional energy barrier for re-emission. In AlGaAs, $E_{\mathrm{T}} \approx 50\,\mathrm{meV}$ and $E_{\mathrm{B}} \approx 300\,\mathrm{meV}$.

When reducing the temperature, formerly free carriers will be 'frozen' and as a consequence, will no longer be available for the channel. This can be described by a shift in threshold voltage:

$$\Delta V_{\mathrm{off}} = -\frac{q}{2\epsilon} N_{\mathrm{DT,ion}}\, d_{\mathrm{d}}^2.$$

The density of ionised traps is, using Fermi–Dirac statistics,

$$N_{\mathrm{DT,ion}} = \frac{N_{\mathrm{DT}}}{1 + \exp\left(\frac{E_{\mathrm{F}} - E_{\mathrm{T}}}{kT}\right)}.$$

Note that the trap density N_{DT} has been experimentally observed to be proportional to the donor density N_{D}. It is now accepted that the donor atoms themselves introduce two different energy levels in the forbidden gap in AlGaAs: a shallow one associated with the Γ-minimum (the direct minimum) – this is the proper donor level – and a deep energy state associated with the L-minimum (an indirect minimum) – this is the DX centre [5].

The effect of DX centres in the supply layer made early HEMT structures very problematic in cryogenic operation.

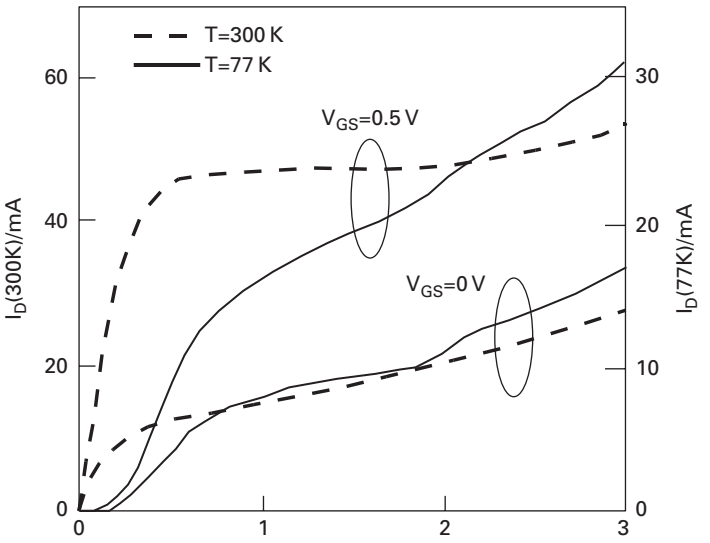

Fig. 2.33 Output I–V characteristics of a HEMT device with a low-temperature current collapse phenomenon (A. Belache, A. Vanoverschelde, G. Salmer and M. Wolny, *IEEE Transactions on Electron Devices*, Vol. ED-38, No.1, pp. 3–13, January 1991. ©1991 IEEE).

Figure 2.33 shows an example of a device showing such a current collapse phenomenon [3]. Apart from the change in output conductance in saturation and the various 'kink' effects, which shall not be discussed here, we note

- At low V_{DS}, the output conductance in the linear regime decreases considerably – this is due to an increase in the source resistance R_S. The decrease in R_S with decreasing T is unexpected, as the mobility itself will increase. The decrease in the free carrier concentration, however, dominates.
- In the saturated regime, $V_{DS} > 0.5\,\text{V}$, the transconductance also decreases significantly with decreasing temperature.

The occurrence of DX centres is closely linked to the use of AlGaAs as the barrier material – other supply layer materials such as GaInP do not show this effect and will correspondingly fare better in their low-temperature performance [7].

2.3.6 Structural HEMT variations

Increasingly, structural variations of the original HEMT concept are being used to circumvent the non-ideal behavioural effects explained above and to improve performance.

Pulse-doped HEMT structure

Due to the severeness of the DX centre limitation, a method to eliminate this limitation has the highest priority.

Because $N_{DT} \sim N_D$, the trap-induced threshold voltage shift is

$$\Delta V_{off} \sim N_D\, d_d^2.$$

On the other hand, the supply layer must be able to supply the necessary carrier density in the 2DEG:

$$N_D\, d_d > n_S.$$

If, therefore, we concentrate the doping in a narrow sheet – increase N_D and decrease d_d – the trap-induced threshold voltage shift can be drastically reduced.

This concept leads to the *delta-doped* (or pulse-doped) HEMT structure (see Figure 2.34).

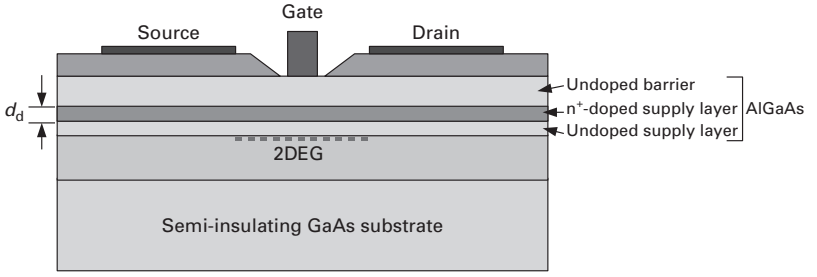

Fig. 2.34 Layer structure of a delta-doped HEMT.

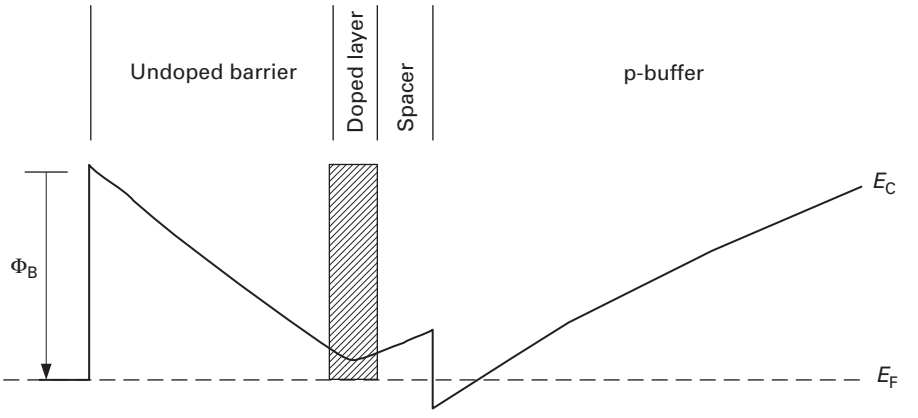

Fig. 2.35 Conduction band diagram of a pulse-doped HEMT structure.

The restriction of doping to only a narrow sheet of the wide-gap layer, of course, also modifies the band structure. The Poisson equation

$$\frac{d^2V}{dy^2} = -\frac{\rho}{\epsilon}$$

tells us that the potential V will have a linear y dependence if $\rho \simeq 0$. Where $\rho \neq 0 = \text{const}$, V will have a parabolic dependence on y. These principles are visible in the conduction band diagram of a pulse-doped HEMT structure (see Figure 2.35).

As an additional advantage, the pulse-doped HEMT can be expected to have lower gate leakage because of the lower doping of the region immediately under the gate.

Pseudomorphic HEMT structure

So far, the AlGaAs/GaAs HEMT structures considered were lattice-matched – a significant advantage of the (Al,Ga)As material system is that the lattice constant is almost independent of the Al content.

We will now deliberately leave the lattice match principle behind and allow for material combinations which are lattice-mismatched, but where the lattice difference is accommodated by elastic deformation of the crystal – *pseudomorphic* structures. This gives us greater flexibility in the choice of materials.

Let us replace the GaAs channel in a conventional HEMT with an InGaAs channel. This leads to a double-heterostructure because the GaAs buffer and substrate shall be maintained. In Figure 2.36, the conduction band diagram of an example structure combining the pseudomorphic channel layer with a pulse-doped barrier is shown.

Compared to the conventional HEMT, this structure has several advantages:

- The significantly higher conduction band discontinuity increases the maximum density of the 2DEG from about $1 \cdot 10^{12}\,\text{cm}^{-2}$ for the conventional HEMT to about $2 \cdot 10^{12}\,\text{cm}^{-2}$ for the pseudomorphic HEMT as shown.
- The low Al content in the supply layer reduces the density of DX centres.

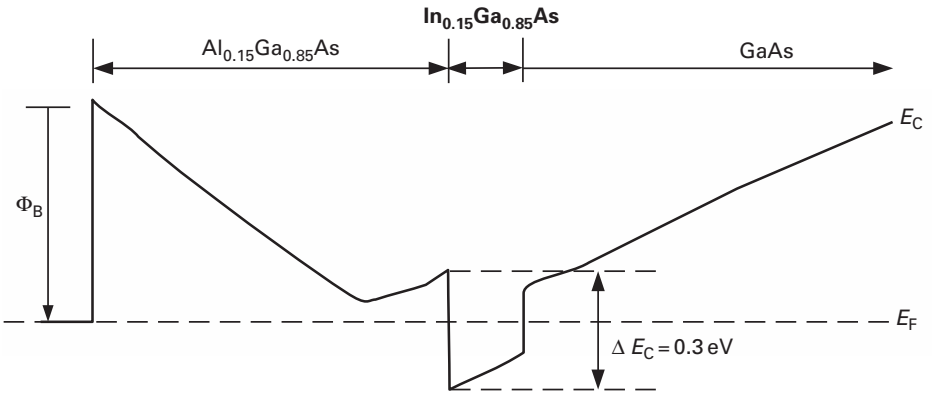

Fig. 2.36 Conduction band diagram of a pseudomorphic HEMT structure with a pulse-doped barrier.

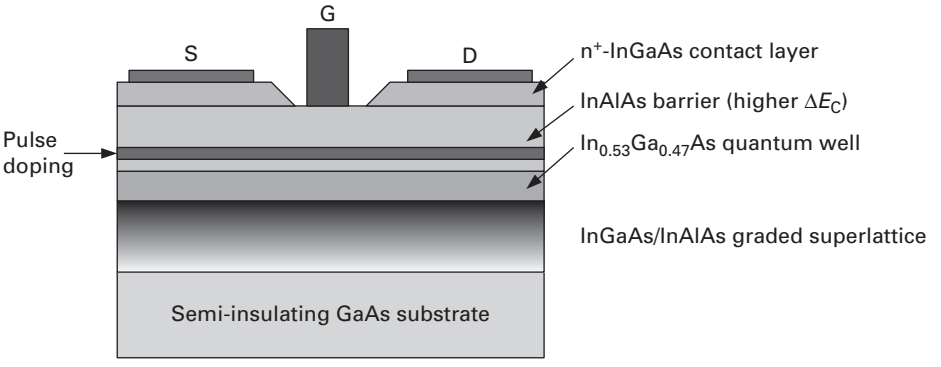

Fig. 2.37 Layer structure of a metamorphic HEMT.

- The addition of In in the channel enhances the *low-field mobility* and, to a lesser extent, the peak velocity in the channel.
- The added heterostructure towards the GaAs buffer reduces injection of carriers into the buffer and substrate.

Higher ΔE_C are possible with higher In concentrations in the channel. However, note that with increasing In content, the InGaAs layer thickness must be reduced. In practical pseudomorphic HEMTs on GaAs substrates, $x_{In} = 0.15 \ldots 0.25$. This increases the maximum density of the 2DEG to $n_{S,max} \simeq 2 \cdot 10^{12}\,cm^{-2}$.

Metamorphic HEMT

Mobility in the channel would benefit from even higher In mole fractions, e.g. $x_{In} = 0.53$, as in InGaAs lattice matched to InP. However, InP substrates are still considerably more expensive than GaAs wafers.

The metamorphic HEMT concept enables high In mole fractions in the channel layer on GaAs substrates, through a modification of the lattice constant in a *graded superlattice*. Such a layer structure is shown in Figure 2.37.

An InAlAs/InGaAs superlattice with varied thickness and composition is grown on top of the GaAs substrate such that the effective lattice constant (modified by composition and built-in mechanical strain) is adapted from that of GaAs to (in this case) the one of InGaAs with an In mole fraction of 0.53. The use of a low-temperature grown superlattice allows the change of the lattice constant while keeping the density of deformation-related crystal defects low.

As in all modern HEMTs, the barrier layer is assumed to be pulse-doped. Another modification is the use of InAlAs instead of AlGaAs as barrier material. InAlAs has a higher conduction band discontinuity towards InGaAs than AlGaAs for comparable Al mole fractions; furthermore, the InAlAs/InGaAs heterostructure stack is easier to grow. As contact layer material, InGaAs is used here because it has a much lower Schottky barrier height than GaAs.

2.3.7 CAD modelling of HEMTs

Due to the similarity of the HEMT to MESFETs and (as we will see in the next section) MOSFETs, CAD models of these two devices are often re-used to simulate HEMTs.

In the discussion of CAD modelling, we will go beyond the rather simple models provided for the MESFET and describe a high-accuracy semi-empirical approach. It is equally suitable for an enhanced precision model of the MESFET.

Static current equations

A HEMT-specific problem is the simulation of transconductance suppression at large V_{GS} (see p. 81). A suitable drain current expression which accommodates this (the discussion follows I. Kallfass [27]) is

$$I_{DS}(V_{GS}) = \beta\ (V_{GS} - V_{off})^{\lambda/(1+\xi\ (V_{GS}-V_{off}))} \tag{2.57}$$

in saturation – neglecting the V_{DS} dependence of I_{DS}.

The non-saturated region at small V_{DS} can be included using the *tanh* term already discussed in the context of the MESFET Curtice model:

$$I_{DS} = \beta\ (V_{GS} - V_{off})^{\lambda/(1+\xi\ (V_{GS}-V_{off}))}\ \tanh\ (\alpha\ V_{DS}). \tag{2.58}$$

In real devices, the drain-source voltage has a non-linear influence (so far neglected) on the current in saturation, e.g. through impact ionisation effects. In the non-saturated regime, on the other hand, the tanh ($\alpha\ V_{DS}$) expression is not always sufficient, because the V_{GS} dependence is not adequately modelled. An effective voltage V_{eff} is introduced, replacing the simple $V_{GS} - V_{off}$ term in Equation (2.58):

$$\begin{aligned} I_{DS} &= \beta\ V_{eff}^{\frac{\lambda}{1+\mu V_{DS}^2+\xi V_{eff}}}\ \tanh\left[\alpha\ V_{DS}\ (1+\zeta\ V_{eff})\right] \\ V_{eff} &= \frac{1}{2}\left(V_{GSt} + \sqrt{V_{GSt}^2 + \delta^2}\right) \\ V_{GSt} &= V_{GS} - (1+\beta_r^2)\ V_{T0} + \gamma\ V_{DS}. \end{aligned} \tag{2.59}$$

This expression, introduced by Cojocaru and Brazil in 1997 [8], is called the *COBRA* current equation. Its advantage is that it is continuous in the entire bias plane, and also its derivatives are continuous, which is very important for simulations of the non-linear behaviour of circuits.

$\beta, \lambda, \mu, \xi, \alpha, \zeta, \delta, \gamma$ and V_{T0} are model parameters to be extracted by measurements. β_r is equal to β, but dimensionless. They affect I_{DS} as follows:

- α, ζ affect the linear regime of the device – α is the main parameter modelling the V_{DS} dependence; ζ modifies the V_{GS}-dependent behaviour.
- β is the main transconductance parameter.
- ξ is the parameter which adjusts the transconductance compression.
- γ introduces a V_{DS} dependence to the drain current in the saturated regime and is hence responsible for the output conductance.
- μ equally introduces a V_{DS} dependence in the linear regime. It is used to model impact ionisation effects in the saturated regime.
- λ adjusts the curvature of $I_{DS}(V_{GS})$ for small V_{DS} and close to threshold.
- V_{T0} is the threshold voltage for small V_{DS}.

The drain current source $I_{DS} = f(V_{GS}, V_{DS})$ is embedded into an equivalent circuit to account for the series resistances and the non-linear gate-source and gate-drain contacts. This is shown in Figure 2.38. Note that the controlling voltages drop between the internal nodes! The diodes, D_{GS} and D_{GD}, are used to model the non-linear gate current. Breakdown behaviour can equally be included here:

$$I_{GS}(V_{GS}) = I_{sgs}\left(\exp\frac{V_{GS}}{n_{id}\,V_T} - 1\right) + I_{bv}\exp\left(-\frac{V_{GS} - V_{bv}}{n_{bv}\,V_T}\right)\frac{V_{GS}}{V_{bv}} \tag{2.60}$$

$$I_{GD}(V_{GD}) = I_{sgd}\left(\exp\frac{V_{GD}}{n_{id}\,V_T} - 1\right) + I_{bv}\exp\left(-\frac{V_{GD} - V_{bv}}{n_{bv}\,V_T}\right)\frac{V_{GD}}{V_{bv}}, \tag{2.61}$$

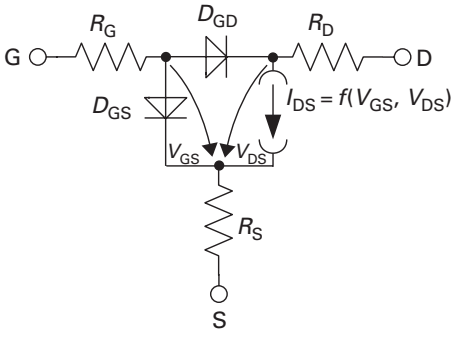

Fig. 2.38 Quasi-static equivalent circuit used in the COBRA model.

where I_{sgs} and I_{sgd} are saturation currents for the gate-source and gate-drain diodes, respectively, and n_{id} is the emission factor for these diodes. The second term in each equation models breakdown with an exponential diode term. V_{bv} is the breakdown voltage and I_{bv} and n_{bv} are used to model the current increase beyond breakdown. The very last product term simply makes sure that the breakdown current is zero, if either V_{GS} or V_{GD} are zero in Equations (2.60) or (2.61), respectively, but has no other major effect.

Non-linear capacitance equations

To properly model the non-linear behaviour in any FET, we need to account for several contributions:

- parasitic (non-bias-dependent) capacitance,
- the junction capacitance,
- the change in channel charge with varying voltage.

The first two are straightforward to model: the parasitic capacitance is C_{pgs} for the gate-source diode and C_{pgd} for the gate-drain diode. For the junction capacitance, the common form also implemented in SPICE is used:

$$C(V) = \frac{C_0}{\left(1 - \frac{V}{V_{bi}}\right)^m},$$

where C_0 is the capacitance without any external voltage, V_{bi} is the built-in voltage of the junction and m is an exponent.

Inclusion of the channel charge is much more complicated. For once, the channel charge depends on V_{GS} and V_{GD} simultaneously. Then, charge conservation needs to be satisfied. This means [28]

$$\frac{\delta C_{GS}}{\delta V_{GS}} = \frac{\delta^2 Q_G}{\delta V_{GS}\delta V_{GD}} = \frac{\delta^2 Q_G}{\delta V_{GD}\delta V_{GS}} = \frac{\delta C_{GD}}{\delta V_{GS}}. \tag{2.62}$$

Any empirical expressions for $C_{GS}(V_{GS}, V_{GD})$ or $C_{GD}(V_{GS}, V_{GD})$ must fulfil Equation (2.62).

Figure 2.39 shows gate-source and gate-drain capacitances experimentally determined from S-parameter measurements, as a function of V_{GS}, for V_{DS} values in the linear and the saturated regime of FET operation. Note the rather strong variation near pinch-off, and generally in the linear regime.

In the following empirical equations [29], the $\tanh(x)$ function is again exploited, similar to the Curtice models.

$$\begin{aligned} C_{GS}(V_{GS}, V_{GD}) = C_{pgs} &+ \frac{C_{gs1}}{\left(1 - \frac{V_{GS}}{V_{bi}}\right)^m} \\ &+ C_{gs2}\{1 + \tanh[\kappa(V_{GS} - V_{t2})]\} \\ &+ C_S(V_{GS})\{1 + \tanh[\iota(V_{GS} - V_{GD} - V_{t4})]\} \\ &- \frac{\delta C_S(V_{GS})}{\delta V_{GS}}\left(V_{GD} - \frac{1}{\iota}\ln\{\cosh[\iota(V_{GS} - V_{GD} - V_{t4})]\}\right) \end{aligned} \tag{2.63}$$

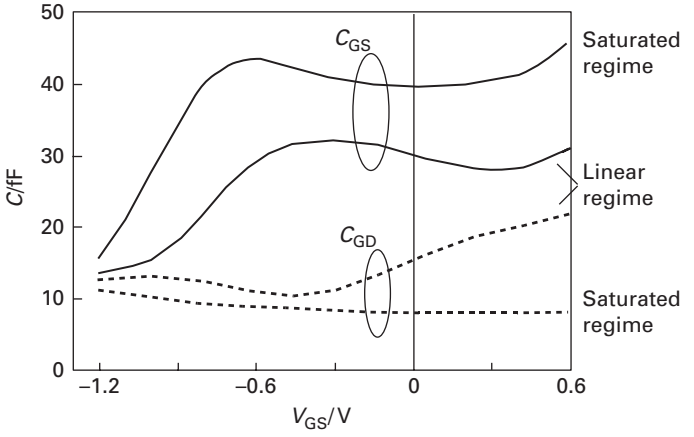

Fig. 2.39 Experimentally determined C_{GS} and C_{GD} of pseudomorphic GaAs HEMT ($L_G = 0.15\,\mu m$, $W_G = 2 \times 20\,\mu m$).

$$C_{GS}(V_{GS}, V_{GD}) = C_{pgd} + \frac{C_{gd1}}{\left(1 - \frac{V_{GD}}{V_{bi}}\right)^m} + C_{gd2}\{1 + \tanh[\kappa(V_{GD} - V_{t5})]\} - C_S(V_{GS})\{1 + \tanh[\iota(V_{GS} - V_{GD} - V_{t4})]\}. \quad (2.64)$$

The capacitance

$$C_S(V_{GS}) = C_3 V_{eff}^{\psi}$$

with

$$V_{eff} = \frac{1}{2}\left(v_{GS} - V_{t3} + \sqrt{(V_{GS} - V_{t3})^2 + \theta^2}\right)$$

is closely related to the drain saturation current (compare Equation (2.59)).
$C_{gs1}, C_{gs2}, C_{gd1}, C_{gd2}, m, V_{bi}, V_{t2}, V_{t3}, V_{t4}, V_{t5}, \iota, \kappa, \theta$ and ψ are fitting parameters.

The non-linear capacitances, along with an additional parasitic drain-source capacitance C_{DS} and a parasitic channel resistance R_i, have been combined in Figure 2.40 to form a basic non-linear dynamic model of the HEMT. More bias-independent parasitic parameters may be added, as needed.

2.3.8 MESFET versus HEMT: a small-signal comparison

When the non-linear equivalent circuit in Figure 2.40 is linearised in a given bias point, the resulting small-signal equivalent circuit is identical to that derived for the MESFET in the previous section, with the exception of the domain capacitance C_{DC}, which is often neglected anyhow. Many results obtained for the MESFET can therefore be directly applied. Rather than repeating the results here, let us discuss how the achievable small-signal performance differs between MESFET and HEMT.

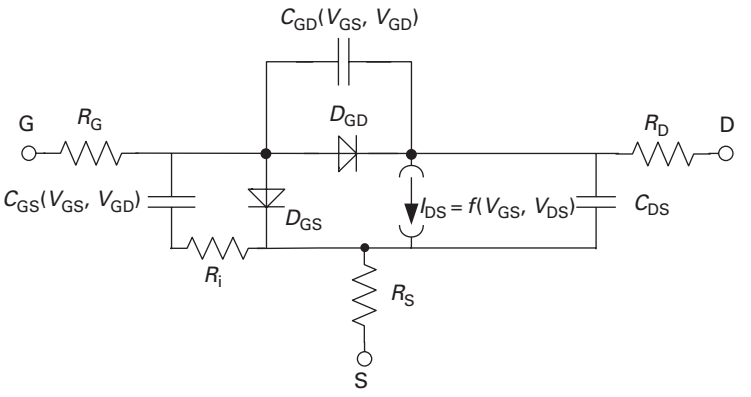

Fig. 2.40 A dynamic non-linear model of the HEMT.

As discussed, the higher low-field mobility affects the series resistances which represent semiconductor regions outside of the velocity-saturated channel. These are R_S, R_i and, of lesser importance, R_D. Equally, it increases the transconductance g_m (see Equation (2.53)).

The larger potential barrier between the channel and the substrate reduces the output conductance g_{ds} in the HEMT.

These findings directly translate into a significant advantage in terms of the maximum frequency of oscillation, f_{max}:

$$f_{max} = \frac{f_T}{2\sqrt{(R_G + R_S + R_i)g_{ds} + 2\pi f_T R_G C_{DG}}}.$$

The gate resistance R_G is, of course, independent of the device structure.

The HEMT structure also has a positive impact on the noise performance. This can be shown using the Fukui equation already introduced for the MESFET:

$$\begin{aligned} F_{min} &= 1 + k_F \frac{f}{f_T}\sqrt{g_m(R_G + R_S)} \\ &= 1 + k_F 2\pi f C_{GS}\sqrt{\frac{R_G + R_S}{g_m}}, \end{aligned}$$

using

$$f_T \sim \frac{g_m}{2\pi C_{GS}}.$$

The noise performance is improved not only by the reduction in R_S and the increase in g_m. The fitting factor k_F, which is typically 2.5 in MESFETs, decreases to $k_F = 1 \ldots 2$ in HEMTs. This is commonly explained by the higher correlation between channel noise and induced gate noise, and the reduction in channel noise due to the smaller degree of freedom of carrier movement in the 2DEG.

2.3.9 A practical HEMT example

The example chosen here is a metamorphic HEMT [4] because it illustrates many of the concepts discussed.

The device structure is shown in Figure 2.41. The rather thick (1 μm) linearly graded buffer adapts the lattice constant of the GaAs substrate to the much larger lattice constant of $In_{0.53}Ga_{0.47}As$ and $Al_{0.48}In_{0.52}As$ (the ternary compounds are lattice-matched to each other). Note the 'double doping' structure – there are δ-doped AlInAs layers below and above the InGaAs quantum well. In this case, it allows a density of the 2DEG of $n_{S,max} = 4 \cdot 10^{12}\,cm^2$, together with the excellent carrier confinement in the quantum well. The excellent confinement is due to the large conduction band discontinuity between $In_{0.53}Ga_{0.47}As$ and $Al_{0.48}In_{0.52}As$.

The ohmic contacts are placed on an $In_{0.53}Ga_{0.47}As$ cap layer, which reduces the contact resistance and shields the metal–semiconductor interface from the Al-containing alloy, which is prone to formation of Al oxides at the exposed surface. The gate contact has a T shape which reduces the series resistance of the gate stripe and hence R_G. It is again shown in Figure 2.42.

The cross-section of the gate metallisation is significantly larger than what would be possible for a simple stripe with a 250 nm footprint, due to the T-gate structure. A refractory metal is used here so that the gate can be fabricated before the ohmic contacts – this allows an easy self-alignment of the ohmic contacts with respect to the T-gate structure, minimising the distance between the source and drain contacts and the channel, reducing R_S and R_D. The surface between the gate and the ohmic contacts is passivated by a SiN layer.

The plot in Figure 2.43 shows the drain current and transconductance of the device, normalised to 1 mm gate width, at $V_{DS} = 1\,V$, which is well into the saturated regime for this device. The actual gate width of the characterised device is 20 μm. The g_m maximum is placed at $V_{GS} = 0$ – this is frequently done as it facilitates gate biasing. The threshold voltage is slightly below 0.6 V. The g_m depression at higher V_{GS} is also clearly visible.

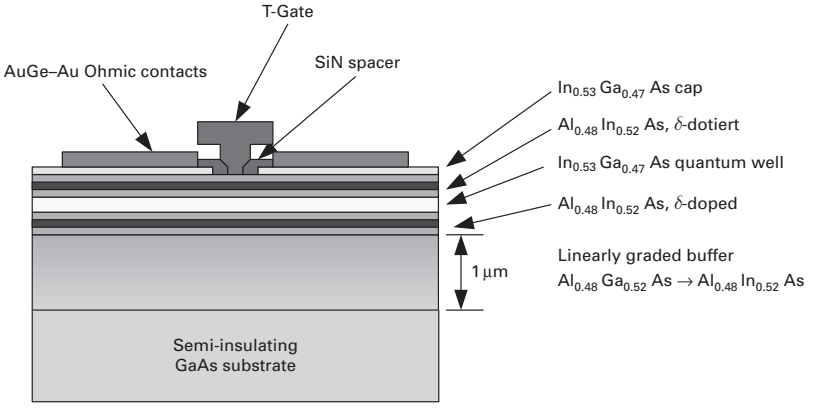

Fig. 2.41 Layer structure of the metamorphic HEMT structure discussed here.

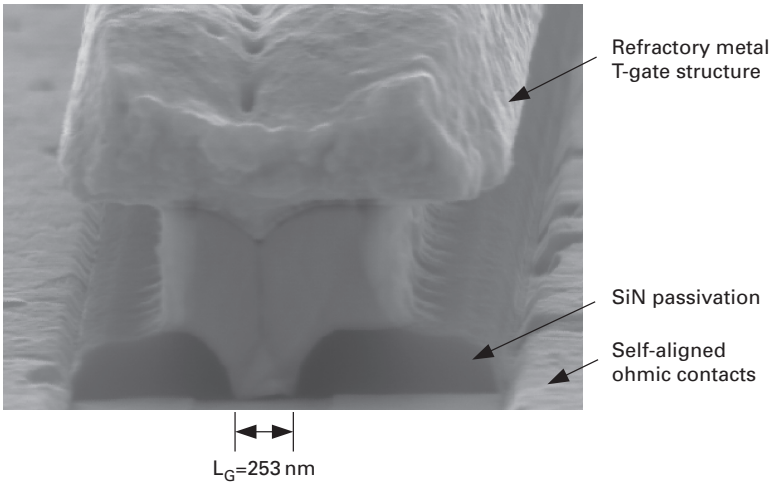

Fig. 2.42 SEM micrograph of the gate structure (F. Benkhelifa, M. Chertouk, M. Dammann, M. Massler, H. Walther and G. Weimann, *International Conference on Semiconductor Manufacturing Technology GaAs MANTECH 2001 Digest*, May 2001).

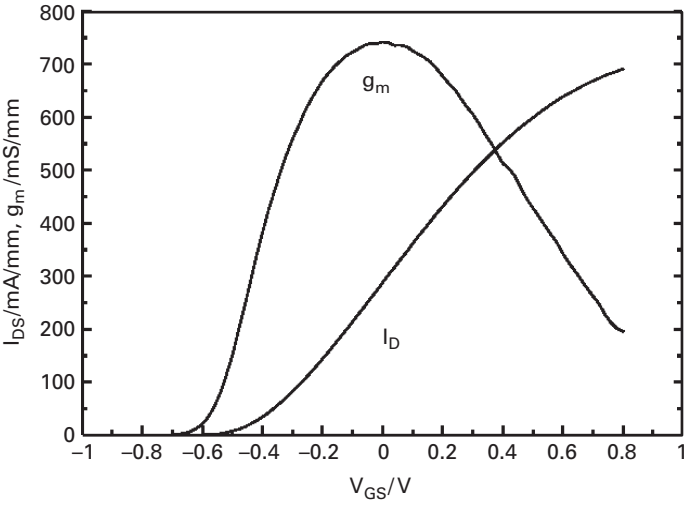

Fig. 2.43 Drain current (I_D) and transconductance (g_m) of the metamorphic HEMT, normalised to 1 mm gate width (F. Benkhelifa, M. Chertouk, M. Dammann, M. Massler, H. Walther and G. Weimann, *International Conference on Semiconductor Manufacturing Technology GaAs MANTECH 2001 Digest*, May 2001).

Figure 2.44, finally, shows the short-circuit current gain h_{21} as well as the maximum available gain (MAG) and the maximum stable gain (MSG) as a function of frequency, on a logarithmic scale. The current gain rolls off with an expected −20 dB/decade, and the transit frequency f_T, measured at $|h_{21}| = 0$ dB, is 110 GHz. The extraction of the claimed f_{max} of 300 GHz is less certain. As is explained in Chapter 5, f_{max} can be

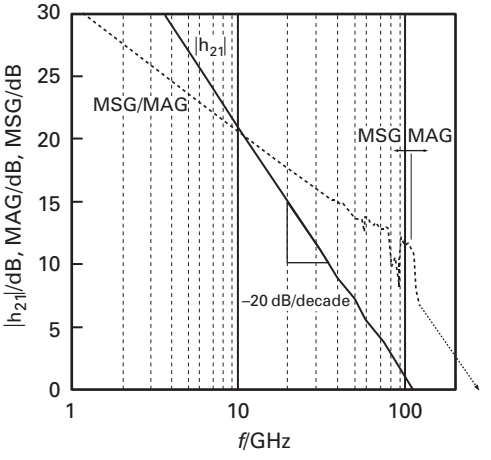

Fig. 2.44 Short-circuit current gain $|h_{21}|$, and power gains MAG and MSG, as a function of frequency, for the metamorphic HEMT structure (F. Benkhelifa, M. Chertouk, M. Dammann, M. Massler, H. Walther and G. Weimann, *International Conference on Semiconductor Manufacturing Technology GaAs MANTECH 2001 Digest*, May 2001).

extracted at the frequency where MAG $= 0$ dB. The problem is that MAG only exists where Rollet's stability factor $k > 1$, otherwise it is replaced by MSG. The change in slope of the MSG/MAG curve suggests that the transition between MSG and MAG happens only above 100 GHz, close to the upper end of the measurement range. From there, $f_{\max}$ seems to be extrapolated also at -20 dB/decade, even though the true roll-off is much steeper. The problem in determining $f_{\max}$ from MAG can be circumvented if an extraction from Mason's unilateral gain u is used. This is also explained elsewhere, and leads to different values for $f_{\max}$.

The important finding, however, is that using a metamorphic HEMT with an *optically defined* gate of $L_G = 0.25\,\mu$m provides sufficient gain for applications at 100 GHz.

2.4 Radio Frequency MOSFETs

2.4.1 Introduction

The silicon MOSFET is by a huge margin the most popular transistor structure. Long confined to either digital circuits or lower-frequency analogue applications, it now makes significant inroads into the realm of micro- and millimetre-wave circuits. With gate lengths below 100 nm, its cutoff frequencies f_T and $f_{\max}$ now rival those of the already introduced HEMTs or advanced HBTs, which will be introduced in the next section of this chapter.

We will briefly review the fundamental aspects of MOSFET operation and then proceed to the analogue aspects of RF CMOS operation which are commonly not covered in texts dealing primarily with MOSFETs as components in digital VLSI and ULSI.

The unparalleled success of silicon as the material of choice in fabricating electronic components is due to several factors:

- Silicon is cheap and has an almost limitless supply.
- Silicon is mechanically robust.
- Silicon has a high thermal conductivity, at least compared to GaAs and InP.
- But most importantly, silicon has a highly stable native oxide, SiO_2, which forms a high-quality interface with silicon.

This latter property led to the unequalled victory of metal-oxide–semiconductor (MOS) technology.

Basic MOSFET structure

Consider a somewhat schematic cross-section of a MOSFET (Figure 2.45).

It is not intended to do justice to the complexity of modern MOSFET devices, but shows their fundamental components. This is an *n-channel* device – the current in the channel will be carried by electrons. We will focus on n-channel devices here as this facilitates comparison with the previously discussed MESFETs and HEMTs (which are almost exclusively n-channel devices), but all findings relate analogously to p-channel devices as well, with appropriate modifications reflecting the differences in doping and free carrier type.

First, we note that the electron channel will actually form in a p-type semiconductor region, called the *bulk*. This can be either the substrate or a p-doped layer formed by epitaxy or diffusion. This will be explained in the next paragraph. Secondly, the stable SiO_2 is used in two different ways:

(i) as a thin *gate oxide* which covers the surface between the source (S) and drain (D) contacts and carries the gate (G) electrode on top; and
(ii) as a thick *field oxide* which covers the remainder of the structure.

The source and drain contacts are non-blocking (ohmic). The gate is physically separated from the semiconductor by the gate oxide – this arrangement is called an *MOS* (*m*etal *o*xide-*s*emiconductor) diode, even though the gate electrode in modern MOSFETs is actually not metallic, but fabricated from highly doped *polycrystalline Silicon* (poly-Si).

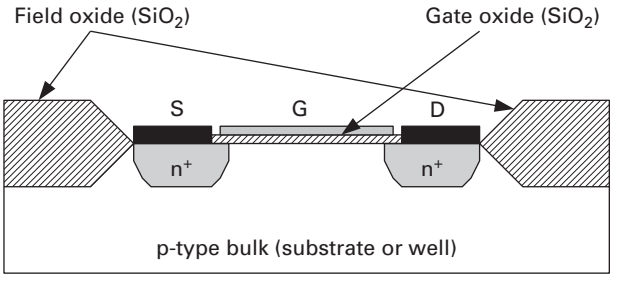

Fig. 2.45 Basic structure of an n-channel MOSFET.

MOS diode operation

Let us now investigate how an electron channel can form in the p-type semiconductor. To this end, we look more closely at the band diagram in a region below the gate electrode. Initially, no external voltages shall be applied to the device.

We construct the band diagram of the MOS diode (Figure 2.46), using *Anderson's rule*.

In the n^+-doped poly-Si gate, we assume that the conduction band energy E_G coincides with the Fermi energy E_F. The bulk Si is p-doped, and we calculate the distance between the Fermi level and the valence band energy E_V, assuming that the Boltzmann approximation to the Fermi–Dirac statistics is valid:

$$E_F - E_V = kT \ln\left(\frac{N_V}{N_A}\right), \tag{2.65}$$

where N_V is the density of states in the valence band and N_A is the acceptor concentration in the p-Si.

The SiO_2 is handled as a semiconductor with a very large band gap.

The distance between the conduction bands and the vacuum level, E_{vac} is given by the electron affinities in the Si and SiO_2 – χ_{Si} and χ_{SiO_2}, respectively. Note $\chi_{Si} > \chi_{SiO_2}$.

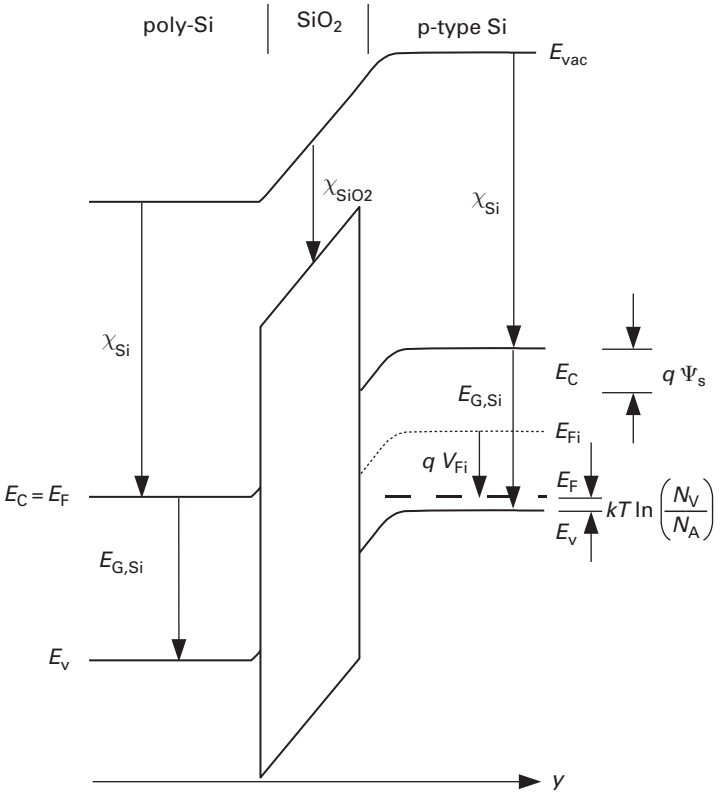

Fig. 2.46 Band diagram of an MOS diode structure.

The continuity of the vacuum level (postulated by Anderson's rule), along with the fact that the poly-Si is n-type, makes the bands in the p-type Si 'dip' towards the SiO_2/Si interface.

We introduced two new potentials and their corresponding energies:

(i) $q\,V_{\mathrm{Fi}}$ is the energy difference between the Fermi levels for the intrinsic and doped semiconductor:

$$q\,V_{\mathrm{Fi}} = \frac{E_{\mathrm{C}} + E_{\mathrm{V}}}{2} + \frac{1}{2}kT\,\ln\frac{N_{\mathrm{V}}}{N_C} - E_{\mathrm{V}} - kT\,\ln\frac{N_{\mathrm{V}}}{N_{\mathrm{A}}} \approx \frac{E_{\mathrm{G}}}{2} - kT\,\ln\frac{N_{\mathrm{V}}}{N_{\mathrm{A}}}. \quad (2.66)$$

(ii) $q\,\Psi_{\mathrm{s}}$ is the energy difference between the undisturbed semiconductor and the Si/SiO_2 interface.

Using these two potentials, we can easily distinguish four different regions:

(i) $\Psi_{\mathrm{s}} < 0$: The bands 'bend upwards' – *accumulation of holes* at the interface – creation of a positive space charge of mobile carriers there.

(ii) $0 < \Psi_{\mathrm{s}} \leq V_{\mathrm{Fi}}$: *Depletion of holes* at the interface – creation of a negative space charge of fixed carriers. The charges are the ionised acceptor atoms. Increase of Ψ_{S} results in an extension of the space charge layer into the semiconductor. In the limit $\Psi_{\mathrm{s}} = V_{\mathrm{Fi}}$, the interface behaves like an intrinsic semiconductor.

(iii) $V_{\mathrm{Fi}} < \Psi_{\mathrm{s}} \leq 2\,V_{\mathrm{Fi}}$: As the Fermi level is now closer to the conduction band than to the valence band at the interface, the conduction type converts from p-type to n-type. This condition is called *light inversion*.

The density of *minority electrons* at the interface increases exponentially with Ψ_{s}:

$$n_{\mathrm{p}} = n_{\mathrm{i}}\,e^{q(\Psi_{\mathrm{s}} - V_{\mathrm{Fi}})/kT}.$$

(iv) At $\Psi_{\mathrm{s}} > 2\,V_{\mathrm{Fi}}$ the interface carrier density rises sharply for small changes in Ψ_{s}, which remains almost constant for large changes in the interface charge. This condition is called *strong inversion*. The *width of the depletion region* stays approximately constant at

$$w_{\max} = 2\sqrt{\frac{\epsilon_{\mathrm{Si}}}{q\,N_{\mathrm{A}}}\,V_{\mathrm{Fi}}}, \quad (2.67)$$

for a homogeneously doped semiconductor.

In Figure 2.46, $0 < \Psi_{\mathrm{s}} < V_{\mathrm{Fi}}$, hence the device is in depletion without externally applied voltages and no channel forms. This is typical of n-channel MOSFET transistors – they have positive threshold voltages, in contrast to MESFETs and HEMTs.

As the gate electrode is isolated from the semiconductor by the gate oxide, we can apply large positive gate-channel voltages without creating a gate current, as would be the case in Schottky diodes. This situation is shown in Figure 2.47.

Note that the conduction band forms a *triangular potential well* at the Si/SiO_2 interface. In this respect, the MOSFET is closely related to the HEMT and will equally form a two-dimensional electron gas in the channel.

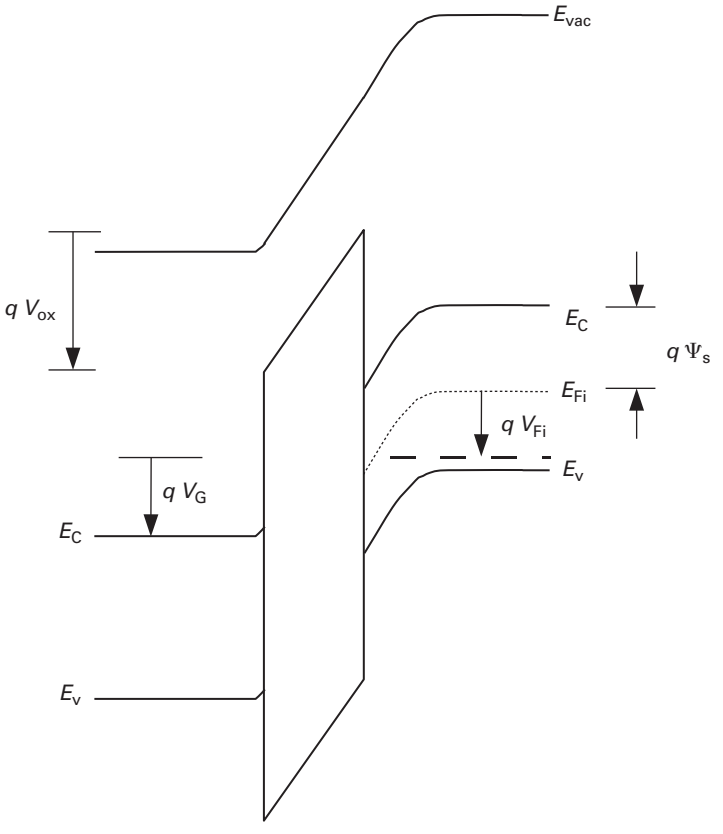

Fig. 2.47 MOS diode band diagram with positive V_G.

The externally applied gate voltage drops partially across the gate oxide, partially across the semiconductor and increases Ψ_s. In the situation drawn in Figure 2.47, the hypothetical intrinsic Fermi level already drops below the Fermi level in the undisturbed p-type semiconductor ($\Psi_s > V_{Fi}$), but $\Psi_s > 2V_{Fi}$ has not been reached – the structure is in light inversion. An even more positive V_G will introduce strong inversion.

If V_{ox} is the voltage drop across the oxide,

$$V_G = V_{ox} + \Psi_s + \frac{1}{q}\left(kT\,\ln\frac{N_V}{N_A} - E_G\right), \tag{2.68}$$

The last term in Equation (2.68) is the *flat-band voltage* V_{FB}:

$$V_{FB} =: \frac{1}{q}\left(kT\,\ln\frac{N_V}{N_A} - E_G\right). \tag{2.69}$$

It is called *flat-band* because for $V_G = V_{FB}$, $\Psi_s + V_{ox} = 0$ and the bands become completely horizontal.

V_{ox} can be calculated from the *gate capacitance* C_{ox} and the *total charge stored under the gate* Q_s:

$$V_{ox} = -\frac{Q_s}{C_{ox}}.$$

The total gate charge is formed by the *space charge in the depleted region* Q_B and the *interface charge* Q_i.

Let us now consider the case where $\Psi_s = 2\,V_{Fi}$ just occurs (onset of strong inversion). Q_i is negligible, and hence $Q_s \approx Q_B$ with

$$Q_B = -q\,N_A\,w_{max}\,L_G\,W_G,$$

where w_{max} is the maximum extension of the space charge region, equal to the extension for $\Psi_s = 2\,V_{Fi}$ (see Equation (2.67)) and $W_G\,L_G$ is the gate footprint, which determines the area of the channel. Hence,

$$Q_B = -2\sqrt{\epsilon_{Si}\,q\,N_A\,V_{Fi}}.$$

We can now calculate the *threshold voltage* V_{th} as the necessary V_G to reach the onset of strong inversion. Recalling that at this point $\Psi_s = 2\,V_{Fi}$, Equation (2.68) yields

$$\begin{aligned} V_{th} &= -\frac{Q_B}{C_{ox}} + 2\,V_{Fi} + V_{FB} \\ &= \frac{2\,W_G\,L_G}{C_{ox}}\sqrt{\epsilon_{Si}\,q\,N_A\,V_{Fi}} + 2\,V_{Fi} + V_{FB}. \end{aligned} \tag{2.70}$$

In strong inversion, the oxide capacitance is easy to calculate, because the mobile charge Q_i is concentrated as a sheet charge at the Si/SiO_2 interface (compare the situation in the HEMT). The sheet charge forms a simple parallel-plate capacitor with the gate electrode:

$$C_{ox} = W_G\,L_G\frac{\epsilon_{SiO_2}}{t_{ox}}, \tag{2.71}$$

where t_{ox} is the gate oxide thickness.

The threshold voltage is then approximately:[5]

$$V_{th} = \frac{2\,t_{ox}}{\epsilon_{SiO_2}}\sqrt{\epsilon_{Si}\,q\,N_A\,V_{Fi}} + 2\,V_{Fi} + V_{FB}. \tag{2.72}$$

2.4.2 Drain current

So far, we considered only the voltage between gate and channel, assuming that the drain and source electrodes are on equal potentials. Now, we apply external voltages $V_{GS}, V_{DS} \neq 0$, as in Figure 2.48. As before in the discussion of MESFET and HEMT, we make certain assumptions for the channel:

- The channel shall be one-dimensional, i.e. the electric field has only a z component.

[5] Because $V_G = V_{th}$, the MOS diode is not strictly in strong inversion.

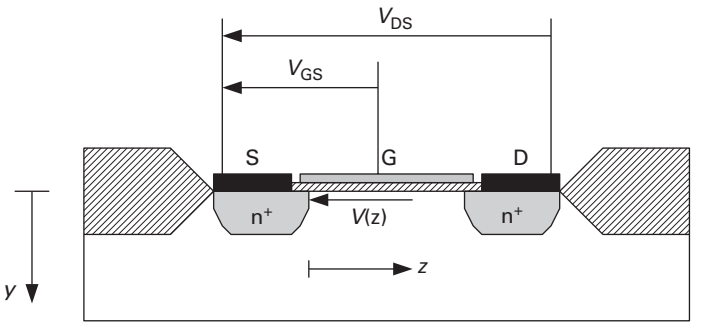

Fig. 2.48 MOSFET structure with externally applied voltages.

- The channel shall be gradual, i.e. the current is driven purely by the electric field and diffusion is neglected.

Now that $V_{DS} > 0$, the voltage between the gate electrode and the semiconductor will depend on the z coordinate along the interface:

$$V_G(z) = V_{GS} - V(z). \tag{2.73}$$

Constant-mobility model

As discussed already, the channel current can be calculated from the local mobile charge and the velocity with which it moves. In case of the MOSFET, the local charge q_i is the mobile interface charge which in strong inversion can be calculated simply from the oxide capacitance and the local gate-channel voltage $V_G(z)$:

$$q_i(z) = \frac{\epsilon_{SiO_2}}{t_{ox}}[V_G(z) - V_{th}]. \tag{2.74}$$

We initially calculate the charge velocity for the low-field case, assuming that $\mu_n =$ const and find

$$I_D(z) = W_G \frac{\epsilon_{SiO_2}}{t_{ox}}[V_G(z) - V_{th}]\,\mu_n' \frac{dV(z)}{dz},$$

where μ_n' is the interface mobility, which is lower than the bulk mobility due to the imperfections of the interface plane.

Applying current continuity, we know that

$$I_D(z) = \text{const} = I_D = \frac{1}{L_G}\int_{z=0}^{z=L_G} I_D(z)dz.$$

Using parameter substitution and noting that $V(z = 0) = 0$, $V(z = L_G) = V_{DS}$, we find

$$I_D(V_{GS}, V_{DS}) = \frac{\epsilon_{SiO_2}}{t_{ox}}\frac{W_G}{L_G}\mu_n'\left[(V_{GS} - V_{th})\,V_{DS} - \frac{V_{DS}^2}{2}\right]. \tag{2.75}$$

The above equation only holds as long as the channel is not fully depleted. Because $V(z)$ increases monotonically with z along the channel, depletion of the channel will

start at the drain, when V_{DS} reaches the knee voltage V_k, in full analogy to the MESFET and HEMT.

$$q_i(z = L_G) = \frac{\epsilon_{SiO_2}}{t_{ox}} (V_{GS} - V_k - V_{th}) = 0$$

yields

$$V_k = V_{GS} - V_{th} \tag{2.76}$$

For $V_{DS} > V_k$, the channel charge no longer depends on V_{DS} in this simple model. With $V_{DS} = V_k$ and using Equation (2.76), we find from Equation (2.75)

$$I_D(V_{GS}) = \frac{\epsilon_{SiO_2}}{t_{ox}} \frac{W_G}{2\,L_G} \mu'_n (V_{GS} - V_{th})^2, \tag{2.77}$$

for $V_{DS} > V_k$.

This simple model of the MOSFET static behaviour is often referred to as the *Shockley model* [57].

Backgating

Another parasitic effect influencing the static performance needs to be considered. Because the MOSFET sits on a conducting silicon layer (the 'bulk', p-type for n-channel, n-type for p-channel transistors), the device is essentially a four-terminal device, where the bulk is the fourth terminal. We had implicitly assumed that the bulk layer would have a fixed potential, which is that of the source contact. In an integrated circuit, however, this cannot always be maintained. Therefore, we need to consider a second control voltage, the bulk-source voltage V_{BS}.

Recall that we defined the threshold voltage via the potential difference between the bulk and the Si/SiO_2 interface. This suggests a very simple way of accommodating backgating – by modifying the threshold voltage:

$$V_{th} = V_{t0} - \gamma\, V_{BS}. \tag{2.78}$$

Here, V_{t0} is the threshold voltage without backgating (i.e. the one considered so far) and γ is a fitting parameter [49].

Non-ideal effects in short-channel MOSFETs

Channel length modulation.

So far, we neglected another effect: the source and drain regions form n–p junctions with the bulk semiconductor. These n–p junctions necessarily create depletion regions, whose width depends on the voltage across the junction. In an n-channel MOSFET, the drain has a positive potential with respect to source. Assuming that the bulk is held on source potential ($V_{BS} = 0$), the drain-bulk region is therefore reverse-biased, which leads to an increase in the width of the space charge region there.

Figure 2.49 shows this situation. The effective gate length L_{eff} is shorter than the 'drawn' gate length L_G. The difference is V_{DS}-dependent due to the drain space charge region:

$$L_{eff} = L_G - \Delta L(V_{DS}). \tag{2.79}$$

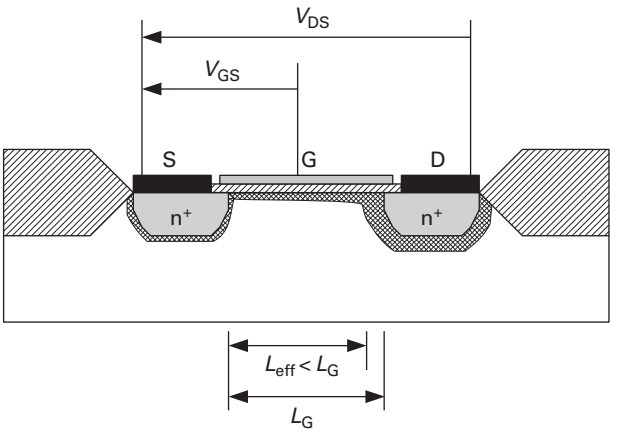

Fig. 2.49 Schematic representation of channel length modulation due to space charge incursion into the channel.

ΔL can be approximated as follows:

$$\Delta L = \frac{1}{2}\sqrt{\frac{2\epsilon}{q\,N_A}(V_{DS} - V_{off})\,\alpha}, \tag{2.80}$$

where α is a factor which depends on the exact geometry of the MOSFET – $\alpha = 0.02\ldots1$. This formula only applies for $V_{DS} > V_{off}$, only then will a part of the channel close to drain be fully depleted.

Recalling Equation (2.77), it is easy to see that the progressive reduction of the effective channel length with increasing V_{DS} will cause the drain current to increase with increasing drain-source voltage. This effect is most pronounced if the 'geometrical' channel length L_G is already small – hence this is a very important effect in high-speed MOSFETs with channel lengths $L_G < 0.5\,\mu m$. This effect is called *channel length modulation* and is conceptually very similar to the *Early effect* which we will introduce for the bipolar transistor (see p. 122).

Short-channel effect.

The calculation of the threshold voltage (Equation (2.72)), assumed that the gate and the mobile sheet charge in the channel form an ideal parallel-plate capacitor: the total interface charge Q_i appears, with opposite sign, at the gate charge Q_G. In reality, some of the field lines emanating from the negative charge in the channel may also terminate on the source and drain areas. Due to the n–p-junctions, the depletion of the channel region progresses more rapidly than predicted from considering the gate potential alone, which leads to a reduction in threshold voltage V_{th}. This is the *short-channel effect* proper, which says that for otherwise unchanged technological parameters, the threshold voltage will decrease with decreasing gate length. The effect is more pronounced, the deeper the source and drain contact regions extend into the bulk material.

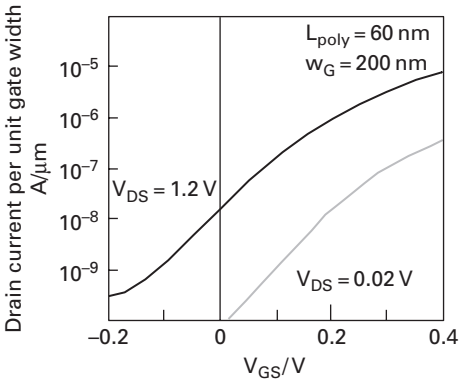

Fig. 2.50 Example for the V_{DS} dependence of subthreshold currents in deep-submicron MOSFETs (T. Sugli, K. Watanabe and S. Sugatani, *Fujitsu Science and Technology Journal*, Vol. 39, No. 6, pp. 9–22, June 2003.).

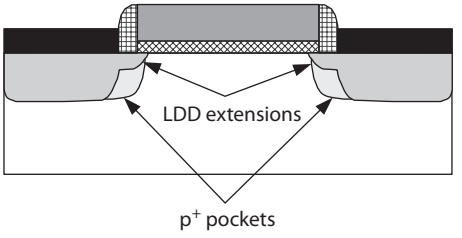

Fig. 2.51 MOSFET structure with a double implant LDD arrangement.

As the shape, specifically of the drain side space charge region, depends on the potential of the drain contact, it is not surprising that V_{DS} also has an effect on V_{th}: as the drain-source voltage is increased, the drain field will deplete the channel more, which leads to a further decrease of the threshold voltage.

The effect of the drain field can best be shown in the *subthreshold regime* (see Figure 2.50). For $V_{GS} < V_{th}$, the channel current does not actually cease to flow, because even before the onset of strong inversion, there are free charge carriers in the channel. Their density and hence the current depend exponentially on $V_{GS} - V_{th}$. The figure shows an example of the subthreshold regime for a deep-submicron MOSFET, for two different values of V_{DS}. We note that even a small increase in V_{DS} increases $I_{off} = I_D(V_{GS} = 0)$ but two orders of magnitude.

A very common modification of the standard MOSFET structure which reduces the short-channel effect and channel length modulation, and also improves the breakdown voltage, is the double implant lightly doped drain (LDD) structure [41] shown in Figure 2.51.

The shallow n-doped drain extensions lower the maximum electric field in the channel and hence increase the breakdown voltage, while the p^+-doped pockets slow the growth of the p–n space charge regions into the channel with increasing V_{DS}.

Mobility degradation.
Short-channel devices benefit from an increase in the bulk doping concentration, because the problem with the high I_{off} and also the channel length modulation can be decreased by increasing the bulk doping. Together with the thin gate oxide, however, this significantly increases the electric field component in y direction. The charge carriers in the channel will then flow, on average, closer to the Si/SiO_2 interface where their mobility is reduced by interface scattering, due to imperfections of the interfacial layer. Therefore, the effective mobility will decrease with increasing V_{GS}, approximated by

$$\mu'_n(V_{GS}) = \frac{\mu'_{n,0}}{1 + m\,(V_{GS} - V_{th})}, \tag{2.81}$$

where $\mu'_{n,0}$ is the mobility at threshold and m is a factor describing the degree of *normal-field mobility degradation*.

In summary, in short-gatelength MOSFETs, the simple one-dimensional approach we started out with is no longer adequate, hence two-dimensional effects have to be incorporated into the physical simulation.

Velocity saturation

As in the typically GaAs-based MESFET and HEMT devices, velocity saturation at high electric fields also has to be considered here. The critical field $\mathcal{E}_{sat}$ for velocity saturation in silicon is $\sim 4 \cdot 10^6\ \mathrm{Vm}^{-1}$ at room temperature, compared to $3 \cdot 10^5\ \mathrm{Vm}^{-1}$ for GaAs, so the onset of velocity saturation is delayed in Si versus GaAs.

If we assume that the channel is fully velocity saturated, i.e. $v_n = v_{sat} \neq f(z)$, the drain current becomes

$$I_D = W_G \frac{\epsilon_{SiO_2}}{t_{ox}} v_{sat}(V_{GS} - V_{th}), \tag{2.82}$$

where v_{sat} is the drift saturation velocity of silicon, which is approximately 10^5 m/s at room temperature.

In the intermediate region, Lee [34] gives the following approximation for the drain current:

$$I_D = W_G \frac{\epsilon_{SiO_2}}{t_{ox}} \frac{v_{sat}}{1 + \frac{L_G \mathcal{E}_{sat}}{V_{GS} - V_{th}}}, \tag{2.83}$$

for $V_{DS} > V_{D,sat}$, where $V_{D,sat}$ is the necessary field for velocity saturation to occur in the channel, approximated as

$$V_{D,sat} \approx \frac{(V_{GS} - V_{th})\, L_G\, \mathcal{E}_{sat}}{(V_{GS} - V_{th}) + L_G\, \mathcal{E}_{sat}}. \tag{2.84}$$

Consider a modern MOSFET with $L_G = 0.09\,\mu$m. $L_G\,\mathcal{E}_{sat}$ is 0.36 V, and assuming $V_{GS} - V_{th} = 0.5$ V, we arrive at $V_{D,sat} = 0.21$ V – considerably smaller than $V_k = V_{GS} - V_{th} = 0.5$ V, as the Shockley model would predict. Velocity saturation is hence a phenomenon with significant importance in deep-submicron MOSFETs.

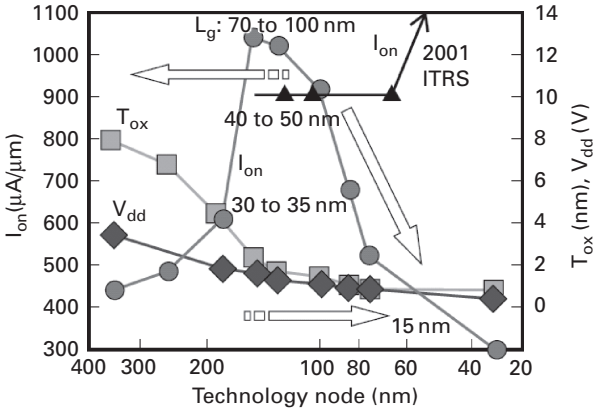

Fig. 2.52 Maximum drain current ('on-current') I_{on}, gate oxide thickness T_{ox} and supply voltage V_{dd} as a function of the technology node (T. Sugli, K. Watanabe and S. Sugatani, *Fujitsu Science and Technology Journal*, Vol. 39, No. 6, pp. 9–22, June 2003.).

Equation (2.82) can also be used to estimate the maximum drain current in a given MOSFET family. In devices with $L_G \leq 130\,\text{nm}$, $t_{ox} \approx 2\,\text{nm}$ typically. For a gate overtravel $V_{GS} - V_{th}$ of 1 V and $V_{sat} = 10^5\,\text{m s}^{-1}$, we find $I_D/W_G = 1.72\,\text{mA}\,\mu\text{m}^{-1}$.

In order to increase the current for a given gate over travel, we have only two options:

(i) Decrease the thickness t_{ox} of the gate oxide. However, as t_{ox} is reduced, the electric field in the dielectric increases, and the gate-channel tunnel current increases dramatically.

(ii) Increase the dielectric constant of the gate dielectric. This can be done by replacing the SiO_2 with a different dielectric, such as HfO_2, which features $\epsilon_{HfO_2} = 25\,\epsilon_0$ instead of $\epsilon_{SiO2} = 3.9\,\epsilon_0$, but with limited thermal stability. Additionally, a major advantage of Si, namely its highly stable native oxide, is given up.

Figure 2.52 presents a literature data review [61] of achieved maximum drain current, gate oxide thickness and power supply voltage as a function of the target technology, as of the year 2003. The decrease in the maximum drain current ('on current', I_{on}) and the supply voltage V_{dd} give proof to the problems CMOS designers face due to the reducing gate oxide thickness.

2.4.3 Large-signal modelling

A MOSFET's non-linear circuit (Figure 2.53), is very similar to what we discussed for the MESFET, or HEMT, except that the substrate ('bulk') node needs to be accounted for. The diodes D_{BS} and D_{BD} represent the source and drain p–n diodes, respectively, and include junction capacitance. The series resistances R_G, R_S and R_D are taken as bias-independent.

As discussed in the previous paragraph, the drain current I_D will be a function of V_{GS} and V_{DS}. In a more precise model, we need to make the threshold voltage a function of V_{DS} and via backgating also of V_{BS}, so the bulk-source voltage needs to be included as

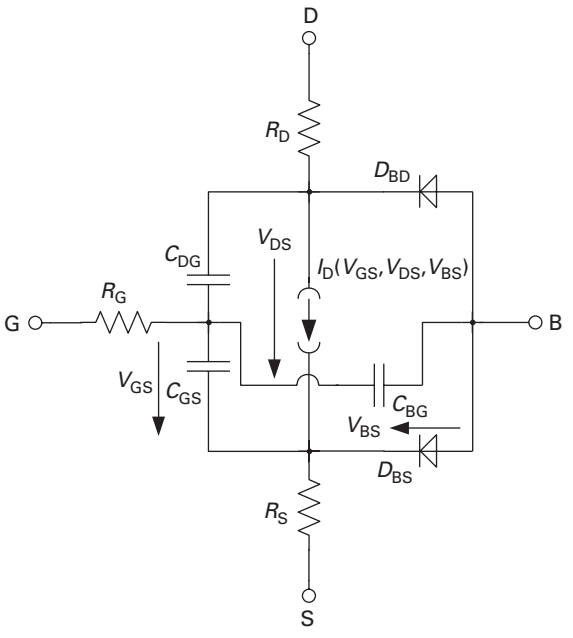

Fig. 2.53 Non-linear equivalent circuit suitable for MOSFETs.

a controlling voltage as well. Please note that the voltages occur between the internal nodes – voltage drops across the series resistances will have to be subtracted.

A particular feature of MOS transistors are the *overlap capacitances*. Referring, for example, to Figure 2.45, note that the gate electrode overlaps the highly doped source and drain regions. Without this overlap, at least on the source side, the channel could not form, as the free charge in the channel is drawn from the source region – unlike in the HEMT, where the free carriers are being introduced through the supply layer on top of the channel. These capacitances will be bias-independent, so that the gate-source and drain-source capacitances can be written as

$$C_{GS} = C_{GSO} + \frac{\delta Q_B}{\delta V_{GS}}$$
$$C_{GD} = C_{GDO} + \frac{\delta Q_B}{\delta V_{GD}}, \tag{2.85}$$

where C_{GSO} and C_{GDO} are the gate-source and gate-drain overlap capacitances and Q_B is the total space charge, which includes the fixed charge ($q\,N_A\,w$) and the mobile interface charge Q_i.

In strong inversion, the change in channel charge is reflected only in the interface charge:

$$\delta Q_B \approx \delta Q_i = W_G \int_{z=0}^{z=L_G} q_i(z) dz.$$

Equation (2.74) indicates that in strong inversion, the total interface charge will only depend on V_{GS}; therefore, C_{GD} will be given only by the overlap capacitance. Non-ideal effects can be included into the capacitance equation by making the threshold voltage V_{th} V_{DS}- and V_{BS}-dependent.

The Meyer capacitance model [38] already introduced for the MESFET is frequently used to model the bias dependence of C_{GS} and C_{GD}. In strong inversion,

- For $V_{DS} < V_k$,

$$C_{GS} = C_{GSO} + \frac{2}{3}C_{GC}\left[1 - \left(\frac{V_k - V_{DS}}{2V_k - V_{DS}}\right)^2\right]$$
$$C_{GD} = C_{GDO} + \frac{2}{3}C_{GC}\left[1 - \left(\frac{V_k}{2V_k - V_{DS}}\right)^2\right] \tag{2.86}$$

- For $V_{DS} > V_k$,

$$C_{GS} = C_{GSO} + \frac{2}{3}C_{GC}$$
$$C_{GD} = C_{GDO}, \tag{2.87}$$

where V_k is the drain-source voltage delineating the linear from the saturated regime and (see Equation (2.76)) C_{GC} is the gate-channel capacitance for $V_{DS} = 0$. In strong inversion, it is simply the total oxide capacitance (see Equation (2.71)).

The gate-bulk capacitance can be similarly expressed; it models the effect of the bulk potential on the channel charge:

$$C_{BG} = \frac{\delta Q_B}{\delta V_{BG}}.$$

In strong inversion, it can be neglected.

The model may be extended with additional elements. Particularly, in RF designs the modelling of the impedance connected to the substrate node deserves particular attention.

At the core of most submicron RF CMOS models is the BSIM3 model, developed at University of California, Berkeley.[6] Unlike e.g. the COBRA model introduced for HEMTs, it uses different sets of equations for the MOSFET's different operating regions. It is also more closely related to device physics, i.e. it is not strictly an empirical model, and its input parameters are partly technological and partly empirical fitting parameters. BSIM's complexity, however, is beyond the scope of a book like this.

A simpler model approach, yet useful for many applications, was published by Sakurai and Newton [49]. The threshold voltage is

$$V_{th} = V_{t0} + \gamma\left(\sqrt{2\Phi_F - V_{BS}} - \sqrt{2\Phi_F}\right), \tag{2.88}$$

where V_{t0}, γ and Φ_F are model parameters and V_{BS} is the bulk-source voltage – the above equation therefore includes backgating.

[6] Web resource at www-device.eecs.berkeley.edu/bsim3/

The drain-source saturation voltage, V_k, is modelled as

$$V_k = K\,(V_{GS} - V_{th})^m, \tag{2.89}$$

where K and m are model parameters. This formulation is more flexible compared to Equation (2.76) and allows to include velocity saturation effects.

The drain current at $V_{DS} = V_k$ is

$$I_{D,sat} = \frac{W_G}{L_{eff}}\,B\,(V_{GS} - V_{th})^n, \tag{2.90}$$

where B and n are model parameters. With variation of n, we can address both the constant-mobility model ($n = 2$) and the constant velocity model ($n = 1$); however, as n is not V_{DS}-dependent, we cannot move from one regime to the other.

A V_{GS}-dependent n, on the other hand, would allow to include the deterioration of mobility at high electric fields normal to the Si/SiO_2 interface (see Equation (2.81)).

The drain current formulation distinguishes between the saturated and non-saturated regions:

- For $V_{DS} > V_k$,

$$I_D = I_{D,sat}\,(1 + \lambda\,V_{DS}), \tag{2.91}$$

  where $\lambda = \lambda_0 - \lambda_1\,V_{BS}$.
- For $V_{DS} < V_k$,

$$I_D = I_{D,sat}\,(1 + \lambda\,V_{DS})\left(2 - \frac{V_{DS}}{V_k}\right)\frac{V_{DS}}{V_k}. \tag{2.92}$$

2.4.4 Small-signal model and RF performance

The small-signal equivalent circuit of the MOSFET is very similar to that used for MESFET or HEMT, but must account for the additional substrate node.

Figure 2.54 shows that a fourth terminal (B) has been added, which is capacitively coupled to the internal gate, source and drain nodes. C_{BG} is shown to facilitate comparison with Figure 2.53; in saturation it is neglected.

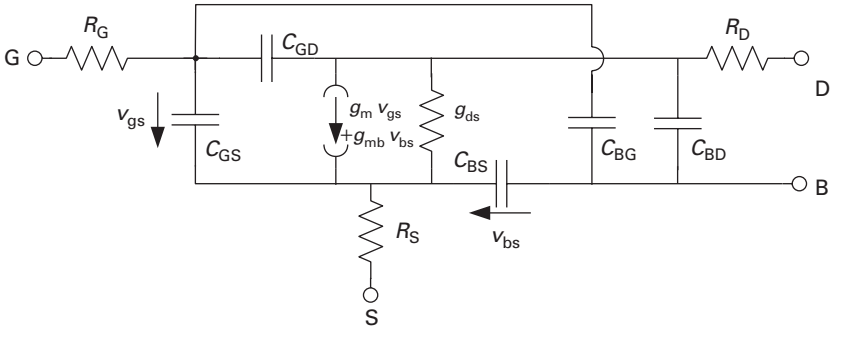

Fig. 2.54 MOSFET small-signal equivalent circuit.

Backgating is included by a special backgating transconductance:

$$g_{\mathrm{mb}} = \frac{\delta I_{\mathrm{D}}}{\delta V_{\mathrm{BS}}}.$$

The transconductance in the constant-mobility limit and for $V_{\mathrm{DS}} > V_{\mathrm{k}}$ is

$$g_{\mathrm{m}} = \frac{dI_{\mathrm{D}}}{dV_{\mathrm{GS}}} = \frac{\epsilon_{\mathrm{SiO_2}}}{t_{\mathrm{ox}}} \frac{W_{\mathrm{G}}\,\mu'_{\mathrm{n}}}{L_{\mathrm{G}}}(V_{\mathrm{GS}} - V_{\mathrm{th}}), \tag{2.93}$$

using Equation (2.77).

In the constant-velocity limit,

$$g_{\mathrm{m}} = \frac{\epsilon_{\mathrm{SiO_2}}}{t_{\mathrm{ox}}} W_{\mathrm{G}}\, v_{\mathrm{sat}}. \tag{2.94}$$

The output conductance is

$$g_{\mathrm{ds}} = \frac{\delta I_{\mathrm{D}}}{\delta V_{\mathrm{DS}}}.$$

In the linear region ($V_{\mathrm{DS}} < V_{\mathrm{k}}$), using the constant-mobility drain current Equation (2.75), we find

$$g_{\mathrm{ds}} = \frac{\epsilon_{\mathrm{SiO_2}}\, W_{\mathrm{G}}\, \mu'_{\mathrm{n}}}{L_{\mathrm{G}}\, t_{\mathrm{ox}}} \left[(V_{\mathrm{GS}} - V_{\mathrm{th}}) - V_{\mathrm{DS}}\right]. \tag{2.95}$$

In the saturated region ($V_{\mathrm{DS}} > V_{\mathrm{k}}$), we use the Sakurai–Newton model Equation (2.91) as our simplified physical regions would predict $g_{\mathrm{ds}} = 0$ there:

$$g_{\mathrm{ds}} = \lambda I_{\mathrm{D,sat}} = \lambda \frac{B\, W_{\mathrm{G}}}{L_{\mathrm{eff}}}(V_{\mathrm{GS}} - V_{\mathrm{th}})^n. \tag{2.96}$$

To reconcile the Sakurai–Newton model with the constant-mobility model, choose $B = (\epsilon_{\mathrm{SiO_2}}\, \mu'_{\mathrm{n}})/t_{\mathrm{ox}}$, $n = 2$, $L_{\mathrm{eff}} = L_{\mathrm{G}}$.

Transit frequency.

We again approximate the transit frequency with

$$f_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{C_{\mathrm{GS}} + C_{\mathrm{GD}}}.$$

In saturation and using the Meyer capacitance equations (2.87),

$$f_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{C_{\mathrm{GSO}} + C_{\mathrm{GDO}} + \frac{2}{3}C_{\mathrm{ox}}}.$$

While recognising the importance of the overlap capacitances, let us assume for simplification that C_{ox} dominates.

For the transconductance, we need to distinguish between the constant-mobility and constant-velocity models. For the constant mobility, using Equation (2.93), we obtain

$$f_{\mathrm{T}} = \frac{3\,\mu'_{\mathrm{n}}}{4\pi\, L_{\mathrm{G}}^2}\,(V_{\mathrm{GS}} - V_{\mathrm{th}}). \tag{2.97}$$

The transit frequency is predicted to increase linearly with the gate overtravel. However, this does not take the mobility degradation with increasing normal field into account (see Equation (2.81)).

In the constant-velocity limit, the transconductance is given by Equation (2.94) and the transit frequency becomes

$$f_T = \frac{3v_{sat}}{4\pi L_G}. \tag{2.98}$$

We already recognised (page 105) that velocity saturation will dominate in short-channel MOSFETs, so we conclude that, like in MESFETs and HEMTs, f_T will scale inversely proportional to the gate length for short L_G.

For a given L_G, the transit frequency may still be increased by improving the mobility, because velocity saturation will be reached sooner and the average velocity in the channel increases.

Both electron and hole mobilities in silicon are enhanced if the silicon layer experiences a tensile strain in the plane parallel to the Si/SiO_2 interface. Semiconductor heterostructures can be used to achieve this: on top of the silicon wafer, first a strain-relaxed $Si_{1-x}Ge_x$ buffer is grown. As was discussed in Section 1.20, the addition of Ge lowers the band gap and at the same time increases the lattice constant. The latter effect is used here – if a thin Si layer is grown on top of the SiGe buffer, it experiences a tensile strain.

Figure 2.55 shows an example of this *strained-layer* technique, here combined with silicon-on-insulator [2]. The use of silicon as the channel layer makes this structure fully compatible with existing gate technology modules.

The axial tensile strain has a significant impact on the electron mobility, as is shown in Figure 2.56, at the expense of extra processing steps, and potential yield limitations when using non-lattice-matched materials.

For p-channel MOSFETs, it is advantageous to place the channel into SiGe layers with significant Ge mole fraction. This will not be discussed here further, however.

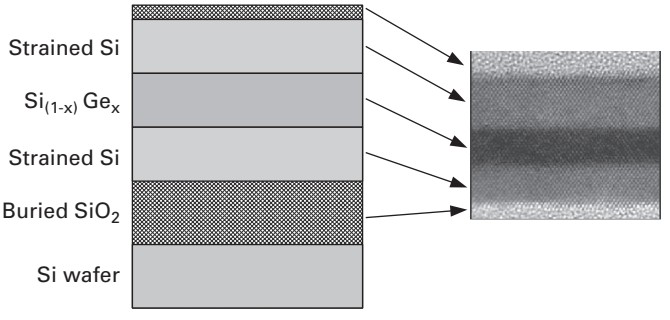

Fig. 2.55 Heterostructure-on-insulator layer stack on a strained-Si MOSFET (left), and corresponding TEM micrograph (TEM micrograph from D. A. Antoniadis, I. Aberg, C. NiCléirigh, O. M. Nayfeh, A. Khakifirooz and J. L. Hoyt, *IBM Journal of Research and Development*, Vol. 50, No. 4/5, pp. 363–377, April–May 2006. ©IBM).

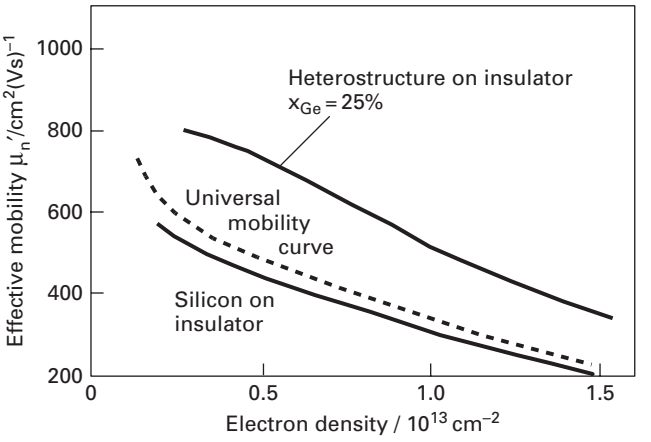

Fig. 2.56 Electron mobility enhancement in strained-Si layer (Data adapted from D. A. Antoniadis, I. Aberg, C. NiCléirigh, O. M. Nayfeh, A. Khakifirooz, J. L. Hoyt, *IBM Journal of Research and Development*, Vol. 50, No. 4/5, pp. 363–377, April–May 2006. ©IBM).

Maximum frequency of oscillation.

As already noted, the maximum frequency of oscillation f_{max} is the more meaningful figure of merit in analogue high-speed applications. As the equivalent circuit for the MOSFET is very similar to that used for MESFET and HEMT, we can easily adapt the f_{max} equation used there:

$$f_{max} = \frac{f_T}{2\sqrt{g_{ds}(R_G + R_S) + 2\pi\ f_T\ R_G\ C_{GDO}}}, \tag{2.99}$$

because in saturation $C_{GD} = C_{GDO}$.

The combination of strained silicon channels and ultrashort gate lengths enables cut-off frequencies above 300 GHz for n-channel CMOS. In a 65 nm technology, a device with $L_G = 29$ nm was reported to have an $f_T = 360$ GHz and an $f_{max} = 420$ GHz [45].

Microwave noise.

The treatment of noise in MOSFETs traditionally neglects the gate and source series resistances and considers only two noise sources [34]:

- The spectral noise current density generated in the channel:

$$\left\langle |i_d|^2 \right\rangle = 8\,kT\ \gamma\ g_{d0}, \tag{2.100}$$

where $g_{d0} = \delta I_D / V_{DS}$ at $V_{DS} = 0$ and γ is a parameter which varies from a value of 1 at $V_{DS} = 0$ to 2/3 at $V_{DS} = V_k$. This model is valid only in the linear region ($V_{DS} \leq V_k$), and was developed for MOSFETs with long gate lengths. In short-channel FETs and in saturation, the observed spectral noise density can be substantially higher. This can be accommodated by making the temperature T larger than the lattice temperature, to account for the significant kinetic energy of the free charge carriers.

- The induced spectral noise current density of the gate current:

$$\left\langle |i_g|^2 \right\rangle = 8\,kT\,\delta\,g_g, \tag{2.101}$$

where

$$g_g = \frac{\omega^2\,C_{GS}^2}{5\,g_{d0}}.$$

In long channel FETs, $\delta = 4/3$.

As discussed for MESFET and HEMT, these two noise sources are partially correlated; in the long gatelength limit, the correlation coefficient is $c = j\,0.395$. From these noise sources, the minimum noise figure can be calculated to be

$$F_{min} = 1 + \frac{2}{\sqrt{5}}\frac{f}{f_T}\sqrt{\gamma\,\delta\,\left(1 - |c|^2\right)}. \tag{2.102}$$

Because f_T in velocity saturation scales $\sim 1/L_G$, we expect the noise figure to vary linearly with the gate length.

An additional noise contribution can come from the substrate. The conducting substrate can be lumped together into a single value, the so-called *spreading resistance*, R_{sub}. This resistance naturally creates thermal noise, with a spectral noise current density:

$$\left\langle |i_{sub}|^2 \right\rangle = \frac{8\,kT}{R_{sub}}. \tag{2.103}$$

This noise current can be capacitively coupled into the transistor via the bulk node. It also leads to a voltage drop across C_{BS}, which will create an additional drain current fluctuation via the backgating effect (see Figure 2.54).

The issue of the source and gate series resistances needs to be re-examined. Modern MOS devices have poly-Si gates. Even highly n-doped poly-Si has specific resistivities which are much higher than for metal films. In RF CMOS technologies, the way around this problem is to connect many very short gate fingers in parallel, e.g. 40 gates of 5 μm gate width each, for a total W_G of 200 μm. As the gate length is more and more decreased, the gate resistance still needs to be recognised with, as Figure 2.57 [48] demonstrates. In this experiment, the noise figure is no longer decreased for $L_G < 0.5\,\mu m$, due to the increase in gate series resistance. Improved gate processes are therefore an important aspect in RF CMOS technology development.

As the gate oxide thickness t_{ox} decreases, the gate current due to Fowler–Nordheim tunnelling increases strongly. It generates a *shot noise contribution* [43], which will have to be accounted for in future MOSFET noise models. If I_G is the gate current, then the spectral noise current density generated is

$$\left\langle |i_g|^2 \right\rangle = 4\,q\,I_G. \tag{2.104}$$

This gate current leads to an additional term in the F_{min} expression [19], compare Equation (2.102):

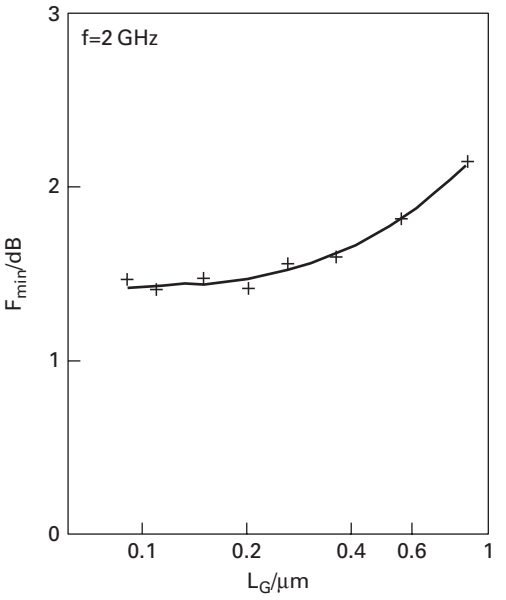

Fig. 2.57 Noise figure versus gate length for n-channel MOSFETs (40 fingers of $W_G = 5\,\mu m$ each) (M. Saito, M. Ono, R. Fujimoto, H. Tanimoto, N. Ito, T. Yoshitomi, T. Ohguro, H. S. Momose and H. Iwai, *IEEE Transactions on Electron Devices*, Vol. ED-45, pp. 737–742, March 1998. ©1998 IEEE).

$$F_{\min} = 1 + \frac{f}{f_T}\sqrt{\frac{\delta\gamma}{5}(1 - c_G^2) + \frac{2q\,I_G g_{d0}\gamma}{16\pi^2\,kT\,f^2\,C_{GS}^2}}. \tag{2.105}$$

For low frequencies, the second term under the root dominates and the minimum noise figure becomes independent of frequency:

$$F_{\min} \approx 1 + \frac{1}{g_m}\sqrt{\frac{2q\,I_G g_{d0}\gamma}{4\,kT}}. \tag{2.106}$$

The appearance of a frequency-independent component in $F_{\min}(f)$ is a tell-tale sign of gate-related shot noise.

A gate current due to Fowler–Nordheim tunnelling is expected to vary with the normal electric field across the gate oxide as

$$I_G \sim \mathcal{E}_{y,SiO_2}^2 \exp\left(-\frac{\phi^{\frac{3}{2}}}{\mathcal{E}_{y,SiO_2}}\right), \tag{2.107}$$

where ϕ is the barrier at the interface. It will therefore be strong function of the gate overtravel $V_{GS} - V_{th}$. The occurrence of gate leakage has thus also important implications on the design of low-noise amplifiers using sub-100 nm CMOS technologies, namely in the choice of the bias point.

2.5 Bipolar and hetero-bipolar transistors

Despite the dominance of MOSFETs in digital circuits, and the significance of HFETs in micro- and millimetre-wave ICs, bipolar transistors have made a strong comeback since the 1990s in high-speed analogue electronics, which is particularly due to the arrival of the Si/SiGe heterostructure bipolar transistor (HBT), but also due to widespread use of GaAs-based HBTs in power amplifiers, e.g. of mobile phone handsets.

One important advantage of bipolar devices is that the current flow is vertical rather than lateral in FETs. This means that the critical geometric dimension (the base layer thickness) is defined by epitaxy or ion implantation. In FETs, the speed-limiting geometry is the gate length which, in present commercially available devices, is defined laterally by lithographic means, at a much higher cost.

We will approach the understanding of HBTs by first considering the standard homojunction homojunction bipolar transistor (BJT), which has long been a corner stone of high-speed electronics, but is gradually being replaced by Si/SiGe HBTs. In particular, we will get a grasp of the shortcomings of the homojunction BJT and how they can be solved by the introduction of bandgap engineering. The HBT is then a straightforward extension of the bipolar transistor concept.

2.5.1 Homojunction bipolar transistors

Homojunction bipolar transistors are the 'classical' bipolar transistors, where all parts of the device are fabricated from the same semiconductor material. Only silicon devices have any market relevance today; for the discussion of high-speed electronics, we can restrict our considerations to n–p–n type devices for reasons which will become clear shortly.

With a suitable permutation in indices, the discussion of n–p–n homojunction bipolar transistors is also valid for p–n–p devices, however.

Figure 2.58 shows the time-honoured[7] one-dimensional representation of a conceptual bipolar transistor of the n–p–n type: the emitter layer is highly donor-doped

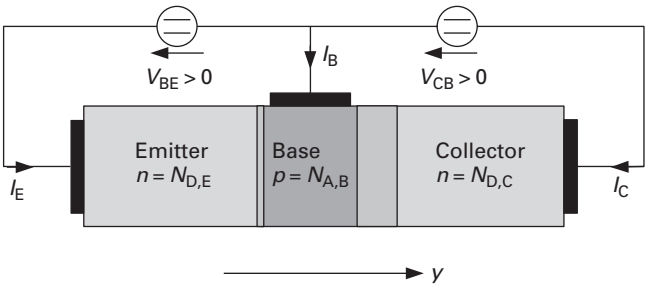

Fig. 2.58 Simplified schematic cross-section of a BJT.

[7] This schematic picture actually dates back to Figure 3 of Shockley's US patent [56].

(n-type), followed by an acceptor-doped (p-type) base layer of medium doping density and a typically lower doped donor-doped (n-type) collector layer.

The emitter, base and collector contacts are non-blocking (ohmic) contacts and are assumed here to be of the recombination type. As shown, two external voltage sources V_{BE} and V_{CB} are connected to the device in such a way that the

- base–emitter p–n junction is forward-biased and
- base–collector p–n junction is reverse-biased.

This mode of operation is called *active forward operation*.

Diffusion triangle

To understand the way in which we control current in the bipolar transistor, let us concentrate initially on the base layer only. The first parameters to introduce are the *diffusion length* of minority charge carriers in the base; in this case the diffusion length of electrons (the base is p-type) L_n, and the thickness of the neutral base layer W_B. W_B is the thickness of the *neutral* base as we have to subtract the space charge regions first. We calculate the diffusion length from the low-field carrier mobility μ, the carrier lifetime τ_r and the absolute temperature T. In the n–p–n transistor, the minority charge in the base are electrons and hence we have to use the electron mobility μ_n and the electron lifetime in the base $\tau_{r,n}$:

$$L_n = \sqrt{\frac{kT}{q}\mu_n \tau_{r,n}}, \tag{2.108}$$

where k is Boltzmann's constant and q the elementary charge.
The term

$$D_n = \frac{kT}{q}\,\mu_n \tag{2.109}$$

is the *diffusion constant for electrons*, hence

$$L_n = \sqrt{D_n \tau_{r,n}}.$$

Equation (2.109) is the Einstein equation introduced earlier, in Equation (1.79).

The emitter–base junction is forward-biased. Therefore, the minority (here, electron) concentration in the base immediately adjacent to the emitter–base space charge region (defined as $y = 0$) is elevated according to

$$n_p(0) = \frac{n_{i,b}^2}{N_{A,B}} e^{qV_{BE}/kT}, \tag{2.110}$$

where $n_{i,b}$ is the intrinsic carrier density in the base.

Provided that the p-layer is infinitely extended in the y direction, the excess minority carrier density decays as

$$n_p(y) = \frac{n_{i,b}^2}{N_{A,B}} + \left[n_p(0) - \frac{n_i^2}{N_{A,B}} \right] e^{-y/L_n}. \tag{2.111}$$

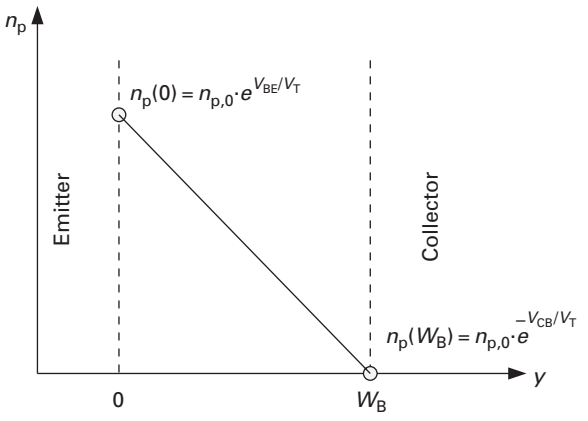

Fig. 2.59 Minority carrier concentration n_p in the base of an n–p–n bipolar transistor, as a function of coordinate y perpendicular to the surface.

In the bipolar transistor, however, the base width W_B is made much shorter than the diffusion length:[8]

$$W_B \ll L_n.$$

It is instructive to estimate a numeric value for L_n. Both μ_n and $\tau_{r,n}$ are strong functions of doping. Let us assume that for a reasonable base doping concentration, $\tau_{r,n} = 1\,\mu s$ and $\mu_n = 500\,cm^2(Vs)^{-1}$. Then the diffusion length amounts to $L_n = 36\,\mu m$, which is certainly much larger than the base width in any microwave bipolar transistor.

The reverse bias across the collector–base junction will cause the minority carrier at the collector side of the neutral base ($y = W_B$) to be significantly smaller than the minority carrier density in the undisturbed semiconductor:

$$n_p(W_B) = \frac{n_{i,b}^2}{N_{A,B}} e^{-qV_{CB}/kT}. \tag{2.112}$$

Provided that $W_B \ll L_n$, which is equivalent to neglecting recombination in the base, the distribution of minority carriers in the base as a function of the coordinate y (which is perpendicular to the surface of the device) is a linear function (see Figure 2.59). Due to its geometric shape, it is sometimes referred to as the *diffusion triangle*.

Collector current equation

In the classic bipolar transistor, the current is carried through the base layer by diffusion only, because the electric field in y direction in the neutral base can be neglected.

We formulate the electron current density flowing through the base layer as a diffusion current:

$$J_{n,B} = q D_n \frac{dn_p}{dy} = q D_n \frac{n_p(0) - n_p(W_B)}{W_B} \approx q D_n \frac{n_p(0)}{W_B}, \tag{2.113}$$

[8] The so-called *short-base diode condition*.

because $n_p(0) \gg n_p(W_B)$. D_n is the diffusion constant of electrons in the base, which has already been defined in Equation (2.109).

Note that due to $dn_p(y)/dy = \text{const}$, the current is constant throughout the base. We obtain the collector current by inserting Equation (2.110) into (2.113) and multiplying with the emitter area A_E:

$$I_C = q A_E D_n \frac{n_{ib}^2}{W_B N_{A,B}} e^{V_{BE}/V_T}, \tag{2.114}$$

still considering $n_p(W_B) \ll n_p(0)$, i.e. under reverse collector–emitter bias. Without this condition, we obtain

$$I_C = q A_E D_n \frac{n_{ib}^2}{W_B N_{A,B}} \left(e^{V_{BE}/V_T} - e^{-V_{CB}/V_T} \right). \tag{2.115}$$

The expression in the denominator of Equation (2.115) $W_B N_{A,B}$ is the *Gummel number of the base layer*, G_B. In the above example and also below, we assume that the base doping concentration is constant across the neutral base. If this is not the case, i.e. $N_{A,B} = f(y)$, we calculate the Gummel number as the integral sheet charge in the base:

$$G_B = \int_0^{W_B} N_{A,B}(y) dy. \tag{2.116}$$

Ideal base current

We will now calculate the ideal base current of a bipolar transistor, i.e. the base current without components due to recombination.

For this, we consider the emitter layer as a short-base diode also: $W_E \ll L_p$, where W_E is the emitter width and L_p is the diffusion constant of minorities (here, holes) in the emitter:

$$L_p = \sqrt{D_p \tau_{r,p}}, \tag{2.117}$$

where $\tau_{r,p}$ is the carrier lifetime of holes in the emitter and

$$D_p = \frac{kT}{q} \mu_p \tag{2.118}$$

is the diffusion constant for holes. μ_p is the hole mobility.

Here, the minority carrier density is assumed to be zero at the emitter contact (ideal recombination contact). The result is a linear dependence of the minority carrier density p_n on y in the emitter (see Figure 2.60).

In analogy to Equation (2.113), we write the hole current in the emitter as a diffusion current:

$$J_{pE} = q D_p \frac{n_{ie}^2}{N_{D,E} W_E} e^{V_{BE}/V_T}, \tag{2.119}$$

where n_{ie} is the intrinsic carrier density in the emitter. Multiplication with the emitter area A_E yields the base current:

$$I_B = q D_p A_E \frac{n_{ie}^2}{N_{D,E} W_E} e^{V_{BE}/V_T}. \tag{2.120}$$

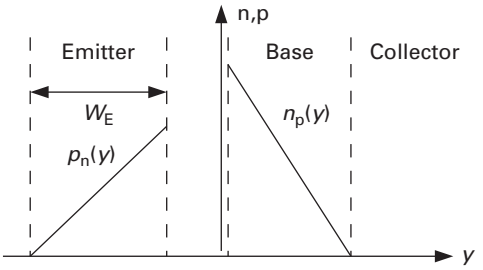

Fig. 2.60 Minority carrier concentration in a bipolar transistor, assuming short-base condition in emitter and base layer.

Ideal current gain

We can now calculate the ideal large-signal forward current gain of the bipolar transistor, dividing Equation (2.115) by (2.120).

$$B_F = \frac{I_C}{I_B} = \frac{D_n W_E}{D_p W_B} \frac{N_{D,E}}{N_{A,B}} \frac{n_{i,b}^2}{n_{i,e}^2}. \tag{2.121}$$

Let us consider the third term in Equation (2.121) first: in a homojunction transistor, where emitter and base are composed of the same material, we can assume $n_{i,b} \approx n_{i,e}$. This is not exactly true because the band gap and hence the intrinsic carrier density also depend weakly on the doping concentration, but it is a useful simplification.

However, we also note that provided we can fabricate the emitter from a different material, it should have a larger band gap (and correspondingly a smaller intrinsic carrier density) such that $n_{i,b} \gg n_{i,e}$. This *wide-gap emitter* is the fundamental idea behind the HBT which we will treat in the next section. It was already included in Shockley's original transistor patent [56] and theoretically expanded upon by Kroemer as early as 1957 [31].

If we consider the second term, we recognise that we cannot arbitrarily increase the base doping concentration unless we also increase the emitter doping concentration, without hurting the current gain. This will lead us to the fundamental limitation of the homojunction bipolar transistor – the inability to lower the base resistance sufficiently for excellent microwave operation.

Non-ideal current contributions

In certain bias conditions, we will have to include additional currents in our considerations. For this, it is instructive to view Figure 2.61.

Especially at low collector currents, *recombination currents* can frequently not be neglected: in bipolar transistors fabricated in direct bandgap semiconductors such as GaAs, the carrier lifetime may be so short that the short-base diode condition ($W_B \ll L_n$ in case of an n–p–n transistor) is never quite fulfilled, so that *volume recombination* in the neutral base has to be accounted for. In other devices, *surface recombination* near the emitter–base p–n junction may play a significant role. These effects are all lumped together in a current contribution J_R, which in an n–p–n transistor exists as an electron

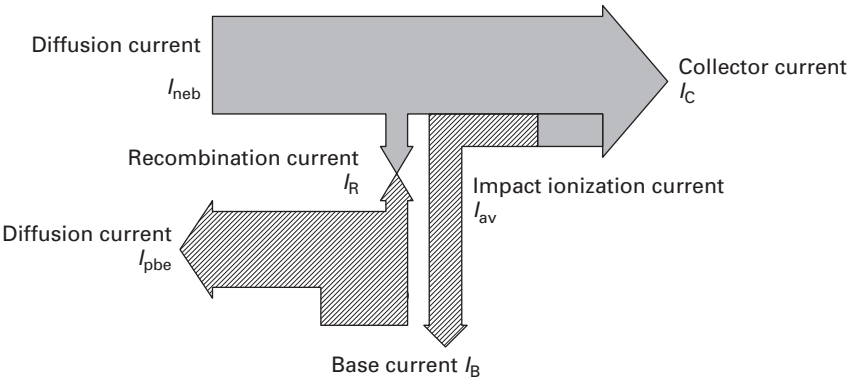

Fig. 2.61 Schematic representation of electron (solid shaded) and hole (hatched) currents in a bipolar transistor.

current (contributing to the emitter current) and a hole current (contributing to the base current).

To account for the recombination current, we add an additional term to the base current Equation (2.120). It has the form of a diode current term with its typical exponential voltage dependence:

$$I_B = \frac{I_C}{B_F} + I_{SR}(\exp^{q \cdot V_{BE}/(N_R \cdot kT)} - 1). \tag{2.122}$$

The emission coefficient N_R is larger than 1. $N_R = 2$ is a typical value for many recombination processes and I_{SR} is the saturation current of the non-ideal base current term.

For high V_{CB}, electrons in the collector space charge region may gain sufficient kinetic energy to elevate a valence electron into the conduction band when colliding with a lattice atom – *impact ionisation* occurs. The charge carriers created by the impact ionisation may again gain sufficient kinetic energy in the strong electric field to cause impact ionisation themselves, leading to a strong increase in current. This is called *avalanching* and is an important breakdown mechanism in bipolar transistors.

Avalanching is accounted for through an additional current term I_{av}, which exists as an electron current adding to the collector current, and a hole current, which *subtracts* from the base current. This is again shown in Figure 2.61.

Non-ideal current gain

The non-ideal current gain can now be written as

$$B = \frac{I_{neb} - I_R + I_{av}}{I_{pbe} + I_R - I_{av}}. \tag{2.123}$$

Even more insight is provided if we express Equation (2.123) in the form of a *common base current gain*.

$$A = \frac{-I_E}{I_C} = \frac{B - 1}{B + 1}$$

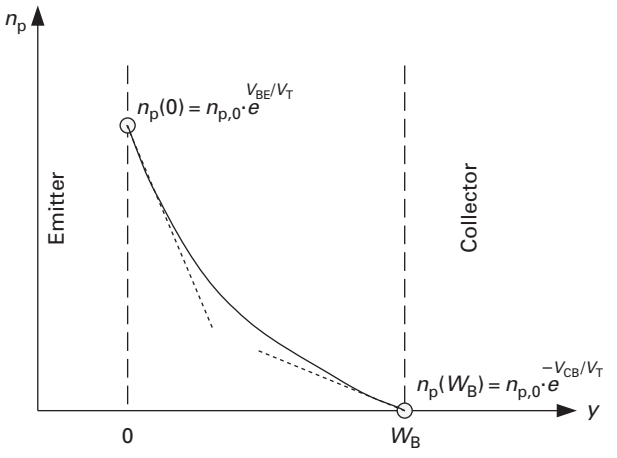

Fig. 2.62 Minority carrier distribution in the base layer in the presence of recombination.

$$A = \frac{I_{neb}}{I_{neb} + I_{pbe}} \cdot \frac{I_{neb} - I_R}{I_{neb}} \cdot \left(1 + \frac{I_{av}}{I_{neb} - I_R}\right). \tag{2.124}$$

In Equation (2.124), the first product term is the *emitter efficiency* γ_E, which describes the electron current from the emitter to the base normalised to the overall current across the emitter–base junction. The second term is the *base transport factor* α_T, which describes the ratio of electron currents across the base–collector and emitter–base junctions. The last term is the *impact ionisation factor* α_M.

Equation (2.124) can hence be rewritten as

$$A = \gamma_E \cdot \alpha_T \cdot \alpha_M.$$

Let us dwell on α_T for a moment. If recombination in the base cannot be neglected, then the minority carrier concentration in the base becomes

$$n_p(y) = \frac{n_{iB}^2}{N_{A,B}} \left[\frac{\sinh\left(\frac{W_B - y}{L_n}\right)}{\sinh\left(\frac{W_B}{L_n}\right)} \left(e^{qV_{BE}/kT} - 1\right) + \frac{\sinh\left(\frac{y}{L_n}\right)}{\sinh\left(\frac{W_B}{L_n}\right)} e^{-qV_{CB}/kT} \right]. \tag{2.125}$$

This is depicted in Figure 2.62.

α_T can be interpreted as the ratio of the minority carrier gradients at $y = W_B$ and $y = 0$:

$$\alpha_T = \frac{\frac{dn_p}{dy}(y = W_B)}{\frac{dn_p}{dy}(y = 0)}. \tag{2.126}$$

In the presence of recombination, the base transport factor is therefore always less than one.

Saturation

In the above discussion, we assumed that the base–collector space charge region was reverse-biased – no carriers were injected into the base from the collector. We will

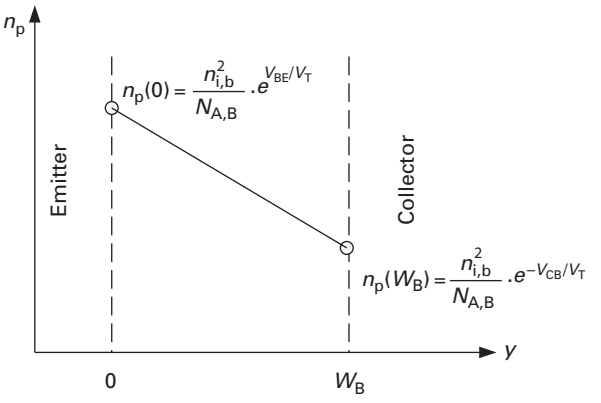

Fig. 2.63 Base minority concentration under saturation conditions.

now drop this condition and allow the base–collector junction to also become forward-biased. In the n–p–n transistor considered here, this means $V_{CB} < 0$. The corresponding bias condition is called *saturation*.

We can conveniently treat this condition again using the diffusion triangle concept introduced in Figure 2.59. Again, recombination across the base is neglected.

Because charge carriers are being injected into the base across the forward-biased collector–base junction, the minority carrier density in the neutral base adjacent to the collector–base space charge region is elevated. For the n–p–n transistor,

$$n_p(W_B) = \frac{n_{i,b}^2}{N_{A,B}} e^{-q \cdot V_{CB}/kT}. \tag{2.127}$$

The resulting diffusion triangle in the base is shown in Figure 2.63.

The ideal collector current under saturation conditions can once again be calculated using a pure diffusion current *ansatz* in the base:

$$I_C = q D_n \frac{dn_p(y)}{dy} = q D_n \frac{n_{i,b}^2}{N_{A,B} W_B} \left(e^{q \cdot V_{BE}/kT} - e^{-q \cdot V_{CB}/kT} \right). \tag{2.128}$$

For inhomogeneous base doping profiles, replace $W_B N_{A,B}$ by the integral Gummel number G_B according to Equation (2.116).

For a fixed base–emitter voltage, the collector current will now strongly decrease with decreasing collector–base voltage, while in the initial discussion of the active forward regime (see Equation 2.115), I_C did not depend on V_{CB}.

In high-speed circuits, the saturation regime has to be carefully avoided due to charge-storage effects whose detailed discussion is beyond the scope of this book.

Early effect

Upon closer examination, the collector current will show a dependence on V_{CB} even in the active forward regime. When deriving Equation (2.115) and the following equations, we had assumed that W_B was constant.

However, consider that W_B is defined as the width of the undepleted (neutral) region bordered by the emitter–base and the base–collector space charge regions. Because the collector current depends exponentially on the base–emitter voltage, $V_{BE} \approx$ const across a wide range of collector currents, and the width of the base–emitter space charge region can be considered constant.

As we change the voltage across the base–collector p–n junction, however, the width of its space charge region will be modulated, which leads to a variation in W_B.

If for simplicity we assume homogeneous doping profiles in base and collector ($N_{A,B} =$ const, $N_{D,C} =$ const), the extension of the base–collector space charge region into the base layer is

$$\delta y_B = \sqrt{2\frac{\epsilon_B}{q} \cdot \frac{N_{D,C}}{N_{A,B}} \cdot \frac{V_D + V_{CB}}{N_{A,B} + N_{D,C}}}, \tag{2.129}$$

where ϵ_B is the dielectric constant of the base layer material and V_D the built-in voltage of the base–collector p–n junction.

The diffusion triangle representation in Figure 2.64 may be helpful again. The dark-shaded areas denote the initial space charge regions. If V_{CB} is increased (with $V_{BE} =$ const), the base–collector space charge region will expand as indicated by the light-shaded areas. Correspondingly, the minority carrier gradient in the neutral base will increase as the neutral base width shrinks.

Hence, the collector current will increase with increasing V_{CB}. This is the *Early effect* [15].

More quantitatively, revisit Equation (2.115), as we consider only $V_{CB} \gg kT/q$. However, now $W_B = f(V_{CB})$:

$$I_C = q A_E D_n \frac{n_{ib}^2}{W_B(V_{CB}) N_{A,B}}. \tag{2.130}$$

We differentiate Equation (2.130) with respect to V_{CB}:

$$\frac{dI_C}{dV_{CB}} = -\frac{I_C}{W_B} \cdot \frac{dW_B}{dV_{CB}}. \tag{2.131}$$

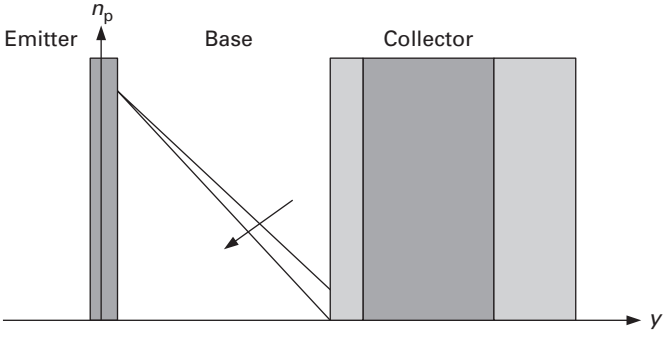

Fig. 2.64 Modification of the base diffusion triangle due to the modulation of the base–collector space charge region.

W_B can be written as $W'_B - \delta y_B$, where W'_B is the layer thickness of the base, i.e. without subtracting the space charge regions.[9] Hence for small changes in the neutral base width ($W'_B \approx W_B$),

$$\frac{dI_C}{dV_{CB}} = -\frac{I_C}{W_B} \cdot \frac{dW_B}{dV_{CB}} = \frac{I_C}{W_B}\frac{d\delta y_B}{dV_{CB}} = \frac{I_C}{N_{A,B}W_B}\sqrt{\frac{\epsilon_B}{2q}\frac{N_{D,C}N_{A,B}}{N_{D,C}+N_{A,B}}\frac{1}{V_D+V_{CB}}}. \tag{2.132}$$

With the Gummel number for homogeneous doping concentration in the base,

$$G_B = N_{A,B}W_B$$

and the base–collector capacitance per unit area

$$C'_{j,BC} = \sqrt{\frac{\epsilon_B q}{2}\frac{N_{A,B}N_{D,C}}{N_{A,B}+N_{D,C}}\frac{1}{V_D+V_{CB}}}$$

it follows that

$$\frac{dI_C}{dV_{CB}} = \frac{I_C}{V_A}, \tag{2.133}$$

where V_A is the Early voltage:

$$V_A = \frac{q \cdot G_B}{C'_{jCB}}. \tag{2.134}$$

The Early voltage is therefore directly proportional to the base Gummel number.

In microwave electronics, the Early voltage is an important factor because it affects the linearity of power amplifiers. For highly linear power amplifiers, a high Early voltage is desired, as it reduces the dependence of the collector current on the collector–emitter voltage. Due to the exponential dependence of the collector current on the base–emitter voltage, V_{BE} is approximately constant even for large variations of I_C, so that $\delta V_{CE} \approx \delta V_{CB}$. A high Early voltage V_A hence reduces the dependence of the collector current on the collector–emitter voltage.

Kirk effect

The last intrinsic effect to be discussed here is the Kirk effect. It can be once again explained using the diffusion triangle (see Figure 2.65).

Despite the fact that the minority carriers traversing the base are being injected into the base–collector space charge region, we had so far assumed that the space charge itself remains unaffected. This is true as long as the density of mobile charge is much smaller than the density of fixed charge.

If we increase the collector current sufficiently to create a density of mobile charge comparable to the fixed space charge density, the mobile charge will start to compensate the space charge and the space charge region will shrink.

[9] Well, actually we have to subtract the extension of the emitter–base space charge region into the base, but that can be neglected here.

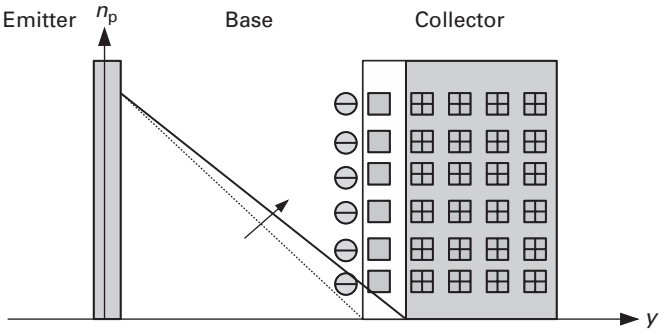

Fig. 2.65 Schematic representation of the Kirk effect influence on the diffusion triangle.

If the width of the space charge region decreases, the neutral base zone will expand – this is called *base push out*. Correspondingly, the minority carrier gradient and with it the collector current for a given V_{BE} will decrease.

To get an estimate of the collector current necessary to cause the Kirk effect, let us assume that the free charge (electrons in our case) will be accelerated to their drift saturation velocity v_{sat} immediately after they enter the base–collector space charge region. Then the electron density n_C corresponding to a collector current density J_C is

$$n_C = \frac{J_C}{q \cdot v_{sat}}.$$

If we now require that the fixed ionised donors in the collector of density $N_{D,C}$ shall be fully compensated by the mobile charge ($n_C = N_{D,C}$), we find for the critical current for the onset of the Kirk effect:

$$J_{C,Kirk} = q \cdot N_{DC} \cdot v_{sat}. \tag{2.135}$$

The onset of Kirk effect hence scales proportionally with the collector doping concentration.

2.5.2 Small-signal dynamic behaviour

Next, we will discuss the dynamic behaviour of the bipolar transistor for the small-signal case, where the non-linear relationships between the terminal currents and the voltages can be described as linear relationships between the deviations of these entities from a given bias point, i.e.

$$i_c = \delta I_C, i_b = \delta I_B, v_{be} = \delta V_{be}, v_{ce} = \delta V_{CE}.$$

The forward-biased emitter–base junction will now be described by a conductance:

$$\frac{d(I_C + I_B)}{dV_{BE}} = -\frac{dI_E}{dV_{BE}} = g_e, \tag{2.136}$$

where g_e will be referred to as the *dynamic emitter conductance*. The minus sign is due to the convention that all currents (I_C, I_B, I_E) are counted positive flowing into the transistor.

From Equations (2.115), (2.120) and (2.121), we conclude that

$$I_E = -(I_C + I_B) = -qA_E D_n \frac{n_{ib}^2}{G_B} e^{V_{BE}/V_T} \left(1 + \frac{1}{BF}\right) \tag{2.137}$$

in the active forward regime, neglecting any non-ideal current contributions.

Differentiating Equation (2.137) with respect to V_{BE}, we find the dynamic emitter conductance to be

$$g_e = -\frac{I_E}{V_T}. \tag{2.138}$$

It can be written as the dynamic emitter resistance:

$$r_e = -\frac{V_T}{I_E}. \tag{2.139}$$

The modulation of stored charge in the neutral base results in the *diffusion capacitance*. In the short-base diode limit, for a homogeneously doped base layer, and if we neglect carrier injection from the collector into the base (i.e. for V_{CB} sufficiently high), the stored minority charge in the base can be easily calculated, refer again to Figure 2.59:

$$Q_B = qA_E \frac{W_B}{2} \frac{n_{ib}^2}{N_{A,B}} e^{V_{BE}/V_T}. \tag{2.140}$$

Differentiating Q_B with respect to V_{BE} results in the diffusion capacitance C_d:

$$C_d = \frac{dQ_B}{dV_{BE}} = qA_E \frac{W_B}{2} \frac{n_{ib}^2}{N_{A,B}} \frac{1}{V_T} e^{V_{BE}/V_T}. \tag{2.141}$$

Considering Equation (2.137) and recalling that in the homogeneously doped case, the base Gummel number is $G_B = N_{A,B} W_B$, we find that Equation (2.141) can be written as

$$C_d = \frac{I_E}{V_T} \frac{W_B^2}{2D_n} = g_e \frac{W_B^2}{2D_n}. \tag{2.142}$$

The ratio of diffusion capacitance to dynamic emitter conductance is bias-independent (at least in the active forward regime) and is called the *base transit time* τ_B:

$$\tau_B = \frac{C_d}{g_e} = \frac{W_B^2}{2D_n}. \tag{2.143}$$

Recall that $D_n = (kT/q)\mu_n$ Equation (2.109); therefore, the base transit time is inversely proportional to the minority carrier mobility in the base. As the electron mobility μ_n is significantly larger than the hole mobility μ_p, in Si as well as the most common compound semiconductors, this justifies the restriction of our discussions to n–p–n-type transistors.

The *output conductance* is related to the Early effect. We find it by differentiating I_C with respect to V_{CE}:

$$g_{ce} = \frac{dI_C}{dV_{CE}} = \frac{dI_C}{dV_{CB}}, \tag{2.144}$$

for constant V_{BE}. Therefore,

$$g_{ce} = \frac{I_C}{V_A}. \tag{2.145}$$

The emitter–base and base–collector p–n junctions also present *junction capacitances* which we have to take into account. The calculation of the junction capacitance can be reduced to the problem of calculating the width of the depletion layer w, because the capacitance is that of a parallel plate capacitor with area A_J and plate separation w:

$$C_J = \epsilon_S \frac{A_J}{w}. \tag{2.146}$$

The depletion layer width is calculated from Poisson's equation:

$$\frac{d^2\Phi}{dy^2} = -\frac{\rho(y)}{y}, \tag{2.147}$$

where Φ is the potential across the junction. A simple analytic solution can be found assuming that within the space charge region the mobile charge can be neglected compared to the fixed charge (i.e. the ionised donor density $N_D^+(y)$ on the n side and the ionised acceptor density $N_A^-(y)$ on the p side), and that within the depletion layer all doping atoms are ionised, while outside all doping atoms are neutral. The total potential difference across the depletion region must be equal to the sum of built-in (or diffusion) voltage V_D and the externally applied voltage V_{ext}.

For the simple case of homogeneous doping on both p and n sides ($N_D(y) = \text{const}$, $N_A(y) = \text{const}$), we obtain

$$w = \sqrt{2\frac{\epsilon_S}{q}\frac{N_A + N_D}{N_A N_D}(V_D + V_{ext})}. \tag{2.148}$$

Note that V_{ext} is defined as a reverse (depleting) voltage here.

In active forward operation, the emitter–base diode is forward-biased ($V_{ext} = -V_{BE}$), while the base–collector diode is reverse-biased ($V_{ext} = V_{CB}$). In practical transistors, the emitter–base junction area A_E will also be different from the base–collector junction area A_C, so that we obtain for the n–p–n transistor:

$$C_{BE} = A_E\sqrt{\frac{q \cdot \epsilon_S}{2}\frac{N_{A,B}N_{D,B}}{N_{A,B} + N_{D,E}}\frac{1}{V_D - V_{BE}}} \quad \text{for } V_{BE} < V_D \tag{2.149}$$

$$C_{CB} = A_C\sqrt{\frac{q \cdot \epsilon_S}{2}\frac{N_{A,B}N_{D,C}}{N_{A,B} + N_{D,C}}\frac{1}{V_D + V_{CB}}}. \tag{2.150}$$

Finally, we have to account for the time lag associated with the transit of free charge carriers through the base–collector space charge region – the *collector transit time*.[10]

The calculation of the collector transit time is not as straightforward as it may seem, as it in principle needs both diffusive (at the base side edge of the space charge region)

[10] Because the emitter–base diode is forward-biased, its depletion region is very thin and the associated transit time can be safely neglected.

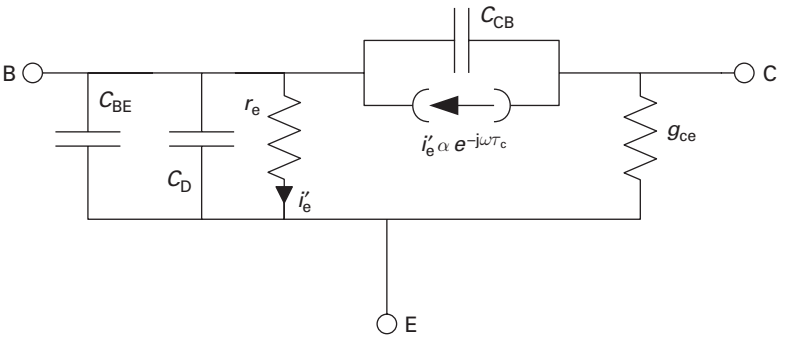

Fig. 2.66 Small-signal equivalent circuit of the intrinsic transistor.

and drift transport components. The field-dependent velocity also needs to be taken into account, as well as the displacement current created by charge moving within the space charge region.

A common assumption is that the carriers reach their drift saturation v_{sat} instantaneously after entering the space charge region. Then, the collector transit time τ_C is

$$\tau_C = \frac{w_C}{2 \cdot v_{sat}}. \tag{2.151}$$

The small-signal components of the intrinsic transistors, which we considered above, can be combined in the *small-signal equivalent circuit* shown in Figure 2.66. α is the small-signal common base current gain in the quasi-static limit.

Figure 2.66 applies only to the intrinsic transistor and will have to be extended by extrinsic parasitic components at microwave frequencies. Most importantly, we have to account for the series resistances. In order to appreciate the problem, please refer to Figure 2.67, which presents a more realistic cross-section of the bipolar transistor, compared to Figure 2.58.

The *emitter resistance* R_E (not to be confused with the dynamic emitter resistance r_e) is composed of the emitter contact resistance and the vertical resistance of the emitter layer, which in homojunction transistors typically is a poly-Si plug. The *collector resistance* R_C is formed by the collector contact resistance, the vertical resistance of the collector 'sinker' implant and the lateral resistance of the subcollector layer. The *base resistance* R_B finally is formed by the base contact resistance, the lateral resistance of the extrinsic base and the lateral resistance of the intrinsic base layer. Of these individual series resistance contributions, the lateral resistance of the intrinsic base layer is the most problematic, as the thickness of this layer is given by the neutral base width W_B and has to be very thin to minimise the base transit time, (see Equation (2.143)).

The series resistances have been added to the small-signal equivalent circuit in Figure 2.68. From an application point of view, R_B and R_E are the most significant.

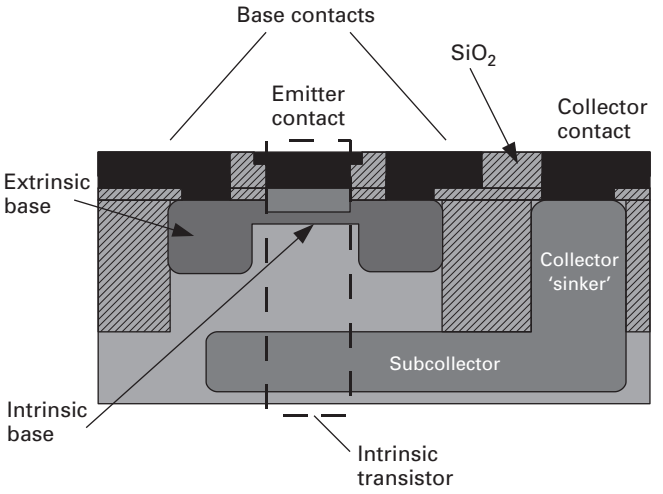

Fig. 2.67 A more realistic schematic cross-section of a typical bipolar transistor with planar contact arrangements. The dashed box indicates the intrinsic transistor – compare with Figure 2.58.

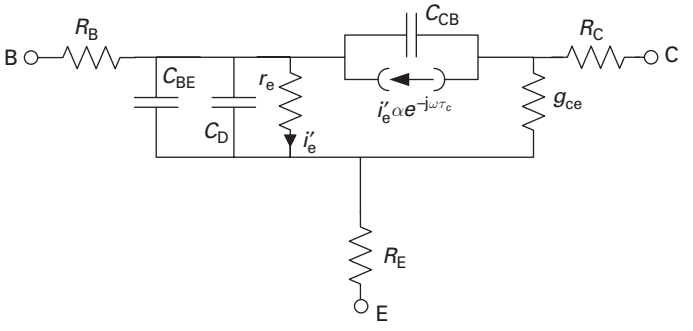

Fig. 2.68 Small-signal equivalent circuit of the bipolar transistor, including the series resistances.

Transit frequency.

The total transit time through the bipolar transistor is calculated from the *transit frequency* f_T, which is the frequency where the magnitude of the short-circuit current gain ($h_{21} = i_c/i_b$ for $v_{ce} = 0$) becomes one: $\tau_T = 1/(2\pi f_T)$.

$$\tau_T = \tau_B + \tau_C + r_e\,(C_{BE} + C_{BC}) + (R_E + R_C)\,(C_{BE} + C_{BC})\,. \tag{2.152}$$

The total transit time can be separated as follows:

- τ_B and τ_C are intrinsic time constants which do not depend on the emitter current (neglecting the Kirk effect).
- $r_e\,(C_{BE} + C_{BC})$ is the intrinsic emitter charging time which is inversely proportional to the emitter current (see Equation (2.139)).
- $(R_E + R_C)\,(C_{BE} + C_{BC})$ is the parasitic charging time due to the emitter and collector series resistances, which is frequently neglected.

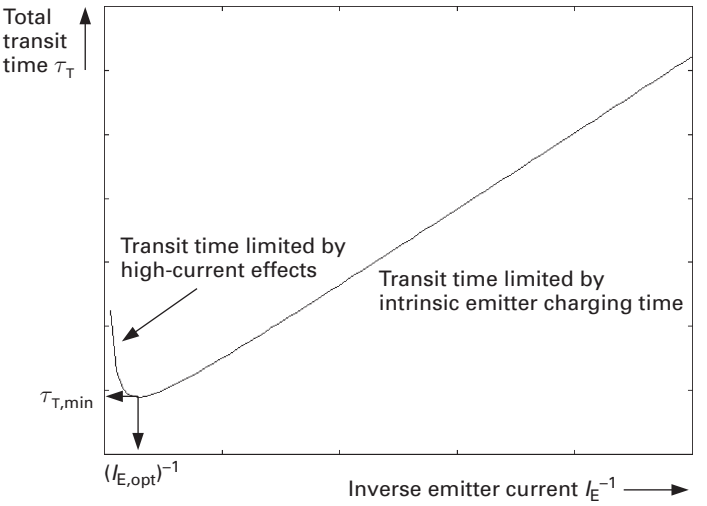

Fig. 2.69 Schematic representation of the total transit time in a bipolar transistor as a function of the inverse emitter current.

The effect of the intrinsic emitter charging time, which depends linearly on the emitter current (or, as $\alpha \approx 1$ in technical transistors, in good approximation on the collector current), leads to a strong bias dependence of the total transit time (and hence f_T), which is shown in Figure 2.69. For $I_E > I_{E,opt}$, high-current effects such as the Kirk effect will again prolong the transit time.

One of the key issues in designing high-speed bipolar circuits is therefore to choose the emitter current as close as possible to the optimum emitter current.

Note that neither the base resistance nor the output conductance have an influence on the total transit time – this is an effect of the definition via f_T and hence h_{21}. The definition of h_{21} assumes an ideal current source at the input and a short circuit at the output. g_{ce} has no effect as it is short-circuited (neglecting R_C here), and R_B is in series with an ideal current source and hence also has no effect.

Maximum frequency of oscillation.

As already discussed, the maximum frequency of oscillation f_{max} is a measure of the power gain cutoff frequency (f_T measures only the current gain behaviour): the frequency where the MAG of a two-port becomes one.

A common approximation of f_{max} for the bipolar transistor is

$$f_{max} = \sqrt{\frac{f_T}{8\pi R_B C_{BC}}}. \tag{2.153}$$

This equation is equivalent to the one introduced for FETs; see Equation (2.27) for very low output conductances and replacing $R_G \to R_B$, $C_{GD} \to C_{BC}$. For an in-depth treatment, see M. B. Das [11]. As explained there, the simplification neglects the distributed nature of the base resistance (as we did in this introductory text) and is only valid if the emitter series resistance R_E and output conductance g_{ce} are sufficiently small.

In practice, however, Equation (2.153) is useful even for today's HBTs with several hundred GHz f_{max}.

Note that now R_B has a strong influence on the maximum frequency of oscillation.

2.5.3 Microwave noise performance of bipolar transistors

We will investigate the microwave noise performance of bipolar transistors using a simplified noise equivalent circuit (see Figure 2.70).

The series resistance R_E and R_C will be neglected, as will be the output conductance g_{ce} and the base–collector capacitance C_{BC}. This leaves three different noise sources to be included:

(i) the thermal noise associated with the base resistance R_B: $\langle |v_{nb}|^2 \rangle$;
(ii) the shot noise associated with the emitter–base p–n junction: $\langle |v_{ne}|^2 \rangle$;
(iii) the shot noise associated with the base–collector p–n junction: $\langle |i_{nc}|^2 \rangle$.

Due to the short base transit time, the emitter–base and base–collector shot noise sources are strongly correlated.

The rationale for the omission of the collector resistance is that its contribution would be divided by the gain of the transistor; further the value of R_C is typically much smaller than R_B. The thermal noise source of the emitter resistance with a squared spectral voltage density of $8\,kT\,R_E$[11] would be in series with the shot noise source of the emitter current, whose spectral voltage density is

$$\left\langle |v_{ne}|^2 \right\rangle = 4\,q\,I_E\,r_e^2 = 4\,kT\,r_e, \tag{2.154}$$

as $r_e = kT/(q\,I_E)$. As long as $r_e \gg 2R_E$, the thermal noise contribution of the emitter resistance can be neglected. Because low-noise bias points for bipolar transistors occur at small I_E, this can generally be assumed.

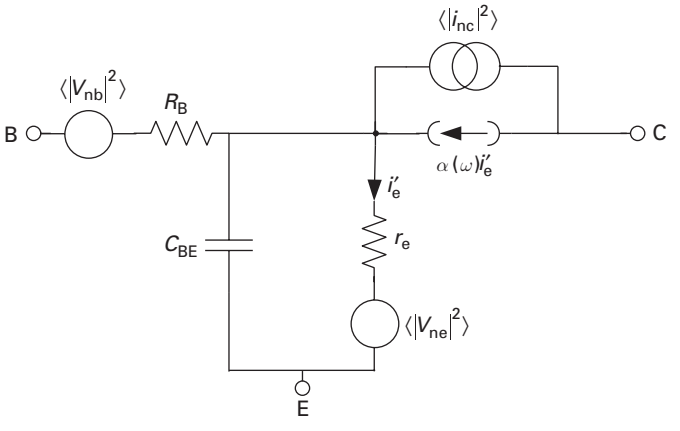

Fig. 2.70 Strongly simplified T-type equivalent noise circuit of a bipolar transistor.

[11] Magnitude of the complex phasor.

Using the equivalent circuit in Figure 2.70, Hawkins [23] derived for the bipolar transistor:

$$F_{\min} = a\,\frac{R_{\mathrm{B}} + R_{\mathrm{opt}}}{r_{\mathrm{e}}} + \frac{\alpha_0}{|\alpha(\omega)|^2}, \tag{2.155}$$

where $\alpha(\omega)$ is the frequency-dependent common base current gain and α_0 its quasi-stationary value:

$$\alpha(\omega) = \frac{\alpha_0}{1 + j\omega/\omega_{\mathrm{b}}}.$$

with the base cutoff frequency $\omega_{\mathrm{b}} = \tau_{\mathrm{B}}^{-1}$ and τ_{B} the base transit time Equation (2.143).

The parameter a is

$$a = \frac{1}{\alpha_0}\left[1 + \left(\frac{\omega}{\omega_{\mathrm{e}}}\right)^2\right]\left[1 + \left(\frac{\omega}{\omega_{\mathrm{b}}}\right)^2\right] - 1.$$

The cutoff frequency ω_{e} represents the emitter charging time:

$$\omega_{\mathrm{e}} = \frac{1}{C_{\mathrm{BE}} r_{\mathrm{e}}} = \frac{q\,I_{\mathrm{E}}}{kT\,C_{\mathrm{BE}}}.$$

R_{opt} is the real part of the noise-optimum generator impedance:

$$R_{\mathrm{opt}} = \sqrt{R_{\mathrm{B}}^2 - X_{\mathrm{opt}}^2 + \frac{\alpha_0}{|\alpha(\omega)|^2}\,\frac{r_{\mathrm{e}}(2R_{\mathrm{B}} + r_{\mathrm{e}})}{a}},$$

and X_{opt} the imaginary part:

$$X_{\mathrm{opt}} = \omega \frac{\alpha_0}{|\alpha(\omega)|^2}\,\frac{C_{\mathrm{BE}}\,r_{\mathrm{e}}^2}{a}.$$

It is instructive to consider the quasi-static case, $\omega \to 0$. In this case,

$$\begin{aligned} a &= \frac{1 - \alpha_0}{\alpha_0} \\ X_{\mathrm{opt}} &= 0 \\ R_{\mathrm{opt}} &= \sqrt{R_{\mathrm{B}}^2 + \frac{r_{\mathrm{e}}\,(2R_{\mathrm{B}} + r_{\mathrm{e}})}{1 - \alpha_0}}. \end{aligned}$$

We obtain therefore

$$F_{\min}(\omega \to 0) = \frac{1}{\alpha_0} + \frac{R_{\mathrm{B}}}{\beta_0 r_{\mathrm{e}}} + \sqrt{\frac{R_{\mathrm{B}}^2}{(\beta_0\,r_{\mathrm{e}})^2} + (1 - \alpha_0)\frac{2\,R_{\mathrm{B}} + r_{\mathrm{e}}}{r_{\mathrm{e}}}}, \tag{2.156}$$

where $\beta_0 = \alpha_0/(1 - \alpha_0)$ is the common-emitter small-signal current gain.

We conclude that for $\omega \to 0$, the minimum noise figure does not converge towards 1, as in FETs, (see e.g. Equation (2.29) for the MESFET), but a higher value which

depends on the current gain and the base resistance. If we further assume small $R_B/(\beta_0 r_e)$, Equation (2.156) reduces to

$$F_{min}(\omega \to 0) \approx \frac{1}{\alpha_0} + \sqrt{(1-\alpha_0)\frac{2R_B + r_e}{r_e}}.$$

It is obvious that the current gain has a very important influence on the noise performance of a bipolar transistor.

Let us now investigate a medium frequency range $\omega_e \ll \omega \ll \omega_b$. To simplify matters, we assume an ideal current gain $\alpha_0 = 1$. In this case,

$$a = \left(\frac{\omega}{\omega_e}\right)^2$$
$$X_{opt} = \frac{\omega_e}{\omega} r_e$$
$$R_{opt} = \sqrt{R_B^2 + 2\, R_B\, r_e \left(\frac{\omega_e}{\omega}\right)^2},$$

and finally for $F_{min}(\alpha_0 = 1, \omega_e \ll \omega \ll \omega_b)$:

$$F_{min} = 1 + \frac{\omega^2}{\omega_e^2}\frac{R_B}{r_e} + \frac{\omega}{\omega_e}\sqrt{\frac{R_B^2}{r_e^2}\frac{\omega^2}{\omega_e^2} + 2\frac{R_B}{r_e}}. \tag{2.157}$$

We find that in this case the increase with frequency is determined by the base resistance R_B, which is therefore a very important parameter for the microwave noise behaviour of bipolar transistors.

Equation (2.155) contains an implicit bias dependence via $r_e = V_T/I_E$, where $V_T = kT/q$, as usual. r_e also determines f_e. As long as $I_E \gg V_T\, \omega\, C_{BE}$, F_{min} will increase proportionally with increasing I_E. For very small I_E, however F_{min} will increase inversely proportional to I_E. We note that here will be an optimum emitter current with respect to the noise performance. This current is usually much lower than the current required for optimum f_T (Figure 2.69) and hence requires a trade-off between device speed and noise in circuit design. Figure 2.71 shows an example calculation. We note that the optimum emitter current is frequency-dependent and moves to higher currents with increasing frequency.

An important noise parameter not considered in Hawkins' theory is the equivalent noise resistance. It determines the sensitivity of the noise figure on deviations from the noise-optimum generator impedance. Therefore, a small R_n facilitates circuit design as it makes exact noise match less critical (Section 5.3). Using Hawkins' equivalent circuit, an expression for R_n was introduced by Pucel and Rohde:

$$R_n = R_B\left(D - \frac{1}{\beta_0}\right) + \frac{r_e}{2}\left\{D + \left(\frac{R_B}{r_e}\right)^2 \cdot \left[1 - \alpha_0 + \left(\frac{f}{f_b}\right)^2 + \left(\frac{f}{f_e}\right)^2 + \left(\frac{1}{\beta_0} - \left(\frac{f}{f_b}\right)\left(\frac{f}{f_b}\right)\right)^2\right]\right\}. \tag{2.158}$$

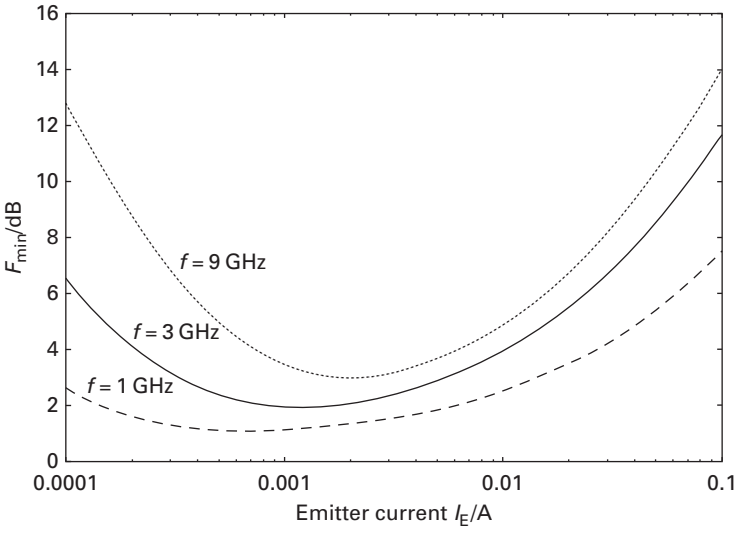

Fig. 2.71 Example calculation of bipolar noise figure dependence on the emitter current. Parameters chosen are: $f_b = 50\,\text{GHz}$, $C_{BE} = 0.2\,\text{pF}$, $\alpha_0 = 0.99$ and $R_B = 10\,\Omega$.

The newly introduced parameter D is

$$D = \frac{1}{\alpha_0}\left[1 + \left(\frac{f}{f_b}\right)^2\right].$$

We note the importance of a low base resistance to achieve a small R_n.

2.5.4 Transit time optimisation

Drift field in the base

We had explicitly assumed that charge carriers traverse the base by diffusion only, that any electric field in the neutral base can be neglected. This is reasonably true provided that the base layer is highly and uniformly doped.

Any significant variation in doping concentration will lead to a built-in electric field which will either enhance or impede the carrier transport in the base. Advantageously, we make the base doping concentration higher at the base–emitter junction than at the base–collector junction, introducing an accelerating field for charge carriers travelling from emitter to collector.

Figure 2.72 shows the schematic band diagram of such a structure. The conduction and valence bands in the neutral base are titled due to the doping variation, adding a drift field force acting upon both electrons and holes. A constant electric field results if the doping concentration is exponentially varied:

$$N_{A,B}(y) \sim e^{-a\,y}.$$

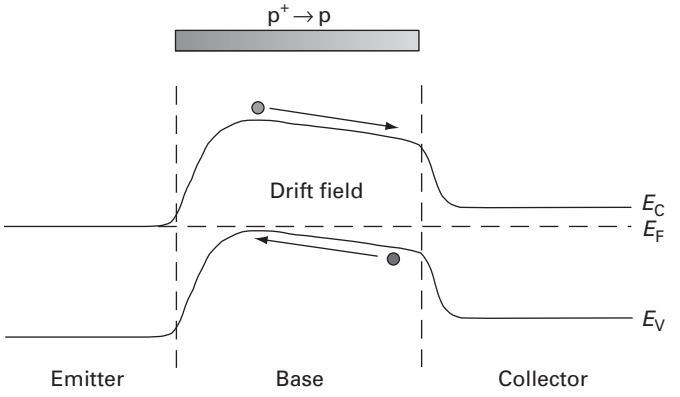

Fig. 2.72 Band diagram of a bipolar transistor with variation of the doping concentration in the base, creating a drift field.

If the base doping is adjusted using ion implantation from the emitter side, as is commonly the case in today's bipolar technologies, a suitable doping profile automatically results.

The built-in field can be easily calculated provided that the Boltzmann approximation is assumed to be valid:

$$\mathcal{E}_{y,bi} = \frac{kT}{2\,q\,W_B} \ln \frac{N_{A,B,max}}{N_{A,B,min}}.$$

The base transit time under the influence of this built-in field is then [65]:

$$\tau_B = \frac{W_B^2}{2\left[1 + \left(q\frac{\mathcal{E}_{y,bi}\,W_B}{kT}\right)^{3/2}\right] D_n}. \tag{2.159}$$

Even modest variations of the base doping concentrations can result in substantial reductions in base transit time.

Collector transit time optimisation

The collector transit time Equation (2.151) can become a significant part of the total transit time, especially in devices with high breakdown voltages. Optimising the collector design involves important design compromises:

- For high f_T, the device needs to be driven to high collector currents, minimising the emitter charging time constant. Therefore, the Kirk effect needs to be pushed to higher currents, demanding a larger collector doping concentration. Equally, the collector transit time needs to be reduced by reducing the depleted collector width W_C. For a given collector–base voltage, this agrees with the demanded increase in collector doping concentration.
- However, a high maximum frequency of oscillation needs a low C_{BC} which, for a given collector–base voltage, demands a decrease in the collector doping concentration.

- Equally, an increase in collector–base breakdown voltage needs a decrease in collector doping concentration and a larger W_C.

The link between breakdown voltage and transit frequency is frequently expressed in terms of the *Johnson limit* [17]. To derive the Johnson limit, let us assume that the collector transit time fully dominates the total transit time. Using Equation (2.151), we find

$$f_T \approx \frac{1}{2\pi\,\tau_C} = \frac{v_{sat}}{\pi\,W_C} \rightarrow W_C = \frac{v_{sat}}{\pi\,f_T}.$$

If $\mathcal{E}_{crit}$ is the critical field for breakdown and if we assume homogeneous doping in the collector, the collector–base breakdown voltage (open emitter terminal) is

$$BV_{CBO} = \mathcal{E}_{crit}\,\frac{W_C}{2}.$$

The product of transit frequency and breakdown voltage will then only depend on the material properties V_{sat} and BV_{CBO}:

$$f_T \cdot BV_{CBO} = \frac{\mathcal{E}_{crit}\,v_{sat}}{2\pi}.$$

The collector–emitter breakdown voltage BV_{CEO} is lower than BV_{CBO} because the impact ionisation current is amplified by the current gain B when entering the base:

$$BV_{CEO} = \frac{BV_{CBO}}{\sqrt[m]{B}}, \tag{2.160}$$

where m is a parameter which depends on the exact geometry and doping of the transistor.

We find for the Johnson limit:

$$f_T \cdot BV_{CEO} = \frac{\mathcal{E}_{crit}\,v_{sat}}{2\pi\,\sqrt[m]{B}}. \tag{2.161}$$

For Si, $\mathcal{E}_{crit} \approx 5 \cdot 10^5\,\mathrm{V\,cm^{-1}}$. The drift saturation velocity at room temperature is $v_{sat} \approx 10^7\,\mathrm{cm\,s^{-1}}$. Assuming a typical $B = 250$ and $m = 4$, we find $f_T \cdot BV_{CEO} = 200\,\mathrm{GHz}$. This is the frequently quoted 'Johnson Limit' for silicon bipolar devices. We readily recognise from Equation (2.161) that it is not a constant and can be significantly different for other values of B and m.

2.5.5 Heterojunction bipolar transistors

The base design dilemma

In our discussion of homojunction bipolar transistors, three main parameters with crucial impact on the high frequency performance were identified:

(i) the base transit time which sets the 'intrinsic speed' of the transistors;
(ii) the base resistance which affects the maximum frequency of oscillation and the noise performance;
(iii) the current gain which not only influences the noise performance, but also has to be typically ≥ 100 to simplify circuit design.

The dilemma is that the base design parameters W_B, $N_{A,B}$ influence these parameters in different ways:

- $\tau_B \sim W_B^2$ and increases, albeit weakly, with increasing N_A due to the reduction in the minority carrier mobility;
- $R_B \sim \frac{1}{W_B N_{A,B}}$;
- $\beta \sim \frac{1}{W_B}\frac{N_{D,E}}{N_{A,B}}$.

A transistor with high f_T and f_{max} would therefore have a thin, highly doped base layer. However, the high base doping concentration will decrease the current gain if the emitter doping concentration cannot be proportionally increased.

On the other hand, there are limits to the increase in emitter doping. The main limitation is *bandgap narrowing*. With increasing doping concentration, the band gap in the emitter will decrease. In silicon,

$$\Delta E_G \approx 22.5 \left(\frac{N_D}{10^{18}\,\mathrm{cm}^{-3}}\frac{300\,\mathrm{K}}{T}\right)^{0.5}. \tag{2.162}$$

The decrease in band gap in the emitter will increase the intrinsic carrier concentration there ($n_i \sim e^{-E_G/2kT}$), which in turn lowers the current gain because the base current due to injection of holes from the base into the emitter is

$$J_B \sim \frac{n_{i,E}^2}{N_{D,E}}.$$

We therefore conclude that the base doping concentration cannot be be strongly increased while keeping a high current gain. Therefore, thin-base microwave bipolar transistors have a problem with rather high base resistances.

Two approaches can be taken to solve the base design dilemma:

(i) We can increase the carrier velocity in the base so that a target transit frequency can be met with a larger base width W_B. This has been discussed already in the context of doping variations in the base layer; we will see further down that the effect can be achieved much more elegantly using bandgap variations.

(ii) We can search for a way to increase the base doping concentration while maintaining a sufficiently high current gain. This we will discuss first.

The wide-gap emitter

The first approach to solving the base design dilemma had already been discussed when deriving the equation for the maximum current gain, Equation (2.121). It was noted that it would be advantageous to fabricate the emitter from a material with a lower intrinsic carrier concentration, $n_{i,E}$. Because

$$n_i = \sqrt{N_C N_V}\, e^{E_G/(2kT)},$$

the bandgap in the emitter needs to be increased. Then the maximum current gain becomes

$$B_{\max} = \frac{J_C}{J_B} = \frac{J_{neb}}{J_{pbe}} = \frac{D_{n,B}}{D_{p,E}} \frac{W_E}{W_B} \frac{N_{D,E}}{N_{A,B}} \frac{n_{i,b}^2}{n_{i,E}^2} \sim \frac{N_{D,E}}{N_{A,B}} e^{\Delta E_G/kT} \tag{2.163}$$

with $\Delta E_G = E_{g,E} - E_{g,B}$, neglecting secondary effects like the differences in effective densities of states in the two materials.

The enhancement factor $e^{\Delta E_G/kT}$ can have very large values. Consider as an example:

Emitter: $Al_{0.25}Ga_{0.75}As$ with $E_G = 1.74\,eV$
Base: GaAs with $E_G = 1.42\,eV$
This results in $e^{\Delta E_G/kT} = 2.2 \times 10^5$ at room temperature!

In a technical transistor, this current gain enhancement is traded for a dramatic increase in base doping with a simultaneous decrease in emitter doping. For a Si homojunction transistor, a typical doping combination is $N_{D,E,typ} = 10^{20}\,cm^{-3}$, $N_{A,B,typ} = 10^{18}\,cm^{-3}$, whereas for an AlGaAs/GaAs HBT, $N_{D,E,typ} = 5 \times 10^{17}\,cm^{-3}$, $N_{A,B,typ} = 4 \times 10^{19}\,cm^{-3}$. The reduction in emitter doping in the HBT is necessary to maintain an adequate reverse breakdown voltage of the base–emitter junction – the effect of current gain on the emitter–collector breakdown voltage BV_{CEO} was already discussed (see Equation (2.160)).

Figure 2.73 shows the band diagram of a wide-gap emitter transistor with a graded emitter–base heterostructure. In a graded heterostructure, the two materials used for the emitter and the base are allowed to intermix over a certain distance. We note that the injection of holes from the base into the emitter, which constitutes a major part of the base current, now faces a much larger potential wall than the injection of electrons from the emitter into the base. This provides the intuitive explanation for the potentially huge increase in current gain.

In an abrupt heterostructure, such as considered for the HEMT (Figure 2.17), we need to consider the effect of the conduction band and valence band discontinuities resulting

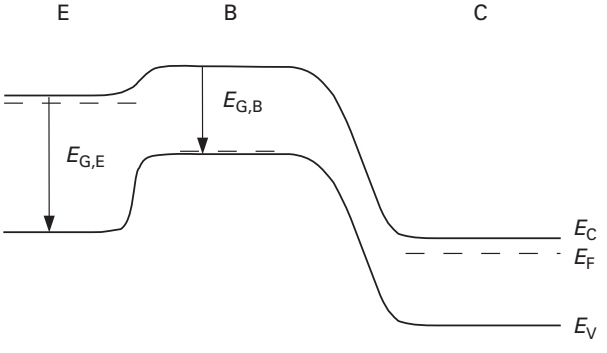

Fig. 2.73 Band diagram of a wide-gap emitter HBT under bias ($V_{BE} > 0$, $V_{CB} > 0$) with graded heterojunction.

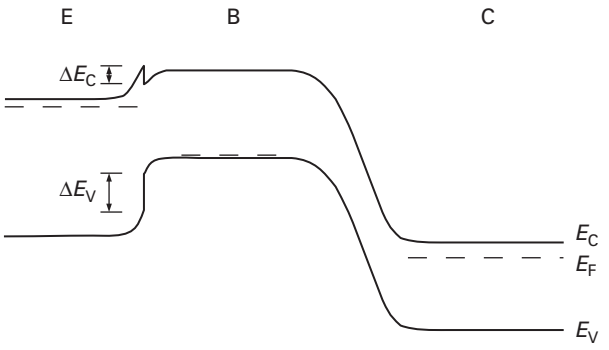

Fig. 2.74 Band diagram of an abrupt heterojunction wide-gap emitter HBT under bias ($V_{BE} > 0$, $V_{CB} > 0$).

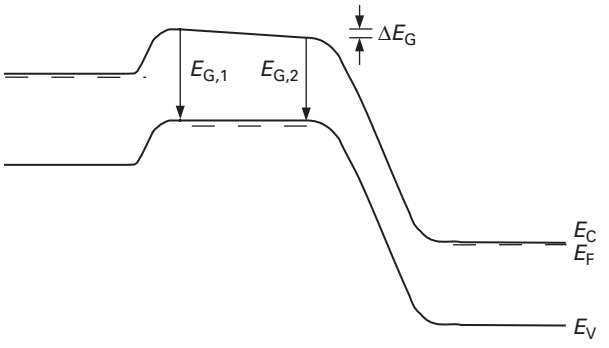

Fig. 2.75 Introduction of a drift field in the base using bandgap variation.

from Anderson's rule (Section 1.20.1). The conduction band discontinuity will lead to an additional energy barrier for electrons (see Figure 2.74).

Assuming purely thermionic emission of electrons over the conduction band barrier, we can write in first order for the maximum current gain of the abrupt HBT, compare Equation (2.163):

$$B_{\max,\text{abrupt}} \approx B_{\max,\text{graded}}\, e^{-\Delta E_C/kT} \sim \frac{N_{D,E}}{N_{A,B}} e^{\Delta E_V/kT}. \tag{2.164}$$

The enhancement factor in this case is only related to the part of ΔE_G which occurs in the valence band. In the above example, $Al_{0.25}Ga_{0.75}As/GaAs$, the enhancement factor is now only 107, because $\Delta E_V = 0.38\,\Delta E_C$.

Drift base

Using compositional grading in the base, we can also introduce a drift field in the base, as shown in Figure 2.75. The emitter has now the same band gap $E_{G,1}$ as the base immediately adjacent to the emitter–base junction. The band gap is then reduced towards the

collector to $E_{G,2} < E_{G,1}$. The resulting bandgap difference ΔE_G reduces the base transit time [9].

$$\tau_{B,\text{graded}} = \tau_{B,\text{ungraded}} \frac{2}{\Delta E_G/kT} \left(1 - \frac{1 - e^{-\Delta E_G/kT}}{\frac{\Delta E_G}{kT}}\right). \tag{2.165}$$

The bandgap reduction is very efficient in reducing the transit time – a modest $\Delta E_G = 4kT$ results in a 62% reduction of the transit time.

HBT implementations

Group III–V HBTs.

HBTs fabricated from group III–V materials such as AlGaAs/GaAs, GaInP/GaAs or InP/InGaAs typically have multiple-mesa structures such as the cross-section shown in Figure 2.76. Due to its cross-sectional shape, it is frequently referred to as a *wedding cake* structure. The structure can be fabricated with a minimum number of masks; the base contacts are usually self-aligned to the emitter mesa using a deliberate undercut of the emitter contact.

While cost-effective in production, this structure has three major drawbacks:

(i) The topology is strongly non-planar. This makes realisations of sub-micron lateral feature sizes difficult, as well as the implementation of multi-level interconnect systems.

(ii) The necessary area for the base contacts and allowances for alignment accuracy necessarily lead to a base–collector area which is substantially larger than the base–emitter junction area. This leads to a larger-than-necessary base–collector capacitance C_{BC}, which in turn lowers the maximum frequency of oscillation (see Equation (2.153)).

(iii) The base–emitter junction is not embedded in semiconductor material, but reaches the less-than-ideal interface with the passivation layer. This gives rise to enhanced surface recombination currents, which increase the non-ideal portion of the base current (see Equation (2.122)). As a consequence, III–V HBTs have a current gain which is strongly dependent on the collector current.

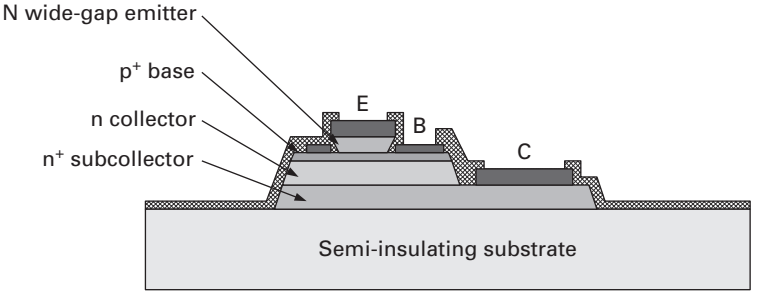

Fig. 2.76 Generic HBT structure typical of III–V semiconductor materials.

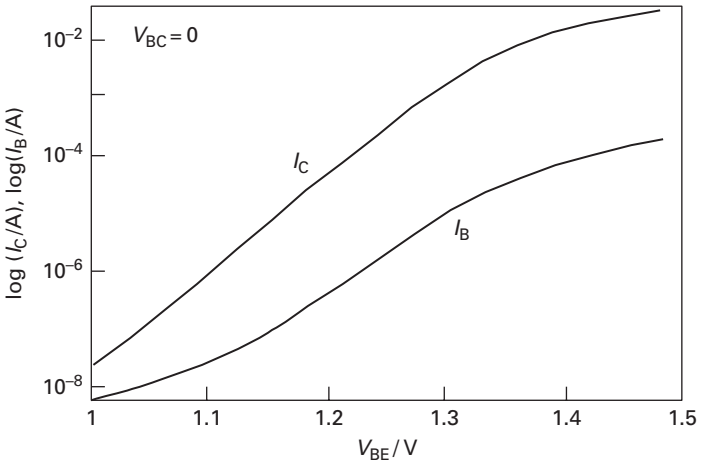

Fig. 2.77 Gummel plot of a typical AlGaAs/GaAs HBT.

Figure 2.77 shows the Gummel plot representation of the collector and base currents of a typical AlGaAs/GaAs HBT. The Gummel plot displays the currents in a semi-logarithmic way as a function of V_{BE}, for $V_{BC} = 0$. Ideally, base current and collector current both have an emission factor (ideality factor) of 1 (see Equation (2.121)). This would result in perfectly parallel curves for $\log I_C$ and $\log I_B$, which is not the case here. The non-ideal base currents, with their emission factor >1, are seen predominantly for very low base–emitter voltages and reduce the current gain there. This is a problem especially for low-noise operation. Furthermore, the surface recombination currents give rise to low-frequency noise ($1/f$ or generation-recombination type noise), with negative impact e.g. on the phase noise of microwave oscillators.

For the wide-gap emitter, AlGaAs was long the material of choice. It was more recently largely replaced with $Ga_{1-x}In_xP$, which for an In mole fraction of 0.5 is lattice-matched to GaAs. GaInP as the emitter material has several advantages:

- The bandgap difference at the GaInP/GaAs junction occurs predominantly in the valence band.
- GaInP can be selectively etched with respect to GaAs, allowing for an automatic etch stop on the base layer when structuring the emitter mesa.
- The reliability of the base–emitter junction under current stress was shown to be substantially higher.

For optoelectronic integration and millimetre-wave applications, HBTs are also being fabricated in InP substrates. The base is now $In_{0.53}Ga_{0.47}As$. As was discussed for the pseudomorphic HEMT structure, the electron mobility is substantially higher than for GaAs, leading to a much shorter base transit time (see Equation (2.143)). The hole mobility in InGaAs, however, is lower than in GaAs, leading to an increased base resistance. The emitter material is either InP or $In_{0.52}Al_{0.48}As$.

HBTs with InGaAs base and collector regions have a major problem with low collector–base breakdown voltages, because the lower band gap of InGaAs lowers

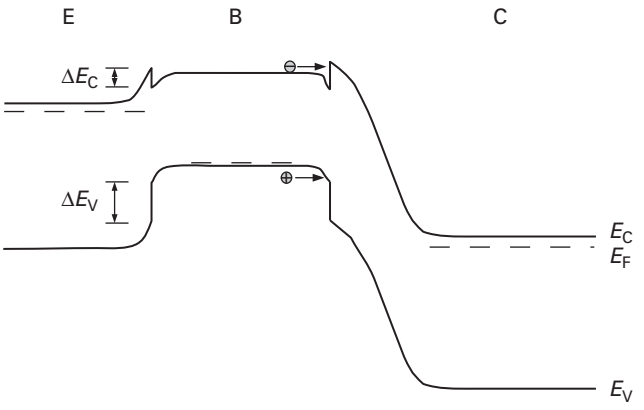

Fig. 2.78 Band diagram of a DHBT under active forward bias.

the threshold for impact ionisation. Therefore, double-heterojunction bipolar transistors (DHBTs) are frequently used in this material system, with either InAlAs or InP as the collector material. DHBTs, however, introduce another problem, which is illustrated in Figure 2.78. The heterostructure at the base–collector interface introduced additional energy barriers in the conduction and valence bands. The conduction band barrier impedes the collection of electrons and lowers the current gain.

The valence band barrier is important especially after the onset of the Kirk effect (see Section 'Kirk effect'). After all fixed donors in the collector have been neutralised, charge neutrality requires that with a further increase in current, free holes from the base are injected into the collector region. Due to the valence band barrier, however, this is restricted in the DHBT. As a result, holes accumulate at the base–collector interface, the bands bend upwards, and the electron barrier in the conduction band becomes higher. This leads to a much more severe deterioration of transistor parameters in the high-current regime. The problem can be avoided either by compositionally grading the base–collector junction, or by introducing a composite collector structure where the heterojunction is offset away from the base–collector p–n junction into the collector [18].

Another problem related to the collector region is the aforementioned substantial C_{BC} due to the triple mesa structure of III–V HBTs (see Figure 2.76). One solution is to fabricate the subcollector in a buried fashion by ion implantation beneath a semi-insulating layer, and to connect it to the collector contacts and to the collector proper via 'sinker' implants.

An example for such a structure is shown in Figure 2.79 [44]. The subcollector is implanted into the semi-insulating InP substrate; the layer above has a drastically increased conductivity due to an Fe implant. Heavily n-doped local implants connect the buried subcollector to the collector contacts and to the collector itself, which is grown together with the base and emitter layers subsequently. While the base–collector area is not changed here, C_{BC} is still reduced because the collector layer itself is depleted in normal operation and therefore the reduced collector–subcollector interface area diminishes the capacitance. Further, C_{BC} is less V_{CE}-dependent, which enhances linearity.

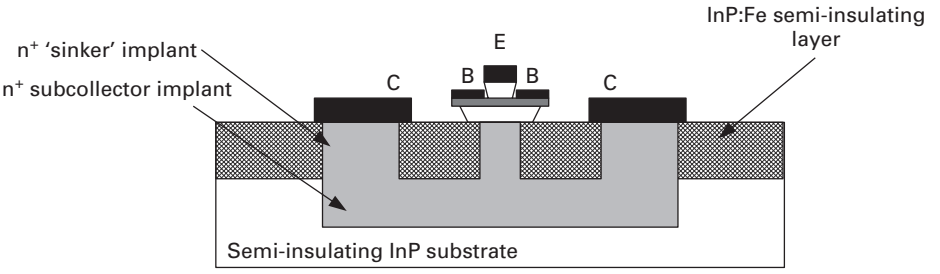

Fig. 2.79 InP/InGaAs HBT structure with a buried subcollector.

Finally, burying the subcollector improves the planarity of the structure. In the example shown, the InP/InGaAs DHBT demonstrated $f_T = 350\,\text{GHz}$ and $f_{max} = 400\,\text{GHz}$.

Another method to restrict C_{BC} is by damage implant through the base contact window prior to metal deposition.

C_{BC} can be further reduced by eliminating the subcollector altogether and attaching the contact directly to the collector layer. To achieve this, the HBT structure must be inverted, i.e. the emitter contact is now at the bottom [50]. The collector must be accurately aligned to the buried emitter structure. A new problem which arises in classical collector-up HBTs is that now the emitter contact must be made laterally through a 'sub-emitter' layer, increasing the crucial emitter series resistance.

The latter problem is addressed in the very aggressive 'transferred substrate' device, where the HBT structure is grown emitter-up. The emitter and base/collector structures are fabricated first. The structure is then flipped around and the emitter is attached to a Au metal structure which provides for the low-resistivity lateral emitter contact. The InP substrate is then removed and the collector contact is structured. The collector layer outside of the contact area is fully depleted and does not add extra capacitance.

A schematic cross-section is shown in Figure 2.80 [47]. Together with submicron scaling ($0.4\,\mu\text{m} \cdot 6\,\mu\text{m}$ emitter area, $0.7\,\mu\text{m} \cdot 10\,\mu\text{m}$ collector area), an InP/InGaAs HBT with a transferred-substrate structure exhibited $f_T = 204\,\text{GHz}$ and $f_{max} = 1080\,\text{GHz}$.

Si/SiGe HBTs.

Unlike III–V HBTs, which are usually fabricated from lattice-matched heterostructures, HBTs in the $Si_{1-x}Ge_x$ material system are necessarily pseudomorphic (Section 1.20), which delayed their practical realisation until the late 1980s. They are commercially available since 1998 and have enjoyed an unparalleled technical and commercial success.

Due to the large difference in lattice constant between Si ($a = 5.43\,\text{Å}$) and Ge ($a = 5.66\,\text{Å}$), an elastically strained SiGe layer will necessarily be very thin, as was shown in Figure 1.35. The use of SiGe compounds is therefore restricted to the base layer – everything else is silicon, making Si/SiGe transistors necessarily DHBTs.

The SiGe alloy can be used in two different ways:

(i) The base may start with a zero Ge mole fraction at the emitter–base junction, and be increased towards the base–collector junction. The corresponding decrease in

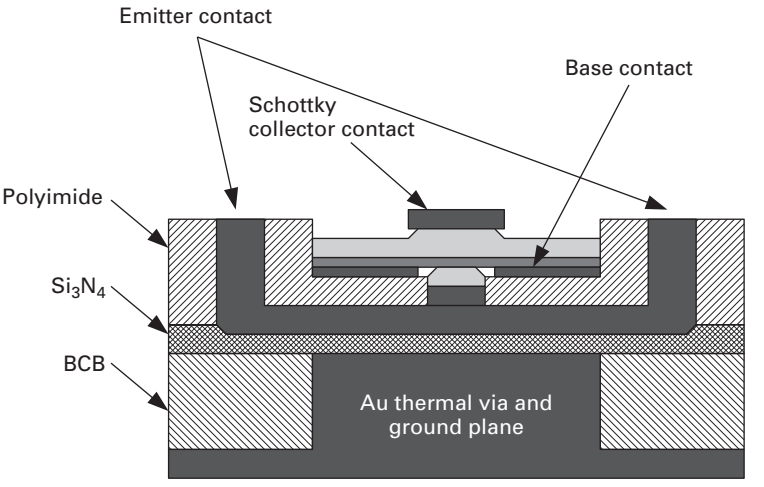

Fig. 2.80 Schematic cross-section of a transferred-substrate collector-up HBT [47].

band gap creates a built-in field for electrons in n–p–n transistors – a drift-base transistor results, with a band diagram similar to the one shown in Figure 2.75, except that the base–collector interface is now a hetero-interface. Pseudomorphic SiGe layers sandwiched between relaxed Si layers have an interesting property: the bandgap difference is almost exclusively in the valence band (see Figure 1.35). Hence, there is no parasitic conduction band barrier, at least not until high-current effects set in and the hole pile-up against the base–collector valence band barrier makes the bands buckle upwards.

The major benefit of the built-in drift field is the reduction in base transit time given by Equation (2.165). Due to the emitter–base interface being a homojunction, it is bound by similar base doping limitations as the homojunction bipolar transistor.

This Si/SiGe drift-base concept has the significant advantage that the average Ge mole fraction in the base, and with it the built-in mechanical strain, is very low. In terms of processing temperatures, these transistors are fully CMOS-compatible. The drift-base heterostructure transistor is therefore the approach of choice in most Si/SiGe BiCMOS processes.

(ii) Si/SiGe heterostructures can, of course, also be used to fabricate a wide-gap emitter structure. In this case, the Ge mole fraction is already significant at the emitter–base junction, leading to a significant valence band discontinuity, which allows to dramatically increase the base doping concentration (see Equation (2.164)). In these transistors, the Ge mole fraction is typically constant across the base.

The major benefit of the wide-gap emitter structure is the high base doping concentration and resulting low base sheet resistance, which allows to achieve high cutoff frequencies despite very relaxed lateral scaling rules, e.g. f_T, $f_{max} = 80\,\text{GHz}$ with $0.8\,\mu\text{m}$ design rules [55] .

The two approaches may be combined, of course – the Ge mole fraction profile may start with a moderate non-zero value at the emitter–base interface and increase towards

the collector to a higher value at the base–collector junction, combining a hole-blocking effect towards the emitter with a built-in drift field towards the collector. This is called a *trapezoidal* Germanium profile in the base. The current gain in this case is [26]:

$$B_{\mathrm{SiGe}} = B_{\mathrm{Si}}\,\eta\,\gamma\,\frac{E_{\mathrm{G}}(y=0) - E_{\mathrm{G}}(y=W_{\mathrm{B}})}{kT}\,\frac{e^{\Delta E_{\mathrm{G}}(y=0)/\mathrm{kT}}}{1 - e^{-[E_{\mathrm{G}}(y=0) - E_{\mathrm{G}}(y=W_{\mathrm{B}})/kT]}}, \quad (2.166)$$

where B_{Si} is the current gain of a homojunction transistor with the same geometry, η is the ratio of the position-averaged minority mobilities in the base of the two transistors, and γ is the position-averaged ratio of the density of states product $(N_{\mathrm{V}} \cdot N_{\mathrm{C}})$ across the base. The emitter–base junction is at $y = 0$, and the base–collector interface at $y = W_{\mathrm{B}}$.

Since

$$\lim_{\mathrm{x}\to 0} \frac{x}{1 - e^{-x}} = 1,$$

Equation (2.166) reverts to Equation (2.163) for $E_{\mathrm{G}}(y = 0) = E_{\mathrm{G}}(y = W_{\mathrm{B}})$. On the other hand, we see that having a pure drift-base profile $(E_{\mathrm{G}}(y = 0) = 0)$ also results in a certain increase in the current gain.

Irrespective of the Ge profile in the base, a major advantage of the Si/SiGe HBTs is that they can harness the full potential of silicon technology, especially aggressive lateral scaling developed predominantly for CMOS process, different isolation techniques, and SiO_2 as a highly stable native oxide.

A typical SiGe HBT in a commercially available technology has a structure similar to the schematic in Figure 2.81. Note the very planar structure compared to III–V HBTs, and the extensive use of SiO_2 isolation. The n^+ subcollector is created by ion implantation, after which a low n-doped Si layer is epitaxially grown and converted to SiO_2 by local oxidisation, except in the areas below the collector contact and where the transistor structure will be. The collector area is doped using selective ion implantation, which allows for several collector doping concentrations on one chip, with different f_{T} versus BV_{CEO} trade-offs. The transistor structure is then grown selectively in the transistor window.

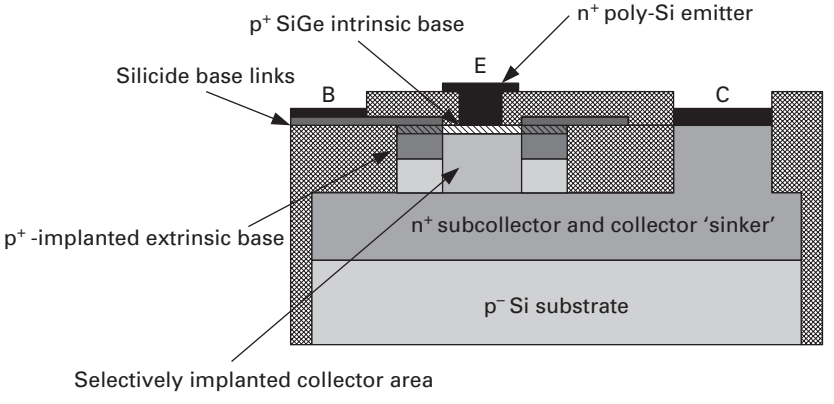

Fig. 2.81 Planar Si/SiGe HBT with implanted extrinsic base region.

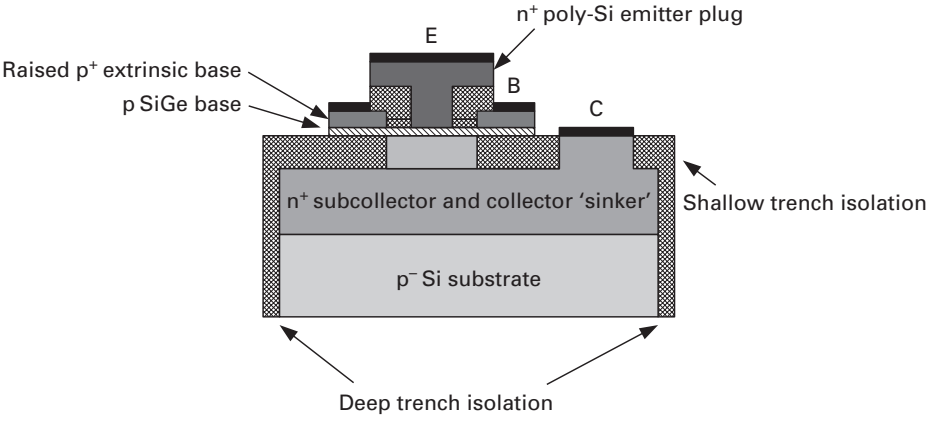

Fig. 2.82 Planar Si/SiGe HBT with a raised extrinsic base structure.

The extrinsic base resistance is reduced by heavy p^+ implantation. This works well if the transistor is not aggressively scaled laterally. It does create, however, crystal faults immediately adjacent to the intrinsic base, which leads to enhanced diffusion of the p-dopant in the base and is a major obstacle to fabricating deep submicron lateral emitter widths. Additionally, the close proximity of the p^+ extrinsic base and the selectively implanted collector increases the base–collector capacitance.

The latter problems are solved using a raised base structure, where a p^+ extrinsic layer is grown selectively on top of the base adjacent to the emitter, as shown schematically in Figure 2.82 [14]. A combination of these techniques with deep submicron scaling led production Si/SiGe HBT technologies to achieve f_T and f_{max} values above 200 GHz.

III–V versus Si/SiGe HBTs – a brief comparison.

Si/SiGe HBTs displaced III–V HBTs in many applications primarily due to their supreme potential for large-scale integration, owing to their technological proximity to very mature Si processes. In terms of raw speed, as measured from f_T and f_{max}, record values are still scored by InP/InGaAs devices, but Si/SiGe HBTs are competitive, because they compensate for material deficiencies (e.g. the much lower electron mobility versus InGaAs), by aggressive lateral scaling and superior suppression of parasitic capacitances. Further, Si has a significantly higher thermal conductivity than either GaAs or InP, which makes the thermal management of dense transistor arrays easier.

In the area of microwave power amplification, however, III–V-based HBTs have an inherent advantage. When deriving the Johnson limit, Equation (2.161), we recognised the importance of the product of drift saturation velocity v_{sat} and the electrical field necessary for impact ionisation $\mathcal{E}_{crit}$. Taking v_{sat} at an electric field of $10\,kV\,cm^{-1}$, and $\mathcal{E}_{sat}$ at a donor doping concentration of $10^{17}\,cm^{-3}$, this product is shown for Si, GaAs and InP in Table 2.1.

When comparing practical transistors, the ratio in $f_T\,BV_{CEO}$ between Si- and GaAs-based HBTs may appear even larger than the factor of 1.6 suggested by Table 2.1; but

Table 2.1 $v_{sat} \cdot \mathcal{E}_{crit}$ product for Si, GaAs and InP

Si	5,000	GHz V
GaAs	8,000	GHz V
InP	22,000	GHz V

this is due to the generally lower current gain in the GaAs devices, which increases BV_{CEO}.

2.5.6 Large-signal modelling

Bipolar transistor models have become increasingly complex. An exhaustive description of popular large-signal formulations is beyond the scope of this book. The following will concentrate on emphasising the major differences between the models, with respect to active forward operation of the transistor, quasi-static non-linear equations and avoiding extreme areas of operation.

The Ebers–Moll model

The Ebers–Moll equivalent circuit model was historically the first compact model of the bipolar transistor [16]. It approximates the intrinsic transistor as a network of two junction diodes and two current-controlled current sources (see Figure 2.83(a)). The parameter A_F is the common-base current gain in forward operation. A_R is the common-base current gain in reverse operation (emitter and base interchanged), which is not being considered here. For active forward operation, the base–collector diode is reverse-biased. Further, $A_R\, I_C$ is much smaller than the forward current through the base–emitter diode and can hence be neglected. The resulting simplified equivalent circuit is shown in Figure 2.83(b).

The emitter current in active forward operation is

$$I_E = -I_{SBE}\left(e^{V_{BE}/(N_E V_T)} - 1\right), \tag{2.167}$$

where I_{SBE} is the base–emitter saturation current, N_E the base–emitter ideality factor and $V_T = kT/q$ the thermal voltage.

Using the full equivalent circuit, the Ebers–Moll equivalent circuit can account for saturation (both diodes are forward-biased), but cannot model Early and Kirk effects. Further, the current dependence of the current gain at low V_{BE} can also not be included.

The Gummel–Poon model

An improved model of the bipolar transistor which is capable of including more of the non-ideal effects of bipolar transistors was introduced by Gummel and Poon in 1970 [22].

The equivalent circuit (Figure 2.84) uses a voltage-controlled current source for the collector current – the transistor is seen in common-emitter configuration here. The

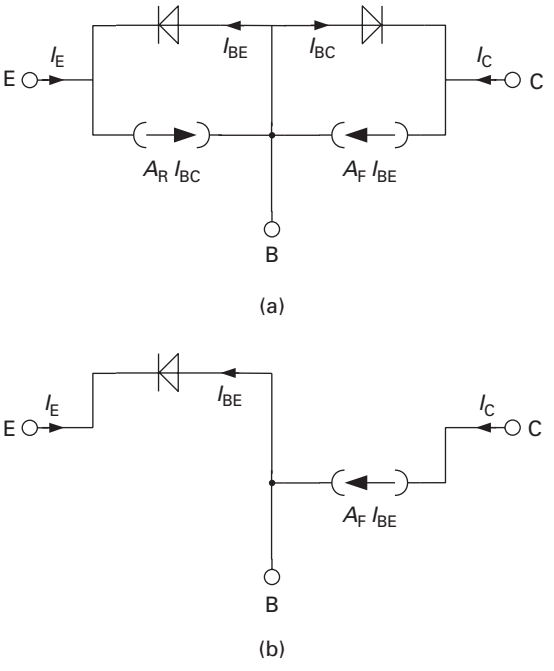

Fig. 2.83 Quasi-static Ebers–Moll equivalent circuit of the intrinsic bipolar transistor: (a) for forward and reverse operation; (b) simplified for forward active operation only.

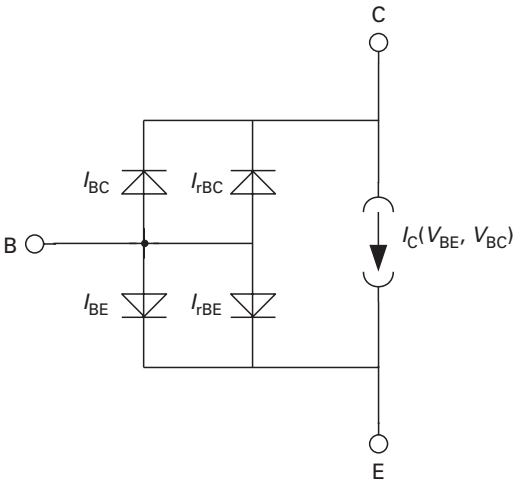

Fig. 2.84 Gummel–Poon quasi-static equivalent circuit of the intrinsic bipolar transistor.

use of two parallel diodes for the base–emitter and base–collector junctions allows to include both the ideal (I_{BE}, I_{BC}) and non-ideal (I_{rBE}, I_{rBC}) current contributions in forward and reverse operations of the transistor, with different emission factors. Hence, the current gain reduction at low V_{BE} (or V_{BC} in reverse operation) can be easily included.

The collector current formulation (shown here for forward operation only) uses a saturation current I_S and a charge control parameter Q_B:

$$I_C = \frac{I_S}{Q_B}\left(e^{V_{BE}/(N_F\,V_T)} - e^{V_{BC}/(N_R\,V_T)}\right), \tag{2.168}$$

where N_F is the ideality factor in forward direction and N_R the ideality factor in reverse direction. Early and Kirk effects are modelled through the charge control parameter:

$$\begin{aligned} Q_B &= \frac{Q_1}{2}\left(1+\sqrt{1+4Q_2}\right) \\ Q_1 &= \left(1-\frac{V_{CB}}{V_{AF}}\right)^{-1} \\ Q_2 &= \frac{I_S}{IKF}\left(e^{V_{BE}/(N_F\,V_T)} - 1\right), \end{aligned} \tag{2.169}$$

where V_{AF} is the Early voltage in forward direction and IKF is the knee current for the onset of high-current effects in the forward direction.

In the active forward regime, I_{BC} and I_{rBC} can be neglected and the base current becomes:

$$\begin{aligned} I_B &= I_{BE} + I_{rBE} \\ &= \frac{I_S}{BF}\left(e^{V_{BE}/(N_F\,V_T)} - 1\right) + I_{SE}\left(e^{V_{BE}/(N_E\,V_T)} - 1\right), \end{aligned} \tag{2.170}$$

where BF is the ideal forward current gain, I_{SE} the saturation current of the non-ideal base current and N_E the emission factor of the non-ideal base current.

Extension of the equivalent circuit to the dynamic case is shown in Figure 2.85. In active forward operation, C_{BE} contains both the diffusion capacitance Equation (2.141) and the junction capacitance of the base–emitter junction, while C_{BC} is a junction capacitance only. The capacitance C_{CS} models the reverse-biased junction between the (sub-)collector region and the substrate node. On semi-insulating substrates, it is not necessary.

The Gummel–Poon model also deals with the bias dependence of the base resistance which is frequently observed – R_B decreases from a higher value at low collector current to a much lower value at high collector current. This effect is due to a concentration of the emitter current towards the emitter periphery with increasing current – due to the lateral voltage drop in the base layer, the local base–emitter voltage is higher and closer to the base contact. As the local current depends exponentially on the local V_{BE}, even a small voltage change can lead to substantial redistributions in current. The base resistance decreases because the inner parts of the emitter–base area get increasingly detached. Due to the much lower base sheet resistance, this effect is less pronounced in wide-gap emitter HBTs. The Gummel–Poon model describes this effect using the parameters RB, RBM and IRB:

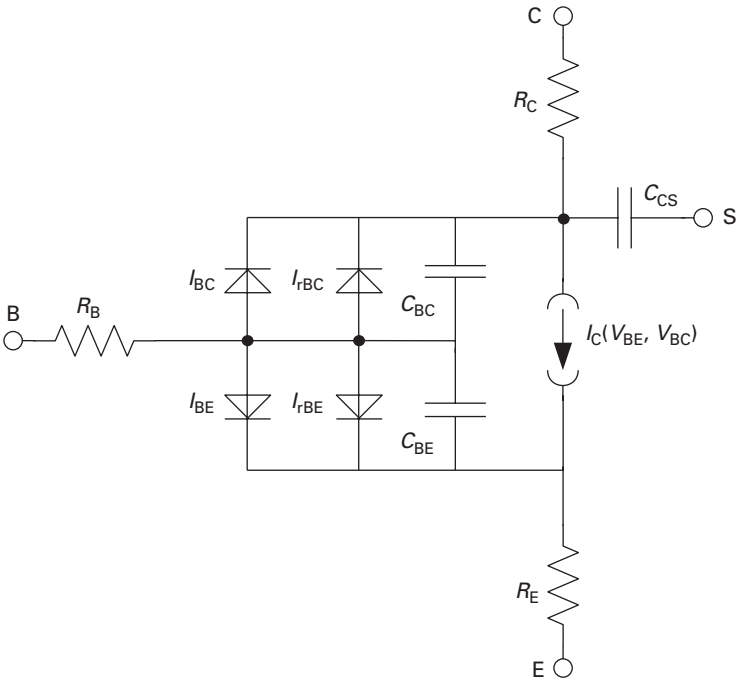

Fig. 2.85 Gummel–Poon equivalent circuit with parasitic elements and substrate node.

$$R_B(I_B) = \mathrm{RBM} + 3\,(\mathrm{RB} - \mathrm{RBM})\frac{\tan(z) - z}{z \cdot \tan^2(z)}$$
$$z = \frac{\sqrt{1 + \left(\frac{12}{\pi}\right)^2 \frac{I_B}{IRB}} - 1}{\left(\frac{24}{\pi^2}\right)\sqrt{\frac{I_B}{IRB}}}. \tag{2.171}$$

In total, the Gummel–Poon model implemented in SPICE contains 42 model parameters. A full discussion is therefore beyond the scope of this book.

The VBIC95 model

The VBIC95 model [37] is an extension of the Gummel–Poon model. Among others, the following problems are being addressed:

- The description of base width modulation using a constant Early voltage is a simplification which only applies to small V_{CE}.
- Self-thermal effects are not included in Gummel–Poon, yet play an important role especially for power amplifiers.
- The collector resistance is not a constant, but depends on V_{CB}, because the undepleted part of the collector increases the series resistance.
- Avalanche breakdown in the collector space charge region needs to be included.

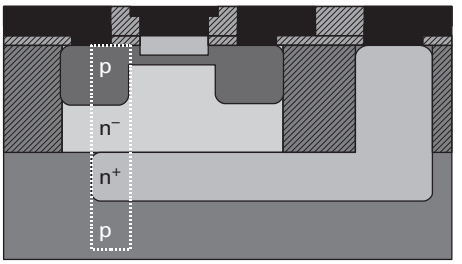

Fig. 2.86 Generic n–p–n BJT cross-section highlighting the parasitic p–n–p transistor.

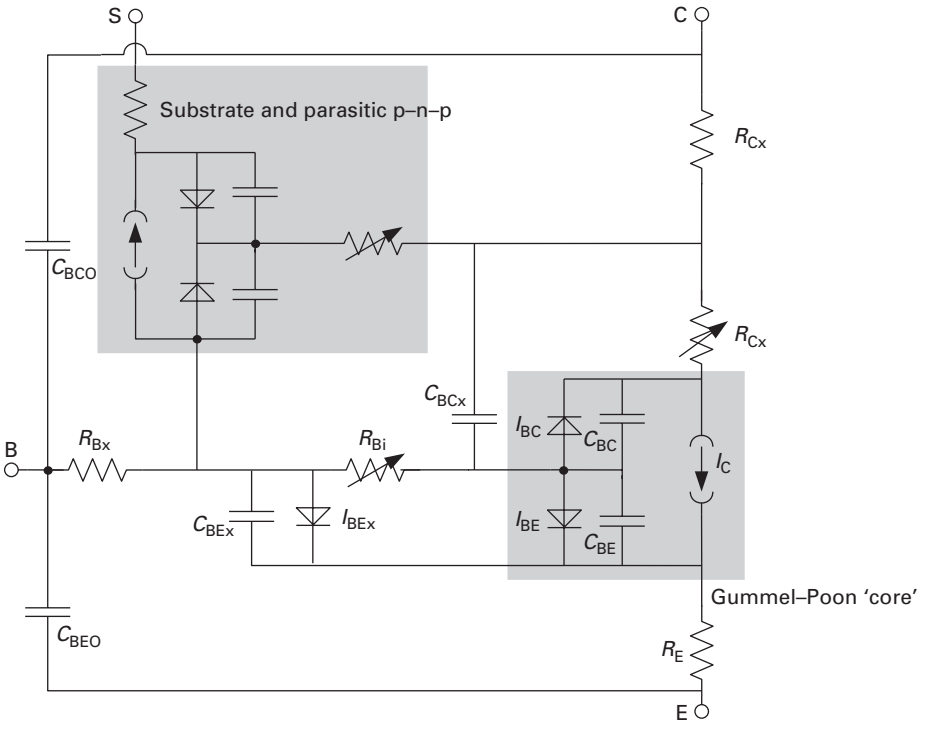

Fig. 2.87 VBIC95 equivalent circuit.

- A major addition has been the implementation of a subcircuit for the parasitic p–n–p transistor, which is formed in Si-based bipolar transistors between the base, the collector and the substrate. This parasitic p–n–p can be easily recognised in Figure 2.86. Under certain bias conditions, it may sink an unexpectedly large current between the base terminal and the substrate node.

Figure 2.87 shows the VBIC95 in a representation which emphasises its Gummel–Poon heritage. The distributed nature of the base resistance is accounted for. The collector resistance is now separated into a bias-dependent part which symbolises the contact, sinker and subcollector resistances, and a bias-dependent internal part modelling the

undepleted part of the collector proper. The substrate network is now much more complex and includes the parasitic p–n–p as a separate Gummel–Poon type equivalent circuit. Additional capacitances C_{BEO} and C_{BCO} have been added to account for overlap capacitances between the poly-Si emitter plug and the base and collector contacts, respectively.

Note that these are the only linear capacitances – all other capacitances are bias-dependent, even though this has not been noted in the equivalent circuit to enhance readability.

The VBIC95 model implemented in newer versions of SPICE has 85 parameters, which also hints at the complexity of setting up such a model from measurements.

The MEXTRAM model

The MEXTRAM bipolar transistor model was created by Philips N. V. [42] and released into the public domain in 1993. It has been implemented in several industry standard simulation environments, such as several versions of SPICE and Agilent Advanced Design System (ADS).

The equivalent circuit (Figure 2.88) shows stronger deviations from the Gummel–Poon topology. The main current equation, however, shows the similarity:

$$I_N = \frac{I_S}{q_b}\left(e^{V_{B2E1}/V_T} - e^{V^*_{B2C2}/V_T}\right). \tag{2.172}$$

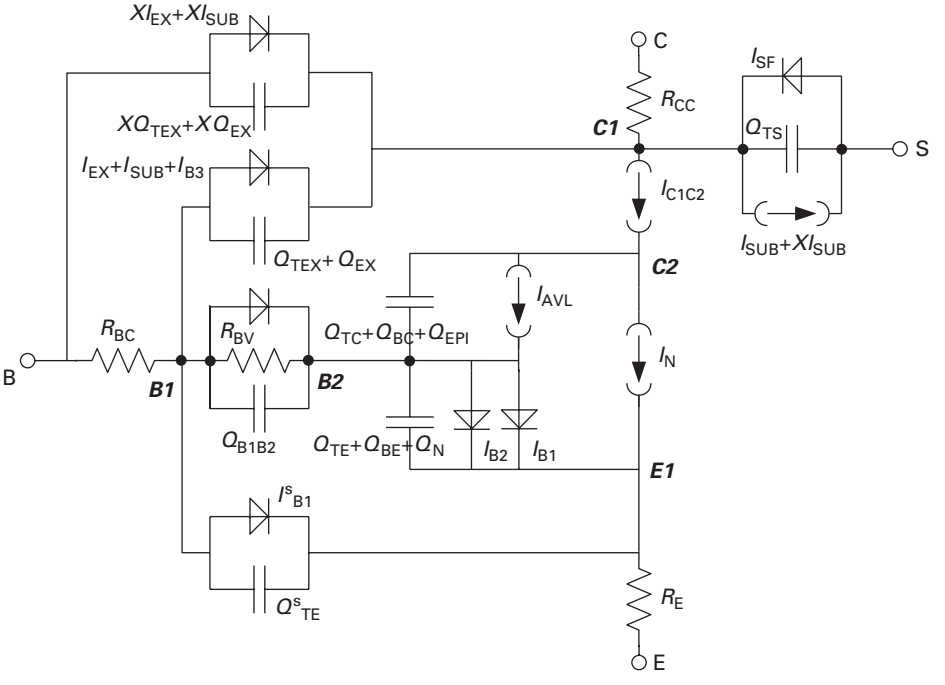

Fig. 2.88 MEXTRAM model equivalent circuit topology.

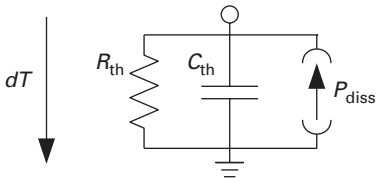

Fig. 2.89 Thermal equivalent circuit used by VBIC95 and MEXTRAM.

Here, V_{B2E1} is the voltage between nodes B2 and E1, while V^*_{B2C2} is a calculated quantity which corresponds to the voltage drop between nodes B2 and C2 – for an explanation of this and other intricacies, please refer to the MEXTRAM documentation [42]. q_b is the normalised base charge, which is used to model both Early and high-current effects. This is conceptually as in Gummel–Poon, but the implemented equations provide a higher level of accuracy, for example in the bias dependence of the Early voltage.

The base is modelled as a distributed structure – this is a must for accurate modelling at elevated frequencies. The base–collector capacitance is split into three partial capacitances. The model does not only distinguish between an external and an internal part of the base, but models the sidewall interface between base and emitter separately (I^S_{B1} and Q^S_{RE}). Two diodes are used to model the ideal and non-ideal parts of the base current.

The parasitic p–n–p transistor is also modelled here, even though this is less obvious – the current source between nodes $C1$ and S is exponentially controlled by the intrinsic base–collector voltage, V_{B1C1}. The reverse behaviour of the parasitic p–n–p, however, is not modelled.

The major claimed advantage over VBIC95 is related to the modelling of high-current effects [30]. This is especially important for double-heterostructure transistors, such as Si/SiGe HBTs (see Figure 2.78 and its associated discussion).

MEXTRAM has also been extended to account for neutral base recombination and base drift fields introduced through bandgap variations, again in an effort to make this model especially useful for Si/SiGe HBTs.

Self-thermal effects are being simulated in VBIC95 and MEXTRAM in the same way (see Figure 2.89). The model calculates the sum of all powers dissipated in resistors and space charge regions as P_{diss}. In the electric equivalent circuit, P_{diss} is treated as a current which creates a voltage drop dT of the parallel connection R_{th}, C_{th}, which establishes the thermal time constant τ_{th}. dT is analogous to the temperature difference between the device (taken to be at one single temperature – a simplification) and the ambient. It is then used as an additional control voltage for the bias-dependent current sources. This is a very common technique to include self-thermal effects, but it neglects the fact that the thermal conductivity of semiconductors, and with it the thermal resistance R_{th}, is temperature-dependent. With increasing temperature T, R_{th} will also increase.

The HICUM model

The acronym of the last model to be discussed here already indicates its major claimed advantage – HICUM [51] stands for **Hi**gh-**C**urrent **M**odel. Aside from being a general purpose non-linear bipolar model, with special emphasis on high-speed applications, it

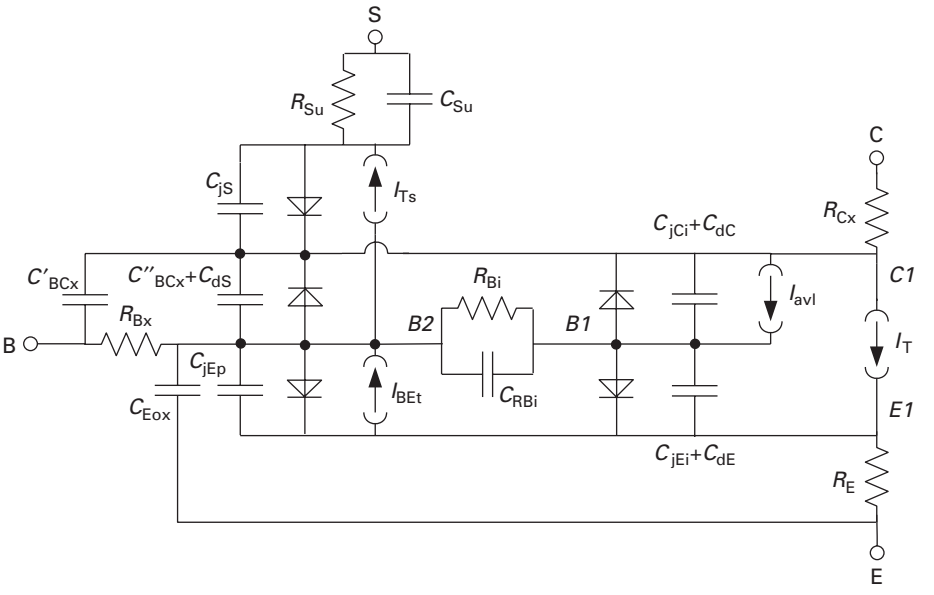

Fig. 2.90 HICUM equivalent circuit (adapted from [52]).

concentrates especially on an accurate prediction of high-current effect. Remember that high-speed bipolar operation will occur at high collector current densities, minimising the emitter charging time (see Figure 2.69). Accurate assessment of high-current effects is therefore a must for any simulator with high-speed emphasis. HICUM's development started in 1980 and it has been implemented in commercial simulation environments since 1994. Its model equations take a semi-physical approach to allow scalability and a certain degree of prediction. A companion program, TRADICA, facilitates the latter two issues.[12]

Figure 2.90 shows the equivalent circuit used in HICUM. The modelling of self-thermal effects is done analogous to Figure 2.89 and is not shown here again.

The HICUM model does not have an equivalent element to the Gummel–Poon ideal current gain BF, but calculates the collector and base currents independently and treats the current gain as a derived entity. The main current in the HICUM model is the transfer current I_Tm, which can be compared to the intrinsic collector current I_C (see Equation (2.168)):

$$I_T = I_S \frac{e^{V_{B1E1}/VT} - e^{V_{B1C1}/V_T}}{\frac{Q_{p,T}}{Q_{p0}}}, \tag{2.173}$$

where Q_{p0} is the total hole charge in the base at zero bias. Note that unlike the Gummel–Poon expression, the exponential function does not have ideality factors. The deviation from non-ideal diode characteristics is handled in the bias-dependent hole charge $Q_{p,T}$.

[12] An introduction to TRADICA is available at www.iee.et.tu-dresden.de/schroter/Trad/features.pdf.

The formulation for $Q_{p,T}$ in HICUM allows to include the effect of strongly varying intrinsic carrier densities across the base layer, as necessary for the simulation of drift-base HBTs [53]. In the model, this is done through different weighting factors being applied to the depletion charges at the base–emitter and base–collector junctions.

The model can also accommodate the hole accumulation at the base–collector heterojunction, with current gain roll-off and transit time deterioration, as needed in DHBTs [54]. For an in-depth treatment of HICUM parameters, refer to [52].

Differences between BJTs and HBTs relevant to large-signal modelling

In general, the standard bipolar models discussed above are also applicable to HBTs, with appropriately chosen parameters.

An important deviation concerns self-heating effects. In homojunction bipolar transistors, both the collector saturation current and the current gain have positive temperature coefficients. The saturation current (see Equation (2.115)) increases because the intrinsic carrier concentration in the base depends exponentially on temperature. The current gain is limited by bandgap narrowing in the heavily doped emitter (see e.g. Equation (2.162)). The bandgap narrowing effect has a negative temperature coefficient, which lets the current gain increase with increasing temperature.

In a wide-gap emitter HBT, the saturation current equally has a positive temperature coefficient. The current gain, however, decreases with increasing temperature. To understand this, investigate again Equation (2.163):

$$B_{\max} \sim e^{\Delta E_G/kT}$$

The enhancement factor therefore decreases with increasing temperature, because the valence band barrier gets less and less efficient.

A frequently observed self-thermal effect in HBTs is the current crush in multi-finger HBTs (see Figure 2.91). At moderate dissipated powers, all fingers will have approximately the same temperature and the current distribution is equal. With increasing V_{CE} and under constant base current, the collector current will gradually decrease due to the negative temperature coefficient of B. However, due to the strongly positive temperature coefficient of the saturation current, a finger which is slightly hotter than the others will draw more and more current, deviating it away from the others, and increase its temperature. This strongly non-linear positive feedback will lead to a situation where the hottest finger takes on all the current, increases dramatically in temperature, with a resulting sudden decrease in current gain. This effect cannot be modelled with the standard bipolar models.

Another important effect which cannot be simulated using the standard models is the rapid onset of high-current effects in DHBTs, discussed in the context of Figure 2.78.

A minor effect, but worth mentioning, is that the non-ideal base current, see Equation (2.122), may have a different V_{BE} dependence in HBTs. This is due to space charge region recombination associated with the conduction band discontinuity at an abrupt emitter–base heterojunction (see Figure 2.74). The potential well on the base side of the junction leads to an increased recombination, which depends in turn on the voltage across the junction.

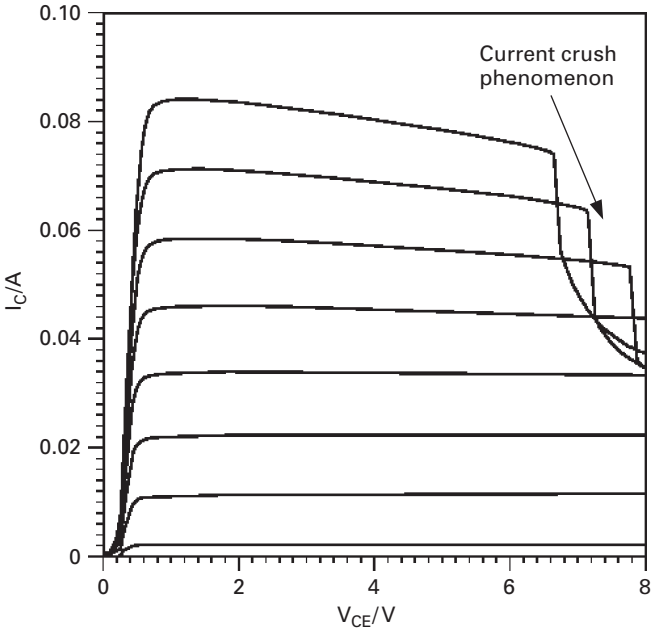

Fig. 2.91 Current crush phenomenon in a multi-finger HBT (http://parts.jpl.nasa.gov/mmic/mmic_complete.pdf). Courtesy NASA/JPL-Caltech.

2.6 Problems

(1) In a MESFET device, the following parameters are known from the fabrication process:

Gate	Channel	Drain/source
Ti, $\Phi_B = 0.7\,\text{eV}$,	$N_D = 1 \cdot 10^{17}\,\text{cm}^{-3}$,	$R_S = R_D = 10\,\Omega$
$L_G = 1\,\mu\text{m}$,	layer thickness	
$W_G = 100\,\mu\text{m}$,	$a = 0.3\,\mu\text{m}$,	
$R_G = 1\,\Omega$	$v_{sat} = 1.2 \cdot 10^7\,\text{cm}\,\text{s}^{-1}$	

a) Draw the qualitative band diagram under the gate in thermodynamic equilibrium.
b) Calculate the Schottky gate built-in voltage and the pinch-off voltage.
c) Assuming constant velocity throughout the channel, calculate the drain current for an applied bias of $V_{GS} - V_P = 2\,\text{V}$ and $V_{DS} = 3\,\text{V}$.
d) Calculate the device transconductance.
e) What is the expected transit frequency?
f) Calculate the expected minimum noise figure F_{min} at a frequency of 2 GHz.

(2) A GaAs MESFET with a gate width of $W_G = 100\,\mu\text{m}$ is specified with a transit frequency $f_T = 35\,\text{GHz}$ and a maximum frequency of oscillation $f_{max} = 50\,\text{GHz}$. The transconductance is $g_m = 21\,\text{mS}$, and the gate resistance is $R_G = 2\,\Omega$.

Estimate the gate-source capacitance and the gate-drain capacitance, assuming that the output conductance is negligible. Can you provide an estimated value for the minimum noise figure?

Due to a fabrication error, the gate resistance is increased to 5 Ω. What is the impact on f_T, f_{max} and F_{min}?

(3) Why does a MOSFET require an overlap between the gate and the source implantation region? What does this imply for device capacitances?

(4) In an n-channel MOSFET with a *metal* gate electrode, the original gate metal is replaced by a metal with a smaller work function. Explain qualitatively the effect on the threshold voltage.

(5) In a MOSFET technology, the thickness of the field oxide is chosen such that under the highest possible voltage between metallisation and substrate, no inversion channel can form at the SiO_2/Si interface. Considering an Al metallisation with a work function of 4.1 eV and a bulk doping concentration of $N_A = 5 \cdot 10^{17}$, calculate the minimum thickness of the field oxide, if the maximum voltage between Al metallisation and the substrate is 5 V.

(6) A silicon-on-insulator n-channel MOSFET has a p-doped 'bulk' layer above the oxide layer with a doping concentration of $N_A = 2 \cdot 10^{16}\,cm^{-3}$. The gate is heavily n-doped poly-Si. The gate oxide thickness is 5 nm.

Calculate the thickness of the doped layer such that it will be fully depleted in active device operation. What is the purpose of the buried oxide layer? Explain its effect(s) on device performance.

(7) In a HEMT, what is the purpose of the spacer layer? Would the device still function without it?

(8) A HEMT device has the following layer structure:

Function	**Composition**	**Thickness**	**Doping concentration**
Supply	$Al_{0.3}Ga_{0.7}As$	80 nm	$N_D = 3 \cdot 10^{17}\,cm^{-3}$
Spacer	$Al_{0.3}Ga_{0.7}As$	5 nm	Nominally undoped
Buffer	GaAs	100 nm	$N_A = 1 \cdot 10^{15}\,cm^{-13}$
Substrate	GaAs	150 µm	Intrinsic

Calculate the threshold voltage of this device at room temperature.

Let now $V_{GS} - V_{off} = 0.5$ V. Calculate the sheet density charge of the 2DEG.

(9) You want to optimise the gain and low-noise behaviour of a HEMT by changing the position of the gate electrode between source and drain. You observe the best performance if the gate is placed

a) in the middle between the contacts
b) closer to the drain contact
c) closer to the source contact

One or none of the statements is true – explain your choice!

(10) Draw the small-signal equivalent circuit of an FET for $V_{DS} = 0$, making appropriate simplifications. What is the relationship between C_{GS} and C_{GD} in this bias point? Would you expect a noticeable difference between MESFETs and HEMTs in this mode of operation?

(11) Explain the two major advantages of a pseudomorphic HEMT structure, compared to the classic AlGaAs/GaAs HEMT. How do they relate to f_T, f_{max} and F_{min}? Is there a potential disadvantage of the lower band gap in the channel?

(12) In order to reduce the series resistance of the gate, FETs (MESFETs, HEMTs and MOSFETs alike) are typically constructed with several gate fingers in parallel. Which effect(s) on device performance will result from this measure? Will this affect the transit frequency f_T? Explain your answer.

(13) A HEMT technology has $f_T = 80\,\text{GHz}$ and $g_m = 600\,\text{mS}\,\text{mm}^{-1}$. For a device with $W_G = 2 \cdot 60\,\text{m}$, the gate resistance is measured to be $R_G = 10\,\Omega$, and the minimum noise figure at 24 GHz is $F_{min} = 1.7\,\text{dB}$. For a device with $W_G = 6 \cdot 60\,\mu\text{m}$, what is the expected noise figure at 30 GHz? You may neglect the source and drain resistances.

(14) Only one of the following answers is correct: in a bipolar transistor in active forward operation, the base transit time is
   a) not a function of V_{CE}
   b) a weak function of V_{CE}
   c) a strong function of V_{CE}.
   Explain your choice!

(15) Explain the following observations on high-speed bipolar transistors:
   a) In devices optimised for record f_T, the maximum frequency of oscillation is often quite low, and the breakdown voltage is also low.
   b) Devices optimised for record f_{max} often have low f_T, higher breakdown voltage, and need relatively high V_{CE} for optimum operation.
   c) The latter devices suffer from significant Kirk effect.

(16) In n–p–n bipolar transistors, a drift field for electrons in the base can reduce the base transit time. This can be done in two ways:
   (a) introduce a continuously varying material composition;
   (b) vary the doping concentration in the base.
   Explain how to achieve a constant field strength in the neutral base using either of the two methods.

(17) Note that Si/SiGe HBTs are always double-heterostructure devices. Why?

References

[1] Allam R., Kolanowski C., Theron D., Crosnier Y. (1994). Large signal model for analysis and design of HEMT gate mixer. *IEEE Microwave Guided Wave Lett. MGWL-4*, 12 (December), 405–407.

[2] Antoniadis D. A., Aberg I., NiCléirigh C., Nayfeh O. M., Khakifirooz A., Hoyt J. L. (2006). Continuous MOSFET performance increase with device scaling: the role of strain and channel material innovations. *IBM J. Res. & Dev. 50*, 4/5 (April–May), 363–377.

[3] Belache A., Vanoverschelde A., Salmer G., Wolny M. (1991). Experimental analysis of HEMT behavior under low-temperature conditions. *IEEE Trans. Electron Devices ED-38*, 1 (January), 3–13.

[4] Benkhelifa F., Chertouk M., Dammann M., Massler M., Walther H., Weimann G. (2001). High performance metamorphic HEMT with 0.25 μm refractory metal gate on 4" GaAs substrate. In *International Conference on Semiconductor Manufacturing Technology GaAs MANTECH 2001 Digest of papers*. Las Vegas, NV: GaAs MANTECH, 230–233. `http://www.csmantech.org/Digests/2001/PDF/11_3_Benkhelifa_V2.pdf`.

[5] Bourgoin J., Mauger A. (1988). Physical origin of the DX center. *Appl. Phys. Lett. 53*, 9 (August), 749–751.

[6] Cappy A. (1988). Noise modeling and measurement techniques. *IEEE Trans. Microw. Theory Tech. MTT-36*, 1 (January), 1–10.

[7] Chan Y.-J., Pavlidis D., Razeghi M., Omnes F. (1990). $Ga_{.51}In_{.49}P$/GaAs HEMT's exhibiting good electrical performance at cryogenic temperatures. *IEEE Trans. Electron Devices ED-37*, 10 (October), 2141–2147.

[8] Cojocaru V. I., Brazil T. J. (1997). Scalable general-purpose model for microwave FET's including DC/AC dispersion effects. *IEEE Trans. Microw. Theory Tech. MTT-45*, 12 (December), 2248–2255.

[9] Cressler J. (2003). `http://extenv.jpl.nasa.gov/presentations/SiGe_HBT_BiCMOS.pdf`.

[10] Curtice W. (1980). A MESFET model for use in the design of GaAs integrated circuits. *IEEE Trans. Microw. Theory Tech. MTT-28*, 5 (May), 448–456.

[11] Das M. B. (1988). High-frequency performance limitations of millimeter-wave heterojunction bipolar transistors. *IEEE Trans. Electron Devices ED-35*, 5 (May), 604–614.

[12] Delagebeaudeuf D., Chevrier I., Laviron M., Delescluse P. (1985). A new relationship between the Fukui coefficient and optimal current value for low noise operation of field effect transistors. *IEEE Electron Device Lett. EDL-6*, 9 (September), 444–445.

[13] Dingle R. (1984). New high-speed III–V devices for integrated circuits. *IEEE Trans. Electron Devices ED-31*, 11 (November), 1662–1667.

[14] Dunn J. S., Ahlgren D. C., Coolbaugh D. D., *et al.* (2003). Foundation of RF CMOS and SiGe BiCMOS technologies. *IBM J. Res.&Dev. 47*, 2/3 (February/March), 101–138.

[15] Early J. M. (1952). Effects of space-charge layer widening in junction transistors. *Proc. IRE 40*, 11 (November), 1401–1406.

[16] Ebers J. J., Moll J. L. (1954). Large-signal behavior of junction transistors. *Proc. IRE 42*, 12 (December), 1761–1772.

[17] Johnson E. O. (1965). Physical limitations on frequency and power parameters of transistors. *RCA Rev. 26*, 6 (June), 163.

[18] Feygenson A., Ritter D., Hamm R. A., *et al.* (1992). InGaAs/InP composite collector heterostructure bipolar transistors. *Electron. Lett. 28*, 7 (March), 607–609.

[19] Fiegna C. (2003). Analysis of gate shot noise in MOSFETs with ultrathin gate oxides. *IEEE Electron Device Lett. EDL-24*, 2 (February), 108–110.

[20] Folkes P. A. (1985). Thermal noise measurements in GaAs MESFETs. *IEEE Electron Device Lett. EDL-6*, 12 (December), 620–622.

[21] Fukui H. (1979). Design of microwave GaAs MESFETs for broad-band low noise amplifiers. *IEEE Trans. Microw. Theory Tech. MTT-27*, 7 (July), 643–650.

[22] Gummel H. K., Poon H. C. (1970). An integral charge-control model for bipolar transistors. *Bell Syst. Tech. J. 49*, 827–852.

[23] Hawkins R. J. (1977). Limitations of Nielsen's and related noise equations applied to microwave bipolar transistors and a new expression for the frequency and current dependent noise figure. *Solid-State Electron. 20*, 3 (March), 191–196.

[24] Hooge F. N. (1969). 1/f noise is no surface effect. *Phys. Lett. A 29*, 3 (April), 139–140.

[25] Hsia H., Tang Z., Caruth D., Becher D., Feng M. (1999). Direct ion-implanted 0.12 μm GaAs MESFET with f_t of 121 GHz and $f_{\max}$ of 160 GHz. *IEEE Electron Device Lett. 20*, 5 (May), 245–247.

[26] Joseph A., Cressler J. D., Richey D. M., Jaeger R. C., Harame D. L. (1997). Neutral base recombination and its influence on the temperature dependence of Early voltage and current gain-Early voltage product in UHV/CVD SiGe heterojunction bipolar transistors. *IEEE Trans. Electron Devices ED-44*, 3 (March), 404–413.

[27] Kallfass I. (2005a). Comprehensive Nonlinear Modelling of Dispersive Heterostructure Field Effect Transistors and their MMIC Applications. Ph.D. thesis, Ulm Universität, Ulm, Germany.

[28] Kallfass I. (2005b). Comprehensive Nonlinear Modelling of Dispersive Heterostructure Field Effect Transistors and their MMIC Applications. Ph.D. thesis, Ulm Universität, Ulm, Germany. Chapter 3.4.1.

[29] Kallfass I., Schumacher H., Brazil T. J. (2006). A unified approach to charge-conservative capacitance modelling in HEMTs. *Microwave and Wireless Components Letters 16*, 12 (December), 678–680.

[30] Kloosterman W. J. (1996). Comparison of Mextram and the VBIC95 Bipolar Transistor Model. Tech. Rep. NL-UR 034/96, Philips Electronics N. V. `http://www.nxp.com/acrobat_download/other/philipsmodels/ur034_96.pdf`.

[31] Kroemer H. (1957). Theory of a wide-gap emitter for transistors. *Proc. IRE 45*, 11 (November), 1535–1537.

[32] Ladbrooke P. H. (1985). The theory and practice of the GaAs microwave MESFET. *GEC J. Res. 3*, 191–200.

[33] Lee K., Shur M., Drummond T., Morkoc H. (1984). Parasitic MESFET in (Al,Ga)As/GaAs modulation doped FET's and MODFET characterization. *IEEE Trans. Electron Devices ED-31*, 1 (January), 29–35.

[34] Lee T. H. (2004). *The design of CMOS Radio-Frequency Integrated Circuits*. Cambridge University Press.

[35] Lilienfeld J. E. (1930). Method and apparatus for controlling electric currents. USA Patent 1,745,175.

[36] Long S. I. (1989). A comparison of the GaAs MESFET and the AlGaAs/GaAs heterojunction bipolar transistor for power microwave amplification. *IEEE Trans. Electron Devices ED-36*, 5 (May), 1274–1279.

[37] McAndrew C. C., Seitchik J. A., Bowers D. F., *et al.* (1996). VBIC95, the vertical bipolar inter-company model. *IEEE J. Solid-State Circ. 31*, 10 (October), 1476–1483.

[38] Meyer J. E. (1971). MOS models and circuit simulations. *RCA Rev. 32*, 3 (March), 42–63.

[39] Mimura T., Hiyamizu S., Fujii T., Nanbu K. (1980). A new field-effect transistor with selectively doped GaAs/n-$Al_xGa_{1-x}As$ heterojunctions. *Jp. J. Appl. Phys. 19*, 5 (May), L225–L227.

[40] Nakajima S., Otobe K., Kuwata N., Shiga N., Matsuzaki K., Hayashi H. (1990). Pulse-doped GaAs MESFETs with planar self-aligned gate for MMIC. *IEEE MTT-S Int. Microwave Symp. Dig. 3*, 1081–1084.

[41] Ogura S., F. Codella C., Rovedo N., Shepard J. F., Riseman J. (1982). A half-micron MOSFET using double implanted LDD. In *Int'l Electron Devices Mtg. Proceedings*, Vol. 28. Piscataway, NJ: IEEE, 718–722.

[42] Paaschens J. C. J., v. d. Toorn R., Kloosterman W. J. (1995). The Mextram Bipolar Model. Tech. Rep. NL-UR 2000/811, Philips Electronics N. V. `http://www.nxp.com/acrobat_download/other/philipsmodels/nlur2000811.pdf`.

[43] Pailloncy G., Iniquez B., Dambrine G., Danneville F. (2004). Influence of a tunneling gate current on the noise performance of SOI MOSFETs. In *Proceedings 2004 IEEE International SOI Conference*. Piscataway, NJ: IEEE, 55–57.

[44] Parthasarathy N., Griffith Z., Kadow C. *et al.* (2006). Collector-pedestal InGaAs/InP DHBTs fabricated in a single-growth, triple-implant process. *IEEE Electron Device Lett. EDL-27*, 5 (May), 313–316.

[45] Post I., Akbar M., Curello G., *et al.* (2006). A 65 nm CMOS SOC technology featuring strained silicon transistors for RF applications. In *Int'l Electron Devices Mtg. Proceedings*. Piscataway, NJ: IEEE, 1–3.

[46] Pucel R. A., Haus H. A., Statz H. (1975). Signal and noise properties of gallium arsenide microwave field effect transistors. In *Advances in Electronics and Electron Physics*, L. Marton, ed. Vol. 38. Academic Press, 195–265.

[47] Rodwell M. J. W., Urteaga M., Mathew T., *et al.* (2001). Submicron scaling of HBTs. *IEEE Trans. Electron Devices ED-48*, 11 (November), 2606–2624.

[48] Saito M., Ono M., Fujimoto R., *et al.* (1998). 0.15 μm RF CMOS technology compatible with logic CMOS for low-voltage operation. *IEEE Trans. Electron Devices ED-45*, 3 (March), 737–742.

[49] Sakurai T., Newton A. R. (1991). A simple MOSFET model for circuit analysis. *IEEE Trans. Electron Devices ED-38*, 4 (April), 887–894.

[50] Sato H., Vlcek J. C., Fonstad C. G., Meskoob B., Prasad S. (1990). InGaAs/InAlAs/InP collector-up microwave heterojunction bipolar transistors. *IEEE Electron Device Lett. EDL-11*, 10 (October), 457–459.

[51] Schröter M. (2002). Staying current with HICUM. *IEEE Circ. Dev. Mag. 18*, 3 (May), 16–25.

[52] Schröter M. (2007). RF Modeling of Bipolar Transistors with HICUM. `http://www.iee.et.tu-dresden.de/~schroter/Conf/hic_ovw.pdf`.

[53] Schröter M., Friedrich M., Rein H.-M. (1993). A generalized integral charge-control relation and its application to compact models for silicon-based HBT's. *IEEE Trans. Electron Devices ED-40*, 11 (November), 2036–2046.

[54] Schröter M., Lee T. Y. (1999). Physics-based minority charge and transit time modeling for bipolar transistors. *IEEE Trans. Electron Devices ED-46*, 2 (February), 288–300.

[55] Schüppen A., Berntgen J., Maier P., Tortschanoff M., Kraus W., Averweg M. (2001). An 80 GHz SiGe production technology. *III–V Review 14*, 6 (August), 42–46.

[56] Shockley W. (1951). Circuit element utilizing semiconductive material. USA Patent 2,569,347.

[57] Shockley W. (1952). A unipolar 'field effect' transistor. *Proc. IRE 40*, 11, 1365–1376.

[58] Statz H., Newman P., Smith I. W., Pucel R. A., Haus H. A. (1987). GaAs FET device and circuit simulation in SPICE. *IEEE Trans. Electron Devices ED-34*, 2 (February), 160–169.

[59] Stillman G., Wolfe C., Dimmock J. (1970). Hall coefficient factor for polar mode scattering in N-type GaAs. *J. Phys. Chem. Solids 31*, 6, 1199–1204.

[60] Südow M., Andersson K., Billström N., *et al.* (2006). An SiC MESFET-based MMIC process. *IEEE Trans. Microw. Theory Tech. MTT-54*, 12 (December), 4072–4078.
[61] Sugli T., Watanabe K., Sugatani S. (2003). Transistor design for 90 nm-generation and beyond. *Fujitsu Sci. Tech. J. 39*, 6 (June), 9–22.
[62] van der Ziel A. (1962). Thermal noise in field effect transistors. *Proc. IRE 50*, 8 (August), 1808–1812.
[63] van der Ziel A. (1963). Gate noise in field effect transistors at moderately high frequencies. *Proc. IEEE 51*, 3 (March), 461–467.
[64] Yngvesson S. (1991). *Microwave Semiconductor Devices*. Kluwer Academic Publishers.
[65] Zeghbroeck B. V. (2004). Principles of Semiconductor Devices. `http://ece-www.colorado.edu/~bart/book/book/chapter5/pdf/ch5_5.pdf`.