Chapter 24
Transistors

> *Frequently, I have been asked if an experiment I have planned is pure or applied research; to me it is more important to know if the experiment will yield new and probably enduring knowledge about nature. If it is likely to yield such knowledge, it is, in my opinion, good fundamental research; and this is much more important than whether the motivation is purely esthetic satisfaction on the part of the experimenter on the one hand or the improvement of the stability of a high-power transistor on the other.*
>
> W.B. Shockley [1744, 1745]

Abstract The device functionalities of bipolar, heterobipolar and field effect transistors (JFET, MESFET and MOSFET) are explained. Within physical models for drift, diffusion and recombination given earlier in the book, the characteristics of these devices are derived. Remarks on integrated circuits, miniaturization and thin film transistors finish this chapter.

24.1 Introduction

Transistors[1] are the key elements for electronic circuits such as amplifiers, memories and microprocessors. Transistors can be realized in bipolar technology (bipolar junction transistor (BJT), Sect. 24.2) or as unipolar devices using the field effect (field-effect transistor (FET), Sect. 24.3) [500, 1746]. The equivalent in vacuum-tube technology to the transistor is the triode (Fig. 24.1a). Transistors can be optimized for their properties in analog circuits such as linearity and frequency response or their properties in digital circuits such as switching speed and power consumption. Transistors for microwave applications are discussed in [1526]. Early commercial models are shown in Fig. 24.2.

[1] The term ‘transistor’ was coined from the combination of ‘transconductance’ or ‘transfer’ and ‘varistor’ after initially such devices were termed ‘semiconductor triodes’. The major breakthrough was achieved in 1947 when the first transistor was realized that showed gain (Figs. 1.9 and 24.1b).

M. Grundmann, *The Physics of Semiconductors*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-23880-7_24

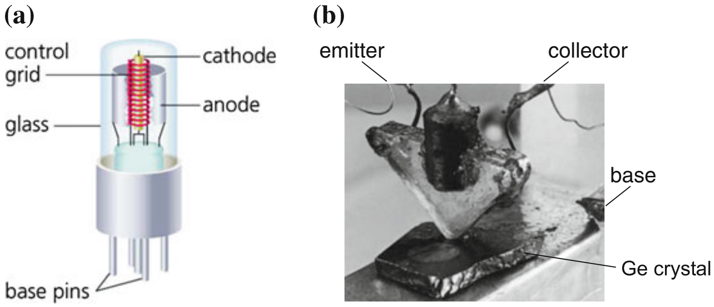


Fig. 24.1 (**a**) Schematic image of a vacuum triode. The electron current flows from the heated cathode to the anode when the latter is at a positive potential. The flow of electrons is controlled with the grid voltage. (**b**) Bell Laboratories' first (experimental) transistor, 1947

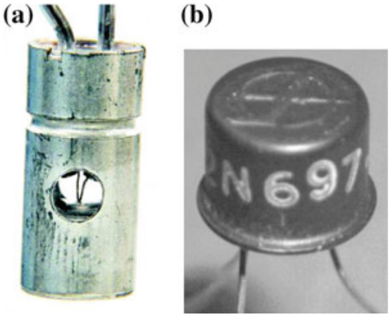


Fig. 24.2 (**a**) First commercial, developmental (point contact) transistor from BTL (Bell Telephone Laboratories) with access holes for adjustment of the whiskers pressing on a piece of Ge, diameter $7/32'' = 5\,\mathrm{mm}$, 1948. (**b**) First high-performance silicon transistor (npn mesa technology), model 2N697 from Fairchild Semiconductor, 1958 (at $200, in 1960 $28.50). The product number is still in use (now $0.95)

24.2 Bipolar Transistors

Bipolar transistors consist of a pnp or npn sequence (Fig. 24.3). The layers (or parts) are named emitter (highly doped), base (thin, highly doped) and collector (normal doping level). The transistor can be considered to consist of two diodes (emitter–base and base–collector) back to back. However, the important point is that the base is sufficiently thin (in relation to its minority carrier diffusion length) and carriers from the emitter (which are minority carriers in the base) can dominantly reach the collector by diffusion.

In Fig. 24.4, the three basic circuits with a transistor are shown. They are classified by the common contact for the input and output circuit. The space charges and band diagram for a pnp transistor in the base circuit configuration are depicted in Fig. 24.5. The emitter–base diode is switched in the forward direction to inject electrons into the

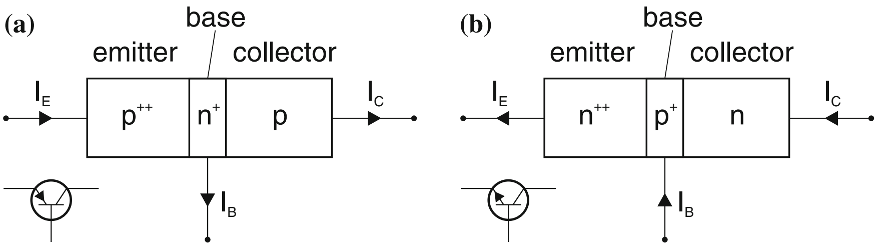


Fig. 24.3 Schematic structure and circuit symbol for (**a**) pnp and (**b**) npn transistors

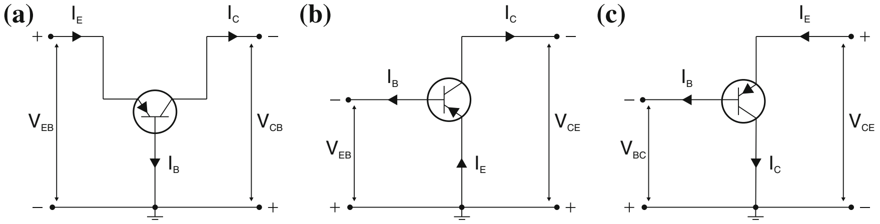
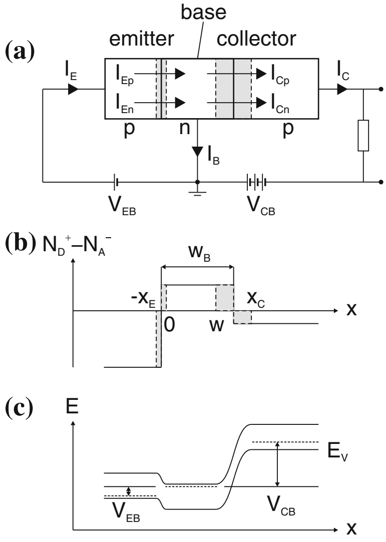


Fig. 24.4 Basic transistor circuits, named after the common contact: (**a**) Common base circuit, (**b**) common emitter circuit and (**c**) common collector circuit

Fig. 24.5 pnp transistor in (**a**) base circuit. (**b**) Doping profile and space charges (abrupt approximation) and (**c**) band diagram for typical operation conditions



base. The base–collector diode is switched in the reverse direction. The electrons that diffuse through the base and reach the neutral region of the collector are transported by the high drift field away from the base.

24.2.1 Carrier Density and Currents

The modeling of transistors is a complex topic. We treat the transistor on the level of the abrupt junction. As an approximation, we assume that all voltages drop at the junctions. Series resistances, capacities and stray capacities and other parasitic impedances are neglected at this point.

The major result is that the emitter–base current from the forward-biased emitter–base diode will be transferred to the collector. The current flowing from the base contact is small compared to the collector current. This explains the most prominent property of the transistor, the current amplification.

For the neutral part of the base region of a pnp transistor, the stationary equations for diffusion and continuity are

$$0 = D_\mathrm{B} \frac{\partial^2 p}{\partial x^2} - \frac{p - p_\mathrm{B}}{\tau_\mathrm{B}} \tag{24.1a}$$

$$j_\mathrm{p} = -e D_\mathrm{B} \frac{\partial p}{\partial x} \tag{24.1b}$$

$$j_\mathrm{tot} = j_\mathrm{n} + j_\mathrm{p}, \tag{24.1c}$$

where p_B is the equilibrium minority carrier density in the base. From the discussion of the pn-diode, we know that at the boundary of the depletion layer the minority carrier density is increased by $\exp(eV/kT)$ (cf. (21.98a, b)). At the boundaries of the emitter–base diode (for geometry see Fig. 24.5a)

$$\delta p(0) = p(0) - p_\mathrm{B} = p_\mathrm{B} \left[\exp\left(\beta V_\mathrm{EB}\right) - 1\right] \tag{24.2a}$$

$$\delta n(-x_\mathrm{E}) = n(-x_\mathrm{E}) - n_\mathrm{E} = n_\mathrm{E} \left[\exp\left(\beta V_\mathrm{EB}\right) - 1\right], \tag{24.2b}$$

where n_E and p_B are the equilibrium minority-carrier densities in the emitter and base, respectively. Accordingly, at the boundaries of the base–collector diode we have

$$\delta p(w) = p(w) - p_\mathrm{B} = p_\mathrm{B} \left[\exp\left(\beta V_\mathrm{CB}\right) - 1\right] \tag{24.3a}$$

$$\delta n(x_\mathrm{C}) = n(x_\mathrm{C}) - n_\mathrm{C} = n_\mathrm{C} \left[\exp\left(\beta V_\mathrm{CB}\right) - 1\right]. \tag{24.3b}$$

These are the boundary conditions for the diffusion equations in the p-doped layers and in the neutral region of the n-doped base. For the p-layers (with infinitely long contacts), the solution is (similar to (21.124)) for $x < -x_\mathrm{E}$ and $x > -x_\mathrm{C}$, respectively

$$n(x) = n_\mathrm{E} + \delta n(-x_\mathrm{E}) \exp\left(\frac{x + x_\mathrm{E}}{L_\mathrm{E}}\right) \tag{24.4a}$$

$$n(x) = n_\mathrm{C} + \delta n(x_\mathrm{C}) \exp\left(-\frac{x - x_\mathrm{C}}{L_\mathrm{C}}\right). \tag{24.4b}$$

L_E and L_C are the minority carrier (electron) diffusion lengths in the emitter and collector, respectively. The solution for the hole density in the neutral region in the base ($0 < x < w$) is

$$\begin{aligned} p(x) = p_\mathrm{B} &+ \left[\frac{\delta p(w) - \delta p(0)\exp\left(-w/L_\mathrm{B}\right)}{2\sinh\left(w/L_\mathrm{B}\right)}\right]\exp\left(\frac{x}{L_\mathrm{B}}\right) \\ &- \left[\frac{\delta p(w) - \delta p(0)\exp\left(w/L_\mathrm{B}\right)}{2\sinh\left(w/L_\mathrm{B}\right)}\right]\exp\left(-\frac{x}{L_\mathrm{B}}\right). \end{aligned} \tag{24.5}$$

We shall denote the excess hole density at $x = 0$ and $x = w$ as $\delta p_\mathrm{E} = \delta p(0)$ and $\delta p_\mathrm{C} = \delta p(w)$, respectively. Typical ('normal') operation condition in the common base circuit is that $\delta p_\mathrm{C} = 0$ (Fig. 24.8a). In the 'inverted' configuration, the role of emitter and collector are reversed and $\delta p_\mathrm{E} = 0$. We can write (24.5) also as

$$p(x) = p_\mathrm{B} + \delta p_\mathrm{E}\,\frac{\sinh\left[(w-x)/L_\mathrm{B})\right]}{\sinh\left[w/L_\mathrm{B}\right]} + \delta p_\mathrm{C}\,\frac{\sinh\left[x/L_\mathrm{B})\right]}{\sinh\left[w/L_\mathrm{B}\right]}. \tag{24.6}$$

If the base is thick, i.e. $w \to \infty$, or at least large compared to the diffusion length ($w/L_\mathrm{B} \gg 1$), the carrier concentration is given by

$$p(x) = p_\mathrm{B} + \delta p(0)\exp\left(-\frac{x}{L_\mathrm{B}}\right) \tag{24.7}$$

and does not depend on the collector. In this case there is no transistor effect. A 'coupling' between emitter and collector currents that are given by the derivative $\partial p/\partial x$ at 0 and w, respectively, is only present for a sufficiently thin base.

From (24.6), the hole current densities at $x = 0$ and $x = w$ are given as[2]

$$j_\mathrm{Ep} = j_\mathrm{p}(0) = e\frac{D_\mathrm{B}}{L_\mathrm{B}}\left[\delta p_\mathrm{E}\,\coth\left(\frac{w}{L_\mathrm{B}}\right) - \delta p_\mathrm{C}\,\mathrm{csch}\left(\frac{w}{L_\mathrm{B}}\right)\right] \tag{24.8a}$$

$$j_\mathrm{Cp} = j_\mathrm{p}(w) = e\frac{D_\mathrm{B}}{L_\mathrm{B}}\left[\delta p_\mathrm{E}\,\mathrm{csch}\left(\frac{w}{L_\mathrm{B}}\right) - \delta p_\mathrm{C}\,\coth\left(\frac{w}{L_\mathrm{B}}\right)\right]. \tag{24.8b}$$

From (24.4a, b), the electron current densities at $x = -x_\mathrm{E}$ and $x = x_\mathrm{C}$ are given (with $\delta n_\mathrm{E} = \delta n(-x_\mathrm{E})$ and $\delta n_\mathrm{C} = \delta n(x_\mathrm{C})$) by

$$j_\mathrm{En} = j_\mathrm{n}(-x_\mathrm{E}) = e\,\frac{D_\mathrm{E}}{L_\mathrm{E}}\,\delta n_\mathrm{E} \tag{24.9a}$$

$$j_\mathrm{Cn} = j_\mathrm{n}(x_\mathrm{C}) = -e\,\frac{D_\mathrm{C}}{L_\mathrm{C}}\,\delta n_\mathrm{C}. \tag{24.9b}$$

[2] $\coth x \equiv \cosh x/\sinh x$, $\mathrm{csch} x \equiv 1/\sinh x$.

The emitter current density is (similar to (21.127))

$$\begin{aligned} j_\mathrm{E} &= j_\mathrm{p}(0) + j_\mathrm{n}(-x_\mathrm{E}) \\ &= e\,\frac{D_\mathrm{B}}{L_\mathrm{B}}\left[\delta p_\mathrm{E}\,\coth\left(\frac{w}{L_\mathrm{B}}\right) - \delta p_\mathrm{C}\,\mathrm{csch}\left(\frac{w}{L_\mathrm{B}}\right)\right] + e\,\frac{D_\mathrm{E}}{L_\mathrm{E}}\,\delta n_\mathrm{E}. \end{aligned} \tag{24.10}$$

The collector current density is given as

$$\begin{aligned} j_\mathrm{C} &= j_\mathrm{p}(w) + j_\mathrm{n}(x_\mathrm{C}) \\ &= e\,\frac{D_\mathrm{B}}{L_\mathrm{B}}\left[\delta p_\mathrm{E}\,\mathrm{csch}\left(\frac{w}{L_\mathrm{B}}\right) - \delta p_\mathrm{C}\,\coth\left(\frac{w}{L_\mathrm{B}}\right)\right] - e\,\frac{D_\mathrm{C}}{L_\mathrm{C}}\,\delta n_\mathrm{C}. \end{aligned} \tag{24.11}$$

In these equations, only the diffusion currents are considered. Additionally, the recombination currents in the depletion layers must be considered, in particular at small junction voltages.

24.2.2 Current Amplification

The emitter current consists of two parts, the hole current I_pE injected from the base and the electron current I_nE that flows from the emitter to the base (Fig. 24.5a). Similarly, the collector current is made up from the hole and electron currents I_pC and I_pC, respectively.

The total emitter current splits into the base and collector currents

$$I_\mathrm{E} = I_\mathrm{B} + I_\mathrm{C}. \tag{24.12}$$

The amplification (gain) in common base circuits

$$\alpha_0 = h_\mathrm{FB} = \frac{\partial I_\mathrm{C}}{\partial I_\mathrm{E}} = \frac{\partial I_\mathrm{pE}}{\partial I_\mathrm{E}}\,\frac{\partial I_\mathrm{pC}}{\partial I_\mathrm{pE}}\,\frac{\partial I_\mathrm{C}}{\partial I_\mathrm{pC}} = \gamma\,\alpha_\mathrm{T}\,M, \tag{24.13}$$

where γ is the emitter efficiency, α_T the base transport factor and M the collector multiplication factor. Since the collector is normally operated below the threshold for avalanche multiplication, $M = 1$.

The current amplification in the common emitter circuit is

$$\beta_0 = h_\mathrm{FE} = \frac{\partial I_\mathrm{C}}{\partial I_\mathrm{B}}. \tag{24.14}$$

Using (24.12), we find

$$\beta_0 = \frac{\partial I_\mathrm{E}}{\partial I_\mathrm{B}} - 1 = \frac{\partial I_\mathrm{E}}{\partial I_\mathrm{C}}\,\frac{\partial I_\mathrm{C}}{\partial I_\mathrm{B}} - 1 = \frac{1}{\alpha_0}\,\beta_0 - 1 = \frac{\alpha_0}{1-\alpha_0}. \tag{24.15}$$

Since α_0 is close to 1 for a well-designed transistor, β_0 is a large number, e.g. $\beta_0 = 99$ for $\alpha_0 = 0.99$.

The emitter efficiency is (A denotes the device area)

$$\gamma = \frac{A j_{\mathrm{Ep}}}{I_{\mathrm{E}}} = \left[1 + \frac{n_{\mathrm{E}}}{p_{\mathrm{B}}} \frac{D_{\mathrm{E}}}{D_{\mathrm{B}}} \frac{L_{\mathrm{B}}}{L_{\mathrm{E}}} \tanh\left(\frac{w}{L_{\mathrm{B}}}\right)\right]^{-1} . \tag{24.16}$$

The base transport factor, i.e. the ratio of minority carriers reaching the collector and the total number of injected minority carriers, is (for reverse bias $|\beta U_{\mathrm{CB}}| \gg kT$)

$$\begin{aligned} \alpha_{\mathrm{T}} = \frac{j_{\mathrm{Cp}}}{j_{\mathrm{Ep}}} &= \frac{\exp(\beta U_{\mathrm{EB}}) - 1 + \cosh w/L_{\mathrm{B}}}{1 + (\exp(\beta U_{\mathrm{EB}}) - 1)\, \cosh w/L_{\mathrm{B}}} \\ &\approx \frac{1}{\cosh\,(w/L_{\mathrm{B}})} \approx 1 - \frac{w^2}{2L_{\mathrm{B}}^2} . \end{aligned} \tag{24.17}$$

The first approximation is for $\beta U_{\mathrm{EB}} \gg 1$ (emitter diode injecting in forward direction), the second approximation is for $w \ll L_{\mathrm{B}}$. If the base length is a tenth of the diffusion length, the base transport factor is $\alpha_{\mathrm{T}} > 0.995$. M is also very close to 1; for reverse bias U_{CB} and $w \ll L_{\mathrm{B}}$ we find

$$M \approx 1 + \frac{w}{L_{\mathrm{C}}} \frac{D_{\mathrm{C}}}{D_{\mathrm{B}}} \frac{\delta n_{\mathrm{C}}}{\delta p_{\mathrm{C}} - \delta p_{\mathrm{E}}} \approx 1 + \frac{w}{L_{\mathrm{C}}} \frac{D_{\mathrm{C}}}{D_{\mathrm{B}}} \frac{n_{\mathrm{C}}}{p_{\mathrm{B}}} \exp(-\beta U_{\mathrm{EB}}) . \tag{24.18}$$

Thus for $w \ll L_{\mathrm{B}}$, α_0 is dominated by γ and given as (approximating (24.16))

$$\alpha_0 \approx \gamma \approx 1 - \frac{w}{L_{\mathrm{E}}} \frac{n_{\mathrm{E}}}{p_{\mathrm{B}}} \frac{D_{\mathrm{E}}}{D_{\mathrm{B}}} . \tag{24.19}$$

The gain β_0 is then also determined by γ:

$$\beta_0 = h_{\mathrm{FE}} \approx \frac{\gamma}{1 - \gamma} \approx \frac{1}{1 - \gamma} \propto \frac{N_{\mathrm{E}}}{N_{\mathrm{B}}} \frac{L_{\mathrm{E}}}{w} , \tag{24.20}$$

N_{E} and N_{B} being the doping levels in the emitter and base, respectively. The base and collector current are shown in Fig. 24.6 as a function of the emitter–base voltage, i.e. the voltage at the injection diode. The collector current is close to the emitter–base diode current and displays a dependence $\propto \exp(eV_{\mathrm{EB}}/kT)$. The base current shows a similar slope but is orders of magnitude smaller in amplitude. For small forward voltages of the emitter–base diode, the current is typically dominated by a nonradiative recombination current that flows through the base contact and has an ideality factor (m in Fig. 24.6) close to 2.

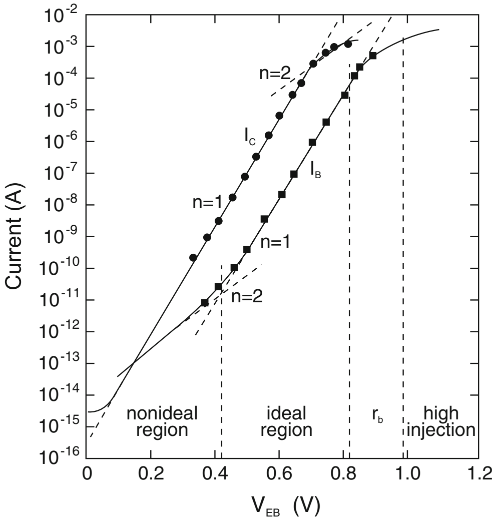


Fig. 24.6 Collector current I_C and base current I_B as a function of the emitter–base voltage V_{EB} (Gummel plot). Adapted from [1747]

24.2.3 Ebers–Moll Model

The Ebers–Moll model (Fig. 24.7) was developed in 1954 and is a relatively simple transistor model that needs, at its simplest level (Fig. 24.7a) just three parameters. It can (and must) be refined (Fig. 24.7b, c). The model considers two ideal diodes ('F' (forward) and 'R' (reverse)) back to back, each feeding a current source. The F diode represents the emitter–base diode and the R diode the collector–base diode. The currents are

$$I_F = I_{F0} \left[\exp(\beta V_{EB}) - 1\right] \tag{24.21a}$$

$$I_R = I_{R0} \left[\exp(\beta V_{CB}) - 1\right]. \tag{24.21b}$$

Using (24.8a, b)–(24.11), the emitter and collector currents are

$$I_E = \hat{a}_{11} \left[\exp(\beta V_{EB}) - 1\right] + \hat{a}_{12} \left[\exp(\beta V_{CB}) - 1\right] \tag{24.22a}$$

$$I_C = \hat{a}_{21} \left[\exp(\beta V_{EB}) - 1\right] + \hat{a}_{22} \left[\exp(\beta V_{CB}) - 1\right], \tag{24.22b}$$

with

$$\hat{a}_{11} = eA \left[p_B \frac{D_B}{L_B} \coth\left(\frac{w}{L_B}\right) + n_E \frac{D_E}{L_E} \right] \tag{24.23a}$$

$$\hat{a}_{12} = -eA p_B \frac{D_B}{L_B} \operatorname{csch}\left(\frac{w}{L_B}\right) \tag{24.23b}$$

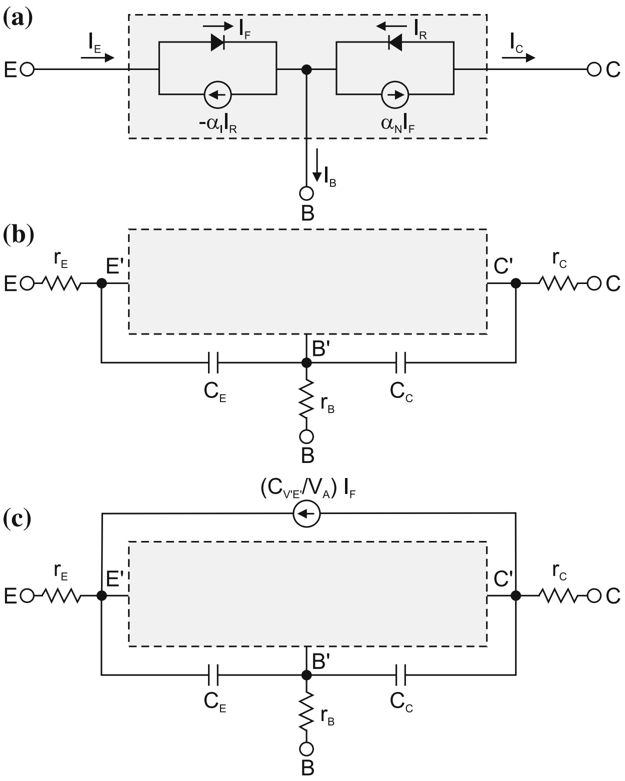


Fig. 24.7 Ebers–Moll model of a transistor, 'E': emitter, 'C': collector and 'B': base. Currents are shown for a pnp transistor. (**a**) Basic model (*grey area* in (**b**, **c**)), (**b**) model with series resistances and depletion-layer capacitances, (**c**) model additionally including the Early effect (V_A: Early voltage)

$$\hat{a}_{21} = eAp_B \frac{D_B}{L_B} \operatorname{csch}\left(\frac{w}{L_B}\right) = -\hat{a}_{12} \tag{24.23c}$$

$$\hat{a}_{22} = -eA\left[p_B \frac{D_B}{L_B} \coth\left(\frac{w}{L_B}\right) + n_C \frac{D_C}{L_C}\right]. \tag{24.23d}$$

The currents at the three contacts are

$$I_E = I_F - \alpha_I I_R \tag{24.24a}$$

$$I_C = \alpha_N I_F - I_R \tag{24.24b}$$

$$I_B = (1 - \alpha_N) I_F + (1 - \alpha_I) I_R. \tag{24.24c}$$

The last equation is obtained from (24.24a, b) using (24.12). By comparison with (24.21a, b) and (24.23a–d) we find

$$I_{\mathrm{F0}} = \hat{a}_{11} \tag{24.25a}$$

$$I_{\mathrm{R0}} = -\hat{a}_{22} \tag{24.25b}$$

$$\alpha_{\mathrm{I}} = \hat{a}_{12}/I_{\mathrm{R0}} \tag{24.25c}$$

$$\alpha_{\mathrm{N}} = \hat{a}_{21}/I_{\mathrm{F0}} = -\hat{a}_{12}/I_{\mathrm{F0}} = -\alpha_{\mathrm{I}} I_{\mathrm{R0}}/I_{\mathrm{F0}}. \tag{24.25d}$$

The constants α_{N} and α_{I} are the forward ('normal') ($\alpha_{\mathrm{N}} = \alpha_0$ from (24.13)) and reverse ('inverted') gains in the common base circuit, respectively. Both constants are larger than zero. Typically, $\alpha_{\mathrm{N}} \approx 0.98 \ldots 0.998 \lessapprox 1$ and $\alpha_{\mathrm{I}} \approx 0.5 \ldots 0.9 < \alpha_{\mathrm{N}}$. The model has three independent parameters, e.g. α_{N}, I_{F0} and I_{R0}. Equation (24.24a, b) can be rewritten as

$$I_{\mathrm{E}} = \alpha_{\mathrm{I}}\, I_{\mathrm{C}} + (1 - \alpha_{\mathrm{I}}\, \alpha_{\mathrm{N}})\, I_{\mathrm{F}} \tag{24.26a}$$

$$I_{\mathrm{C}} = \alpha_{\mathrm{N}}\, I_{\mathrm{E}} - (1 - \alpha_{\mathrm{I}}\, \alpha_{\mathrm{N}})\, I_{\mathrm{R}}. \tag{24.26b}$$

Under normal operation we have

$$I_{\mathrm{E}} = I_{\mathrm{F}} \tag{24.27a}$$

$$I_{\mathrm{C}} = \alpha_{\mathrm{N}}\, I_{\mathrm{E}}. \tag{24.27b}$$

The model can be refined and made more realistic by including the effect of series resistances and depletion-layer capacitances, increasing the number of parameters to eight. The Early effect (see p. 794) can be included by adding a further current source. This level is the 'standard' Ebers–Moll model with a total of nine parameters. Further parameters can be added. However, as is always the case with simulations, there is a tradeoff between the simplicity of the model and to what detail a real situation is approximated.

24.2.4 Current–Voltage Characteristics

In Fig. 24.8, the hole density in the base (of a pnp transistor) is shown for various voltage conditions. In Fig. 24.9, the I–V characteristics of a bipolar transistor in common base and common collector circuit are shown. In the common base circuit (Fig. 24.9a), the collector current is practically equal to the emitter current and is almost independent of the collector–base voltage. From (24.26b), the dependence of the collector current on the collector–base voltage is given (within the Ebers–Moll model) as

$$I_{\mathrm{C}} = \alpha_{\mathrm{N}} I_{\mathrm{E}} - (1 - \alpha_{\mathrm{I}}\alpha_{\mathrm{N}}) I_{\mathrm{R0}} \left[\exp\left(\beta V_{\mathrm{CB}}\right) - 1\right]. \tag{24.28}$$

V_{CB} is in the reverse direction. Therefore, the second term is zero for normal operating conditions. Since $\alpha_{\mathrm{N}} \lessapprox 1$, the collector current is almost equal to the emitter current.

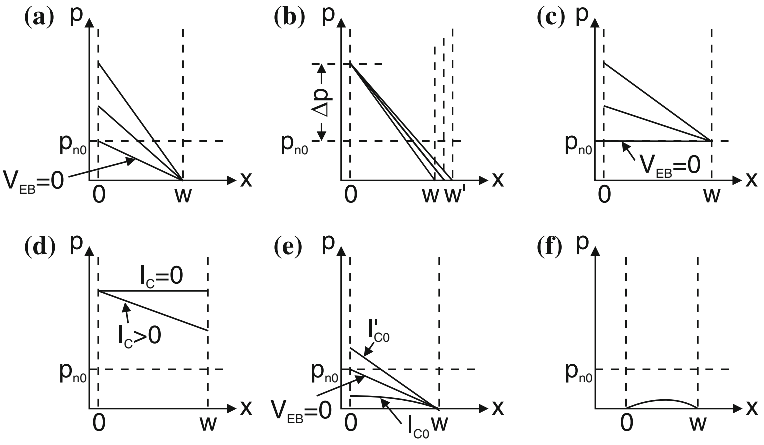


Fig. 24.8 Hole density (linear scale) in the base region (the neutral part of the base ranges from 0 to w) of a pnp transistor for various voltages. (**a**) normal voltages, $V_{\rm CB}$ = const. and various $V_{\rm EB}$ (in forward direction). (**b**) $V_{\rm EB}$ = const. and various values of $V_{\rm CB}$. (**c**) Various values of $V_{\rm EB} > 0$, $V_{\rm CB} = 0$. (**d**) Both pn-junctions in forward direction. (**e**) Conditions for $I_{\rm C0}$ and $I'_{\rm C0}$. (**f**) Both junctions in reverse direction. Adapted from [500]

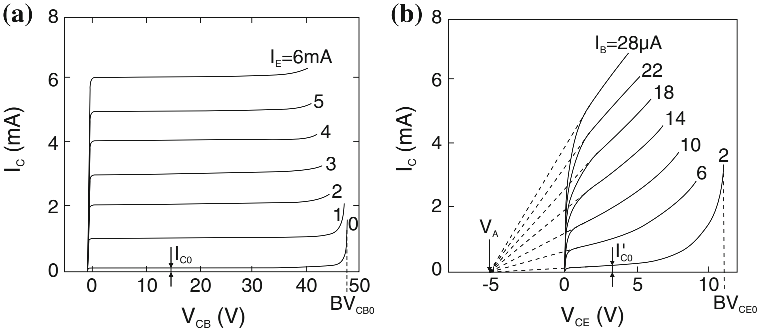


Fig. 24.9 Characteristics ($I_{\rm C}$ vs. $V_{\rm CB}$) of a pnp transistor in (**a**) common base (CB) circuit (Fig. 24.4a) for various values of the emitter current as labeled. Adapted from [1748]. (**b**) Characteristics in common emitter (CE) circuit (Fig. 24.4b). Adapted from [1749]

Even at $V_{\rm CB} = 0$ (the case of (Fig. 24.8c), holes are extracted from the base since $\partial p/\partial x|_{x=w} > 0$. A small forward voltage must be applied to the collector–base diode in order to make the current zero, i.e. $\partial p/\partial x|_{x=w} = 0$ (Fig. 24.8d). The collector saturation current $I_{\rm C0}$ is measured with an open emitter side. This current is smaller than the saturation current of the CB diode, since at the emitter side of the basis a vanishing gradient of the hole density is present (Fig. 24.8e). This reduces the gradient (and thus the current) at the collector side. The current $I_{\rm C0}$ is therefore smaller than the collector current for shorted emitter–base contact ($V_{\rm EB} = 0$). At high collector voltage, the current increases rapidly at $BV_{\rm CB0}$ due to breakdown of the collector–base diode. It can also occur that the width of the neutral base region

w becomes zero (punch-through). In this case, the emitter and collector are short-circuited.

In the common emitter circuit (Fig. 24.9b), there is a high current amplification $I_\mathrm{C}/I_\mathrm{B}$. Note that the collector current is given in mA and the base current in µA. The current increases with increasing V_CE because the base width w decreases and β_0 increases. There is no saturation of the I–V characteristics (Early effect [1750]). Instead, the I–V curves look as if they start at a negative collector–emitter voltage, the so-called Early voltage V_A. In the linear regime, the characteristic can be approximated by

$$I_\mathrm{C} = \left(1 + \frac{V_\mathrm{CE}}{V_\mathrm{A}}\right) \beta_0 \, I_\mathrm{B} . \tag{24.29}$$

Here, β_0 is the current gain for $V_\mathrm{CE} \approx 0$.

The physical reason for the increase of the collector current with increasing V_CE is the increasing reverse voltage at the collector–base diode that causes the so-called 'base-width modulation', as shown in Fig. 24.8b. The expansion of the CB depletion layer leads subsequently to a reduction of the neutral base width w. w will be smaller and smaller compared to the geometrical base width w_B. When w is reduced, the common base gain α_0 (24.19) becomes closer to 1 and the current gain increases. Therefore, the collector current increases with V_CE for a given (fixed) base current. The Early voltage is the coefficient of the increase of collector current with V_CE,

$$\frac{\beta_0 \, I_\mathrm{B}}{V_\mathrm{A}} = \frac{\partial I_\mathrm{C}}{\partial V_\mathrm{CE}} = \frac{\partial I_\mathrm{C}}{\partial V_\mathrm{CB}} \frac{\partial V_\mathrm{CB}}{\partial V_\mathrm{CE}} \approx \frac{\partial I_\mathrm{C}}{\partial V_\mathrm{CB}} . \tag{24.30}$$

For constant base current, the emitter-base voltage is almost constant and the approximation in (24.30) holds. The dependence of the CB depletion layer width on the base side x_C^n on U_CB is given by (21.106a) for a pnp-transistor. Typical values for the Early voltage are 50–300 V. The modeling of the Early effect in the SPICE simulation program is discussed in [1751].

For small collector–emitter voltage, the current quickly drops to zero. V_CE is typically split in such a way that the emitter–base diode is well biased forward and the CB diode has a high reverse voltage. If V_CE drops below a certain value (≈1 V for silicon transistors), there is no longer any bias at the CB diode. A further reduction of V_CE biases the CB diode in the forward direction and quickly brings the collector current down to zero.

24.2.5 Basic Circuits

Common Base Circuit

In the common base configuration, there is no current amplification since the currents flowing through emitter and base are almost the same. However, there is voltage gain since the collector current causes a large voltage drop across the load resistor.

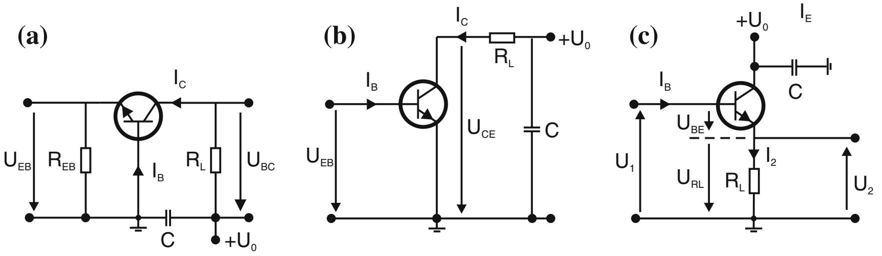


Fig. 24.10 (**a**) Common base, (**b**) common emitter and (**c**) common collector circuits with external loads

Common Emitter Circuit

The input resistance of the common emitter circuit (Fig. 24.10a) depends on the emitter–base diode and varies between a value of the order of 100 kΩ at small current and a few Ω at larger current and high V_{EB}. The voltage gain is

$$r_{\mathrm{V}} = \frac{V_{\mathrm{CE}}}{V_{\mathrm{EB}}} = \frac{I_{\mathrm{C}}}{V_{\mathrm{EB}}} R_{\mathrm{L}}, \tag{24.31}$$

where R_{L} is the load resistance in the output circuit (see Fig. 24.4b). The ratio $g_{\mathrm{m}} = I_{\mathrm{C}}/V_{\mathrm{EB}}$ is called the forward transconductance. Also, the differential transconductance $g'_{\mathrm{m}} = \partial I_{\mathrm{C}}/\partial V_{\mathrm{EB}}$ is used. The voltage gain of the common emitter circuit is typically 10^2–10^3. Since current *and* voltage are amplified, this circuit has the highest power gain.

If the input voltage V_{EB} (U_1 in Fig. 24.10a) is increased, the collector current rises. This increase causes an increase of the voltage drop across the load resistance R_{L} and a decrease of the output voltage U_2. Therefore, the phase of the input signal is reversed and the amplifier is inverting.

Common Collector Circuit

In Fig. 24.10c, the collector is connected to mass for alternating currents. Input and output current flow through the load resistance at which part of the input voltage drops. The input voltage is divided between the load resistor R_{L} and the emitter–base diode. At the transistor, the voltage $V_{\mathrm{BE}} = V_1 - V_{\mathrm{RL}}$ is applied. If the input voltage is increased, I_2 increases. This leads to a larger voltage drop at the load resistor and therefore to a decrease of V_{BE}, working against the original increase. The input resistance R_1 is large despite a small load resistance, $R_1 \approx \beta_0 R_{\mathrm{L}}$. The input voltage is larger than V_{RL}, thus no voltage gain occurs (actually it is a little smaller than 1). The current amplification is $(\beta + 1)$. The output resistance R_2 is small, $R_2 = U_2/I_2 = R_{\mathrm{L}} \approx R_1/\beta_0$. Therefore, this circuit is also called an impedance amplifier that allows high-impedance sources to be connected low-impedance loads.

Since an increase of the input voltage leads to an increase of the output voltage that is present at the emitter, this circuit is a direct amplifier and is also called an emitter follower.

24.2.6 High-Frequency Properties

Transistors for amplification of high-frequency signals are typically chosen as npn transistors since electrons, the minority carriers in the base, have higher mobility than holes. The active area and parasitic capacitance must be minimized. The emitter is formed in the shape of a stripe, nowadays in the 100 nm regime. The base width is in the 10 nm range. High p-doping of GaAs with low diffusion of the dopant is accomplished with carbon. Defects that would short emitter and collector at such thin base width must be avoided.

An important figure of merit is the cutoff frequency f_T for which h_FE is unity in the common emitter configuration. The cutoff frequency is related to the emitter–collector delay time τ_EC by

$$f_\mathrm{T} = \frac{1}{2\pi\,\tau_\mathrm{EC}}. \tag{24.32}$$

The delay time is determined by the charging time of the emitter–base depletion layer, the base capacitance, and the transport through the base–collector depletion layer. It is favorable if all times are short and similar. It does not help to minimize only one or two of the three processes since the longest time determines the transistor performance.

Another important figure of merit is the maximum frequency with which the transistor can oscillate in a feedback circuit with zero loss. This frequency is denoted by f_max. Approximately,

$$f_\mathrm{max} \simeq \sqrt{\frac{f_\mathrm{T}}{8\pi\,R_\mathrm{B}\,C_\mathrm{CB}}}, \tag{24.33}$$

where R_B is the base resistance and C_CB is the collector–base capacitance. f_max is larger than f_T, by a factor of the order of three.

24.2.7 Heterojunction Bipolar Transistors

In a heterojunction bipolar transistor (HBT), the emitter–base diode is formed with a heterostructure diode. The desired functionality is obtained when the emitter is made from the higher-bandgap material and the base from the lower-bandgap material. The

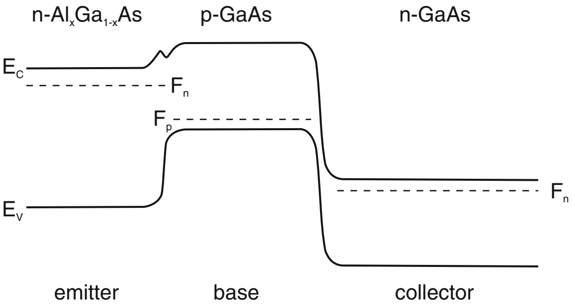


Fig. 24.11 Schematic band diagram of a heterojunction bipolar transistor

schematic band diagram is shown in Fig. 24.11 (see Fig. 21.57c for the emitter–base diode).

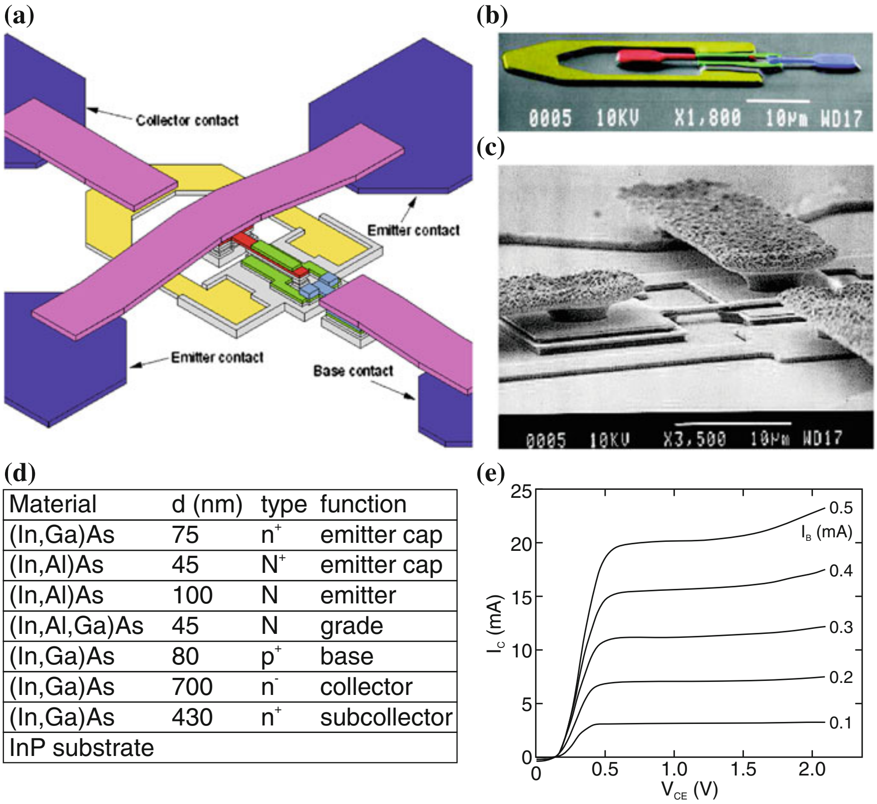
The higher discontinuity in the valence band, compared to a homojunction with the base material, provides a higher barrier for hole transport from the base to the emitter. Thus, the emitter efficiency is increased. Another advantage is the possibility for higher doping of the base without loss of emitter efficiency. This reduces the base series resistance and leads to better high-frequency behavior due to higher current gain and a smaller RC time constant. Also, operation at higher temperature is possible when the emitter has a larger band gap. Current InP/InGaAs-based HBTs have cutoff frequencies beyond 30 GHz, SiGe-HBTs beyond 80 GHz. The high-frequency performance is influenced by the velocity-overshoot effect (cf. Sect. 8.4.3) [1752].

In Fig. 24.12, an InAlAs/InGaAs HBT is shown [1754]. The cutoff frequency is 90 GHz. For the layer design, a fairly thick collector with low doping was chosen. This design allows a broad depletion layer with fairly small maximum electric field and thus a high breakdown voltage of $BV_{\mathrm{CE0}} > 8.5\,\mathrm{V}$. The base is not too thin (80 instead of maybe 60 nm) to reduce the series resistance. A graded region between emitter and base was chosen to avoid a spike occurring in the conduction band (Fig. 21.57b) and keep the turn-on voltage low.

24.2.8 Light-Emitting Transistors

The base current has two components. One is the recombination current in the neutral region of the emitter; this current can be suppressed in the HBT. The other is the recombination in the base region itself.[3] If quantum wells are introduced into the

[3] Also, a recombination current in the emitter–base depletion region is possible. However, since in normal operating conditions this diode is forward biased, the depletion layer is short and the associated recombination current is small, cf. p. 638.



Material	d (nm)	type	function
(In,Ga)As	75	n^+	emitter cap
(In,Al)As	45	N^+	emitter cap
(In,Al)As	100	N	emitter
(In,Al,Ga)As	45	N	grade
(In,Ga)As	80	p^+	base
(In,Ga)As	700	n^-	collector
(In,Ga)As	430	n^+	subcollector
InP substrate			

Fig. 24.12 (**a**) Schematic layout of a high-frequency HBT and SEM images (**b**) without and (**c**) with contacts. (**d**) Epitaxial layer sequence and (**e**) static performance data. Parts (**a**, **b**) from [1753], parts (**d**, **e**) from [1754]

base region, this recombination can occur radiatively between electrons and holes captured into the quantum well (Fig. 24.13). The spectrum exhibits two peaks from the QWs and the GaAs barrier.

24.3 Field-Effect Transistors

Next to the bipolar transistors, the field-effect transistors (FET) are another large class of transistors. FETs were conceptualized first but due to technological difficulties with semiconductor surfaces, realized second. The principle is fairly simple: A current flows through a channel from source to drain. The current is varied via the channel conductivity upon the change of the gate voltage. The gate needs to make a nonohmic contact to the semiconductor. Since the conductivity in the channel is a

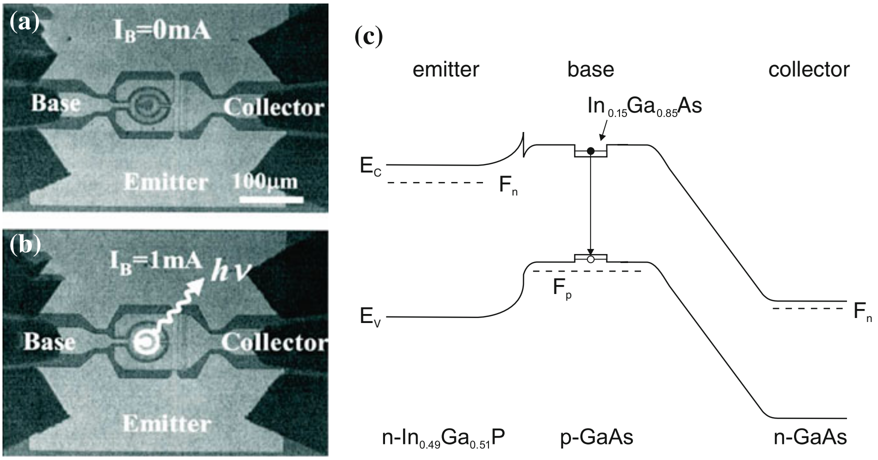


Fig. 24.13 Microscopic image of an InGaP/GaAs HBT with two 5-nm InGaAs/GaAs QWs in the 30-nm wide base at (**a**) zero base current and (**b**) at 1 mA base current in the common emitter configuration with Si CCD image of light emission. (**c**) Schematic band diagram of a HBT with a single InGaAs/GaAs quantum well in the base. Parts (**a**, **b**) from [1755], part (**c**) adapted from [1756]

property related to the majority charge carriers, FETs are called unipolar transistors. FETs feature a higher input impedance than bipolar transistors, a good linearity, and a negative temperature coefficient and thus a more homogeneous temperature distribution. According to the structure of the gate diode we distinguish JFETs, MESFETs and MOSFETs, as discussed in the following.

In the junction FET (JFET), the variation of channel conductivity is accomplished via the extension of the depletion layer of the pn-junction formed by the gate and the channel material (Fig. 24.14a). The JFET was analyzed by Schottky in 1952 [98] and realized by Dacey and Ross in 1953 [99].

In a MESFET, a metal–semiconductor diode (Schottky diode) is used as rectifying contact instead of a pn-diode. Otherwise, the principle is the same as that of the JFET. After the proposal by Mead in 1966 [123], the first (epitaxial) GaAs MESFET was realized by Hooper and Lehrer in 1967 [126]. The MESFET offers some advantages, such as the fabrication of the metal gate at lower temperature than necessary for the (diffusion or epitaxy of the) pn-diode, lower resistance, good thermal contact. The JFET can be made with a heterostructure gate to improve the frequency response.

In a MISFET, the gate diode is a metal–insulator–semiconductor diode (Fig. 24.14b). If the insulator is an oxide, the related FET is a MOSFET. When the gate is put at a positive voltage (for a p-channel), an inversion layer is formed close to the insulator–semiconductor interface. This layer is an n-conducting channel allowing conduction between the two oppositely biased pn-diodes. It can carry a high current. The MOSFET was theoretically envisioned early by Lilienfeld in 1925 [48] and realized only in 1960 by Kahng and Atalla [111].

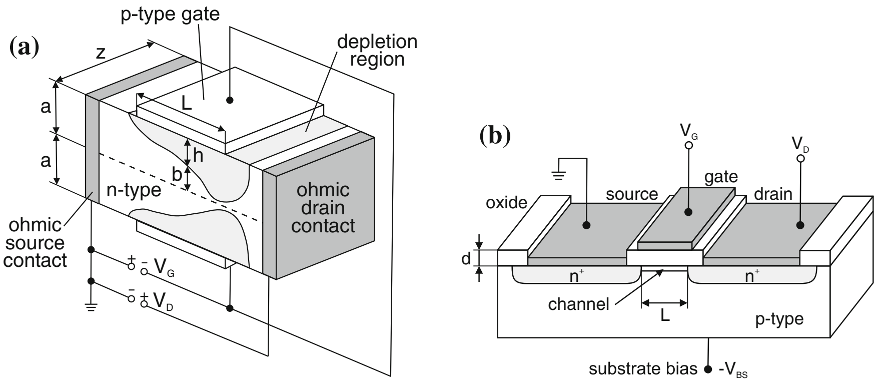


Fig. 24.14 (**a**) Shockley's model of a JFET. The *dashed line* represents the middle of the symmetric channel of total thickness $2a$. The *light grey area* is the depletion layer with thickness h. The gate length is L. The *dark grey* areas are ohmic metal contacts. Based on [99]. (**b**) Scheme of a MOSFET with channel length L and oxide thickness d. The *dark grey* areas are ohmic metal contacts. Adapted from [500]

FETs come in 'n' and 'p' flavors, depending on the conductivity type of the channel. For high-frequency applications, typically an n-channel is used due to the higher mobility or drift velocity. In CMOS (complementary MOS) technology, both n-FETs and p-FETs are integrated in high density, allowing the effective realization of logic gates with minimized power consumption.

24.4 JFET and MESFET

24.4.1 General Principle

The principal characteristic of a JFET is shown in Fig. 24.15. At $V_D = 0$ and $V_G = 0$, the transistor is in thermodynamic equilibrium and there are no net currents. Underneath the gate diode, a depletion layer is present. If for zero gate voltage the source–drain voltage is applied to the channel, the current increases linearly. The positive voltage at the drain contact causes the expansion of the depletion layer of the (reversely biased) gate–drain pn-diode. When the two (the upper and the lower) depletion regions meet (pinch-off), the current saturates. The respective source–drain voltage is denoted as $V_{D,sat}$. For high gate–drain (reverse) voltage V_D, breakdown occurs with a strong increase of the source–drain current. A variation of the gate voltage V_G leads to a variation of the source–drain current. A reverse voltage leads to a reduction of the saturation current and saturation at lower source–drain voltage. For a certain gate voltage V_P, the pinch-off voltage, no current can flow in the channel any longer since pinch-off exists even for $V_D = 0$.

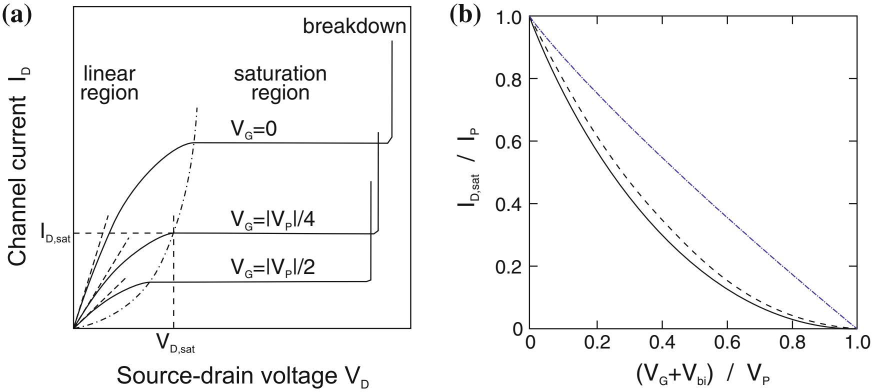


Fig. 24.15 (**a**) Principal characteristics of a JFET. The channel current I_D is shown as a function of the source–drain voltage V_D for three different values of the (absolute value of the) gate voltage V_G. The saturation values $V_{D,sat}$ and $I_{D,sat}$ are indicated for one curve. The intersections with the *dash-dotted* line yield the saturation voltage. Adapted from [500]. (**b**) Transfer behavior of a JFET for two different carrier distributions, homogeneous (*solid* line) and δ-like (*dashed* line). The *blue, dash-dotted* line is $\sqrt{I_D/I_P}$ versus the gate voltage. After [1756, 1757]

24.4.2 Static Characteristics

Here, we will calculate the general static behavior outlined in the previous section. We assume a long channel ($L \gg a$), the abrupt approximation for the depletion layer, the gradual channel approximation, i.e. the depletion layer depth changes slowly along x, and a field-independent, constant mobility. In this case, the two-dimensional Poisson equation for the potential distribution V can be used by solving it along the y direction (channel depth) for all x-positions (adiabatic approximation),

$$\frac{\partial^2 V}{\partial y^2} = -\frac{\rho(y)}{\epsilon_s}. \tag{24.34}$$

The geometry is shown in the inset of Fig. 24.15b.

The depth h of the depletion layer in the abrupt approximation is given by (cf. (21.107), reverse voltages are counted as positive here)

$$h = \sqrt{\frac{2\epsilon_s}{eN_D}\left(V_{bi} + V_G + V(x)\right)}. \tag{24.35}$$

Here, we have assumed homogeneous doping, i.e. N_D does not depend on y (or x). The built-in voltage (for a p$^+$n gate diode) is given by $V_{bi} = \beta^{-1} \ln\left(N_D/n_i\right)$ (21.97a). The voltage V is the applied source–drain voltage in relation to the source. The depth of the depletion layer at $x = 0$ (source) and $x = L$ (drain) is given by

$$y_1 = h(0) = \sqrt{\frac{2\,\epsilon_\mathrm{s}}{e\,N_\mathrm{D}}\,(V_\mathrm{bi} + V_\mathrm{G})} \tag{24.36a}$$

$$y_2 = h(L) = \sqrt{\frac{2\,\epsilon_\mathrm{s}}{e\,N_\mathrm{D}}\,(V_\mathrm{bi} + V_\mathrm{G} + V_\mathrm{D})}. \tag{24.36b}$$

The maximum value of h is a. Therefore, the pinch-off voltage V_P, at which $V_\mathrm{P} = V_\mathrm{bi} + V_\mathrm{G} + V_\mathrm{D}$ is such that $h = a$, is given by

$$V_\mathrm{P} = \frac{e\,N_\mathrm{D}\,a^2}{2\,\epsilon_\mathrm{s}}. \tag{24.37}$$

The (drift) current density along x is given by (cf. (8.53a))

$$j_x = -e\,N_\mathrm{D}\,\mu_\mathrm{n}\,E_x = e\,N_\mathrm{D}\,\mu_\mathrm{n}\,\frac{\partial V}{\partial x} \tag{24.38}$$

for the neutral part of the semiconductor. Therefore, the current in the upper half of the channel is given by

$$I_\mathrm{D} = e\,N_\mathrm{D}\,\mu_\mathrm{n}\,\frac{\partial V(x)}{\partial x}\,Z\,\left[a - h(x)\right], \tag{24.39}$$

where Z is the width of the channel (Fig. 24.14a). Although it seems that I_D depends on x, it is of course constant along the channel due to Kirchhoff's law.[4] Using the triviality $\int_0^L I_\mathrm{D}\,\mathrm{d}x = L\,I_\mathrm{D}$ and $\frac{\partial V}{\partial x} = \frac{\partial V}{\partial h}\,\frac{\partial h}{\partial x}$ with $\frac{\partial V}{\partial h} = eN_\mathrm{D}h/\epsilon_\mathrm{s}$ from (24.35), we find from (24.39)

$$I_\mathrm{D} = \frac{e^2\,\mu_\mathrm{n}\,N_\mathrm{D}^2\,Z\,a^3}{6\,\epsilon_\mathrm{s}\,L}\left[\frac{3}{a^2}\,\left(y_2^2 - y_1^2\right) - \frac{2}{a^3}\,\left(y_1^3 - y_2^3\right)\right]. \tag{24.40}$$

[4] We neglect recombination, in particular since the current is a majority-carrier current.

This equation can also be written, using (24.37) and

$$I_\mathrm{P} = \frac{e^2\,\mu_\mathrm{n}\,N_\mathrm{D}^2\,Z\,a^3}{6\,\epsilon_\mathrm{s}\,L}, \tag{24.41}$$

as

$$I_\mathrm{D} = I_\mathrm{P}\left[\frac{3V_\mathrm{D}}{V_\mathrm{P}} - 2\,\frac{(V_\mathrm{bi}+V_\mathrm{G}+V_\mathrm{D})^{3/2}-(V_\mathrm{bi}+V_\mathrm{G})^{3/2}}{V_\mathrm{P}^{3/2}}\right]. \tag{24.42}$$

The saturation current is reached for $y_2 = a$ or $V_\mathrm{bi}+V_\mathrm{G}+V_\mathrm{D} = V_\mathrm{P}$ and is given by

$$I_\mathrm{D,sat} = I_\mathrm{P}\left[1 - 3\,\frac{V_\mathrm{bi}+V_\mathrm{G}}{V_\mathrm{P}} + 2\left(\frac{V_\mathrm{bi}+V_\mathrm{G}}{V_\mathrm{P}}\right)^{3/2}\right]. \tag{24.43}$$

The dependence of the saturation current on $(V_\mathrm{G}+V_\mathrm{bi})/V_\mathrm{P}$ is depicted in Fig. 24.15b. For the threshold (gate) voltage of

$$V_\mathrm{T} = V_\mathrm{P} - V_\mathrm{bi}, \tag{24.44}$$

the saturation current is zero since then $V_\mathrm{D} = 0$.[5] Around the threshold voltage, the drain saturation current is given in lowest order of V_G as

$$I_\mathrm{D,sat} \approx \frac{3\,I_\mathrm{P}}{4}\left(\frac{V_\mathrm{G}-V_\mathrm{T}}{V_\mathrm{P}}\right)^2. \tag{24.45}$$

Thus, in order to experimentally determine the threshold voltage, $\sqrt{I_\mathrm{D}}$ is plotted versus the gate voltage and extrapolated to $I_\mathrm{D} = 0$ (dash-dotted line in Figs. 24.15 and 24.16).

The source–drain voltage at the saturation point decreases with decreasing saturation current, shown as dashed parabola-like line in Fig. 24.15a.

If the charge-carrier distribution differs from the homogeneous distribution assumed so far, a change of transistor properties arises, as shown in Fig. 24.15b for a δ-like carrier distribution. The I–V characteristic is slightly less curved, but not linear. A linear characteristic is only achievable in the drift velocity saturation regime (cf. Sect. 24.4.4).

For high source–drain voltage $V_\mathrm{D} > V_\mathrm{P} - V_\mathrm{bi} - V_\mathrm{G}$, the current remains essentially at its saturation value. For very high source–drain voltage, breakdown in the gate–drain diode can occur, when the maximum voltage, which is given by $V_\mathrm{G}+V_\mathrm{D}$ at the end of the channel, is equal to the breakdown voltage V_B.

[5] The threshold voltage can also be obtained from the condition $g_\mathrm{D0} = 0$ (cf. (24.49)).

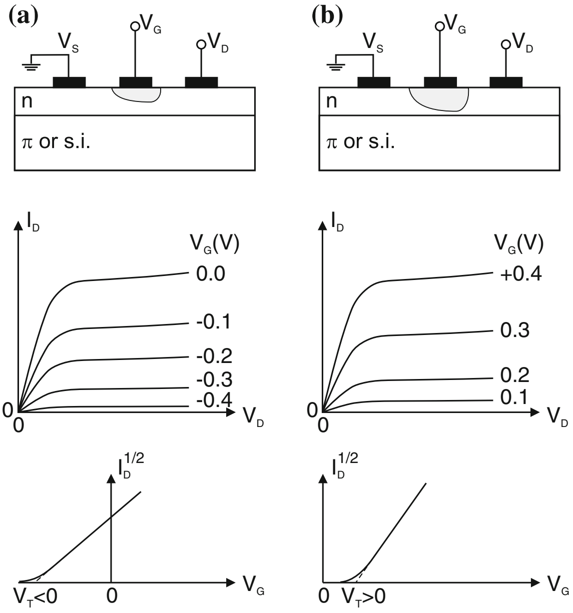


Fig. 24.16 Scheme (*top*), I_D versus V_D (*center*) and $I_D^{1/2}$ versus V_G (*bottom*) I–V characteristics for (**a**) normally on (depletion) and (**b**) normally off (accumulation) n-type JFET. Adapted from [500]

The forward transconductance g_m and the drain transconductance g_D are given by

$$g_m = \frac{\partial I_D}{\partial V_G} = g_{max} \left[\sqrt{\frac{V_{bi} + V_G}{V_P}} - \sqrt{\frac{V_{bi} + V_G + V_D}{V_P}} \right] \tag{24.46}$$

$$g_D = \frac{\partial I_D}{\partial V_D} = g_{max} \left[1 - \sqrt{\frac{V_{bi} + V_G + V_D}{V_P}} \right], \tag{24.47}$$

where

$$g_{max} = \frac{3\, I_P}{V_P} = \frac{e\, N_D\, \mu\, Z\, a}{L}. \tag{24.48}$$

The drain transconductance for $V_D \to 0$ (linear regime, dashed straight lines in Fig. 24.15a) is given by

$$g_{D0} = g_{max} \left[1 - \sqrt{\frac{V_{bi} + V_G}{V_P}} \right] = g_{m,sat}, \tag{24.49}$$

which is equal[6] to the forward transconductance in the saturation regime $g_{m,sat} = \partial I_{D,sat}/\partial V_G$.

[6] Technically, here $g_{D0} = -g_{m,sat}$, however, we had counted V_G positive for the reverse direction.

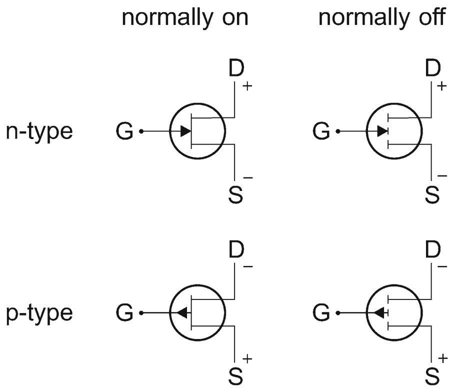


Fig. 24.17 Circuit symbols for various types of FETs

24.4.3 *Normally on and Normally Off FETs*

The JFET discussed so far had an n-conductive channel and was conductive at $V_G = 0$. It is termed an 'n-type, normally on' (or depletion) FET. If the channel is p-conductive, the FET is called 'p-type'. A FET that has a nonconductive channel at $V_G = 0$ is called 'normally off' (or accumulation) FET. In this case, the built-in voltage must be large enough to cause pinch-off. For a positive gate voltage (in the forward direction of the gate–drain diode), current begins to flow. The I–V characteristics of the four FET-types are depicted in Fig. 24.16. The circuit symbols for the four different FET types are shown in Fig. 24.17.

24.4.4 *Field-Dependent Mobility*

So far, we have considered FETs with long channels ($L \gg a$). This situation is often not the case, in particular for high-integration or high-frequency applications. For short channels, the I–V characteristics exhibit changes. The theory needs to be modified to take into account, among other effects, the electric-field dependence of the mobility (Fig. 8.10) that was discussed in Sect. 8.4.1.

Drift-Velocity Saturation

A material without negative differential mobility, such as Si or Ge, can be described with a drift-velocity model

$$v_d = \mu E \, \frac{1}{1 + \mu E / v_s} . \tag{24.50}$$

In this model, μ denotes the low-field (ohmic) mobility and v_s the drift-saturation velocity reached for $E \gg v_s/\mu$. The fraction in (24.50) describes the drift-velocity saturation.

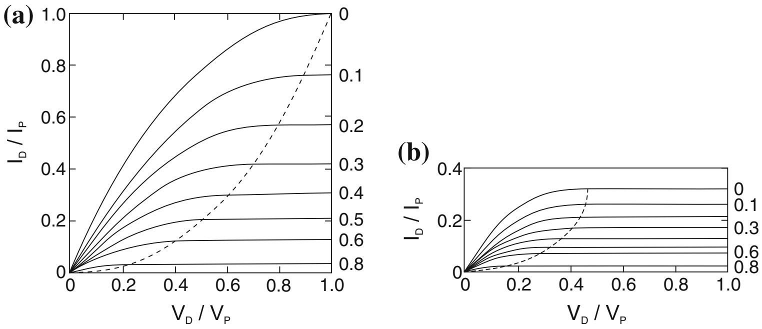


Fig. 24.18 *I–V* characteristic (**a**) without consideration of drift saturation ($z = 0$) and (**b**) with drift saturation ($z = 3$) for various values of $(V_G + V_{bi})/V_P = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8$ as indicated at the right side. The intersections of the *dashed line* and the *solid lines* indicate the beginning of saturation. Adapted from [1758]

By inserting (24.50) into (24.39), we obtain (for a n-channel)

$$I_D = -e\, N_D\, \mu_n\, E(x)\, \frac{1}{1 + \mu\, E(x)/v_s}\, [a - h(x)]\, Z, \tag{24.51}$$

and after a short calculation the drain current is given by (cf. (24.42))

$$I_D = I_P \left(1 + \frac{\mu\, V_G}{v_s\, L}\right)^{-1} \left[\frac{3V_D}{V_P} - 2\, \frac{(V_{bi} + V_G + V_D)^{3/2} - (V_{bi} + V_G)^{3/2}}{V_P^{3/2}}\right]. \tag{24.52}$$

The factor $1/(1 + z)$ with $z = \mu V_G / v_s L$ reduces the channel current due to the drift saturation effect. The effect of the parameter z is depicted in Fig. 24.18 in comparison to $z = 0$, i.e. without the drift saturation effect (or $v_s \to \infty$). The forward conductance $g_{m,sat}$ decreases with increasing z, as shown in Fig. 24.19.

Two-Region Model

In order to model the GaAs drift velocity versus field characteristic, a two-region model is used. In the front region of the channel (region I), the field is small enough and a constant mobility μ is used. In the back region of the channel (region II) is the high-field region where a constant drift velocity v_s is used. With increasing source–drain voltage, the region II (I) increases (decreases) in length. The relative width of region II is also increased with decreasing channel length.

Saturated-Drift Model

Here, the drift velocity is taken everywhere as v_s, i.e. complete drift saturation. This is a good approximation for short channels (high fields) that are in current saturation.

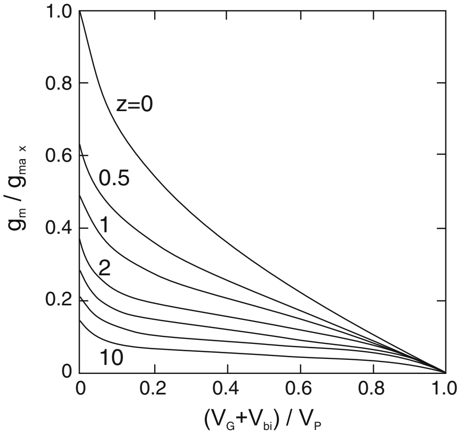


Fig. 24.19 Decrease of (saturated) forward conductance with gate voltage (according to (24.49)) and parametric dependence on z for $z = 0$, 0.5, 1, 2, 3, 5 and 10. Adapted from [1758]

In this case, the current is given by

$$I_\mathrm{D} = -e\, N_\mathrm{D}\, v_\mathrm{s}\, [a - h(x)]\, Z. \tag{24.53}$$

Equation (24.53) is valid for homogeneous doping. For other doping profiles, the current is given by

$$I_\mathrm{D} = v_\mathrm{s}\, Z \int_h^a \rho(y)\, \mathrm{d}y. \tag{24.54}$$

The forward conductance is given by

$$g_\mathrm{m} = \frac{v_\mathrm{s}\, Z\, \epsilon_\mathrm{s}}{h(V_\mathrm{G})}. \tag{24.55}$$

The transistor is more linear if the depletion-layer depth only weakly depends on the gate voltage. This can be accomplished with a doping profile with increasing doping with depth. An increase with a power law and a step-wise or exponential increase lead to a more linear $I(V)$-dependence. In the limit of δ-like doping, a linear $I_\mathrm{D,sat}$ versus V_G relation develops. Indeed, FETs with graded or stepped doping profiles exhibit improved linearity and are used for analog circuits.

Nonequilibrium Velocity

Below the electric field for which the drift velocity in GaAs peaks, the carriers can be considered to be in equilibrium. If the field is higher, velocity overshoot (Sect. 24.4.4 and Fig. 8.13) occurs. The carriers have a higher velocity (and ballistic transport) before they relax to the lower equilibrium (or steady-state) velocity after intervalley scattering. This effect will shorten the transit time in short-channel FETs.

24.4.5 High-Frequency Properties

Two factors limit the high-frequency performance of a FET: The transit time and the RC time constant. The transit time t_r is the time that the carrier needs to go from source to drain. For the case of constant mobility (long channel) and constant drift velocity (short channel), the transit time is given by (24.56a and b), respectively.

$$t_\mathrm{r} = \frac{L}{\mu\, E} \approx \frac{L^2}{\mu\, V_\mathrm{G}} \tag{24.56a}$$

$$t_\mathrm{r} = \frac{L}{v_\mathrm{s}} . \tag{24.56b}$$

For a 1-µm long gate in a GaAs FET, the transit time is of the order of 10 ps. This time is typically small compared to the RC time constant due to the capacitance C_GS and transconductance. The cutoff frequency is given by

$$f_\mathrm{T} = \frac{g_\mathrm{m}}{2\pi\, C_\mathrm{GS}} . \tag{24.57}$$

24.5 MOSFETs

The MOSFET has four terminals. In Fig. 24.14b, two n-type regions (source and drain) are within a p-type substrate. The n-type channel (length L) forms underneath a MIS diode. A forth electrode sets the substrate bias. The source electrode is considered to be at zero potential. The important parameters are the substrate doping N_A, the insulator thickness d and the depth r_j of the n-type regions. Around the MOSFET structure is an oxide to insulate the transistor from neighboring devices.

24.5.1 Operation Principle

When there is no applied gate voltage, only the saturation current of the pn-diode(s) between source and drain flows. In thermodynamic equilibrium (Fig. 24.20c), the necessary surface potential for inversion at the MIS diode is $\Psi_\mathrm{s}^\mathrm{inv} \approx 2\Psi_\mathrm{B}$. If there is a finite drain voltage, a current flows and there is no longer equilibrium. In this case, the quasi-Fermi level of the electrons (or generally of the minority carriers) is lowered and a higher gate voltage is needed to create inversion. The situation at the drain is depicted in Fig. 24.21.

In nonequilibrium, the depletion layer width is a function of the drain voltage V_D. In order to reach strong inversion at the drain, the surface potential must be at least $\Psi_\mathrm{s}^\mathrm{inv} \approx V_\mathrm{D} + 2\Psi_\mathrm{B}$.

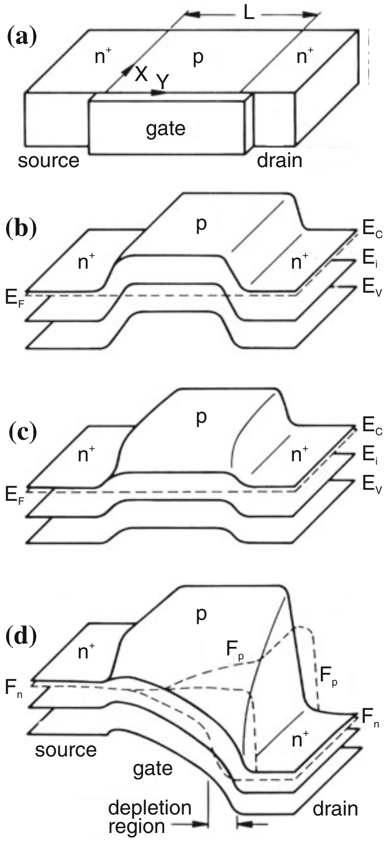


Fig. 24.20 (**a**) Schematic geometry of a MOSFET and its band diagram for (**b**) flat-band conditions for zero gate voltage (and $V_D = 0$), (**c**) thermodynamic equilibrium with reverse gate voltage (weak inversion, still $V_D = 0$) and (**d**) nonequilibrium with nonzero drain voltage and gate voltages (with most of the channel being inverted, the depletion region is indicated). Adapted from [1759]

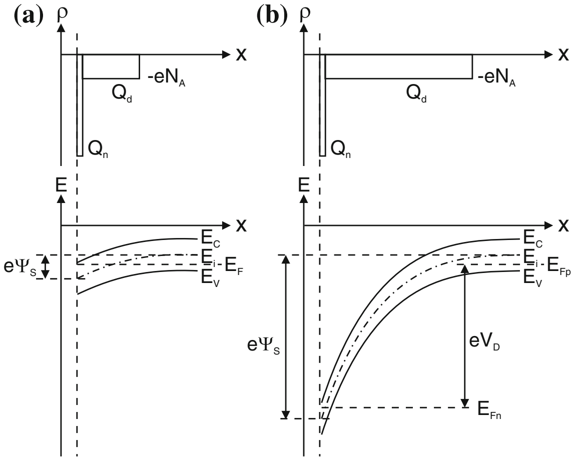


Fig. 24.21 Charge-carrier distribution (*top*) and band diagram (*bottom*) at the inverted p-region of a MOSFET for (**a**) thermodynamic equilibrium ($V_D = 0$) and (**b**) nonequilibrium at drain

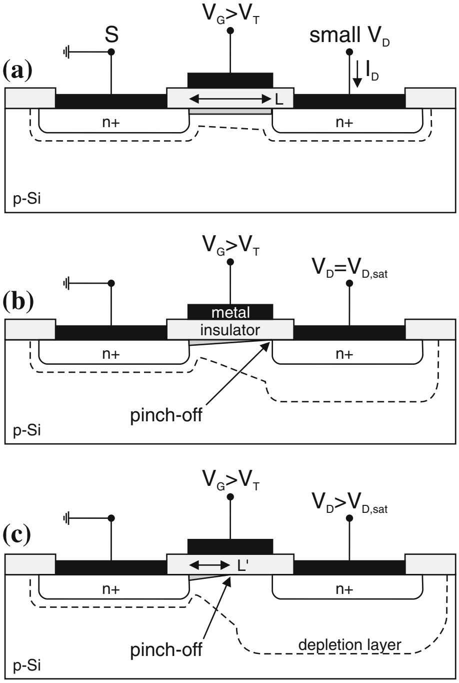


Fig. 24.22 (**a**) MOSFET with inverted channel (*dark grey*) of length L for small source–drain voltage V_D in linear regime, (**b**) at the start of saturation at pinch-off, (**c**) in the saturation regime with reduced channel length L'. The pinch-off point is denoted by an arrow in (**b**) and (**c**). The *dashed lines* denote the extension of the depletion region. Adapted from [500]

If the gate voltage is such that an inversion channel is present from source to drain, a current will flow for a small drain voltage (Fig. 24.22a). Initially, the current will increase linearly with V_D, depending on the conductivity of the channel. With increasing drain voltage, the quasi-Fermi level of the electrons is lowered until, finally at $V_D = V_{D,sat}$, the inversion channel depth becomes zero (pinch-off at the point denoted with an arrow in Fig. 24.22b). The current at this condition is denoted as $I_{D,sat}$. For a further increase of V_D, the pinch-off point moves closer to the source contact and the channel length (inverted region) is shortened (arrow in Fig. 24.22c). The voltage at the pinch-off point remains $V_{D,sat}$ and thus the current in the channel remains constant at $I_{D,sat}$.

24.5.2 Current–Voltage Characteristics

We assume now that the potential $V(y)$ varies along the channel from $V = 0$ at $y = 0$ to $V = V_D$ at $y = L$. In the gradual-channel approximation, the voltage drop V_i across the oxide is

$$V_i(y) = V_G - \Psi_s(y), \tag{24.58}$$

where Ψ_s is the surface potential in the semiconductor (see Fig. 21.33). The total charge induced in the semiconductor (per unit area) is, using (21.88), given by

$$Q_\mathrm{s}(y) = -\left[V_\mathrm{G} - \Psi_\mathrm{s}(y)\right] C_\mathrm{i}, \tag{24.59}$$

with C_i being the insulator capacitance (per unit area), as given in (21.89).

The inversion surface potential can be approximated by $\Psi_\mathrm{s}(y) \approx 2\Psi_\mathrm{B} + V(y)$ (see Fig. 24.21). With (21.93) the depletion-layer charge is

$$Q_\mathrm{d}(y) = -e\, N_\mathrm{A}\, w_\mathrm{m} = -\sqrt{2\, \epsilon_\mathrm{s}\, e\, N_\mathrm{A}\, \left[2\Psi_\mathrm{B} + V(y)\right]}, \tag{24.60}$$

such that, using (24.59), the inversion layer charge is

$$\begin{aligned} Q_\mathrm{n}(y) &= Q_\mathrm{s}(y) - Q_\mathrm{d}(y) \\ &= -\left[V_\mathrm{G} - V(y) - 2\Psi_\mathrm{B}\right] C_\mathrm{i} + \sqrt{2\, \epsilon_\mathrm{s}\, e\, N_\mathrm{A}\, \left[2\Psi_\mathrm{B} + V(y)\right]}. \end{aligned} \tag{24.61}$$

For the calculation of the drain current, we consider the increase of channel resistance $\mathrm{d}R(y)$ along a line element $\mathrm{d}y$ of the channel. The integral of the conductivity over the cross section A of the channel (width Z) is

$$\iint_A \sigma(x,z)\mathrm{d}x\, \mathrm{d}z = -e\, \mu_\mathrm{n} \iint_A n(x,z)\, \mathrm{d}x\, \mathrm{d}z = Z\, \mu_\mathrm{n}\, |Q_\mathrm{n}(y)|. \tag{24.62}$$

Therefore,

$$\mathrm{d}R(y) = \mathrm{d}y\, \frac{1}{Z\, \mu_\mathrm{n}\, |Q_\mathrm{n}(y)|}. \tag{24.63}$$

Here we have assumed that the mobility is constant along the channel, i.e. not field dependent. The change of voltage across the line element $\mathrm{d}x$ is

$$\mathrm{d}V(y) = I_\mathrm{D}\, \mathrm{d}R = \frac{I_\mathrm{D}\, \mathrm{d}y}{Z\, \mu_\mathrm{n}\, |Q_\mathrm{n}(y)|}. \tag{24.64}$$

We note the drain current is independent of x. Using (24.61) and performing the integral of (24.64) from $V(y=0) = 0$ to $V(y=L) = V_\mathrm{D}$, we find

$$\begin{aligned} I_\mathrm{D} = \mu_\mathrm{n}\, C_\mathrm{i}\, \frac{Z}{L} \Bigg\{ \left(V_\mathrm{G} - 2\Psi_\mathrm{B} - \frac{V_\mathrm{D}}{2}\right) - \frac{2}{3}\, \frac{\sqrt{2e\epsilon_\mathrm{s}N_\mathrm{A}}}{C_\mathrm{i}} \Big[(V_\mathrm{D} + 2\Psi_\mathrm{B})^{3/2} \\ - (2\Psi_\mathrm{B})^{3/2} \Big] \Bigg\}. \end{aligned} \tag{24.65}$$

This characteristic is depicted in Fig. 24.23a. In the linear regime (small drain voltage, $V_\mathrm{D} \ll (V_\mathrm{G} - V_\mathrm{T})$), the drain current is given by

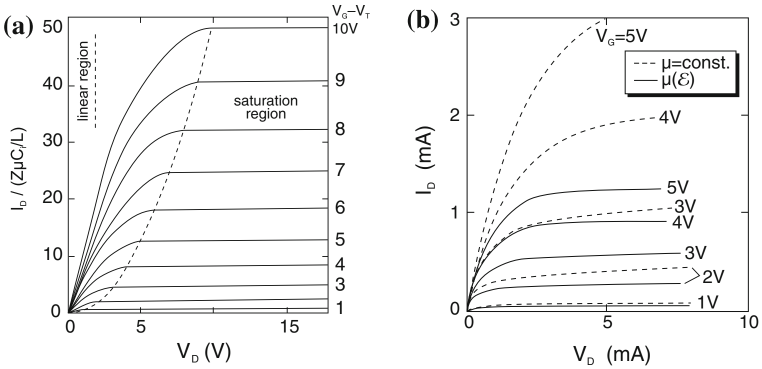


Fig. 24.23 (**a**) Idealized I–V characteristics for a MOSFET with constant mobility. The *dashed line* visualizes the drain (saturation) voltage for which the current is equal to $I_{\mathrm{D,sat}}$. The *solid lines* are for various values of the gate voltage $V_{\mathrm{G}} - V_{\mathrm{T}} = 1$–$10$ V. Adapted from [500] (**b**) I–V characteristics taking into account the effect of field-dependent mobility (*solid lines*) in comparison to the constant-mobility model (*dashed lines*) for various gate voltages as labeled. Adapted from [1760]

$$I_{\mathrm{D}} \cong \mu_{\mathrm{n}}\, C_{\mathrm{i}}\, \frac{Z}{L}\, (V_{\mathrm{G}} - V_{\mathrm{T}})\, V_{\mathrm{D}}. \tag{24.66}$$

The threshold voltage V_{T}, i.e. the gate voltage for which the channel is opened and a current can flow, is given for small drain voltage (linear regime) by

$$V_{\mathrm{T}} = 2\,\Psi_{\mathrm{B}} + \frac{\sqrt{4\,e\,\epsilon_{\mathrm{s}}\,N_{\mathrm{A}}\,\Psi_{\mathrm{B}}}}{C_{\mathrm{i}}}. \tag{24.67}$$

The transconductances in the linear regime are easily obtained as

$$g_{\mathrm{m}} = \mu_{\mathrm{n}}\, C_{\mathrm{i}}\, \frac{Z}{L}\, V_{\mathrm{D}} \tag{24.68a}$$

$$g_{\mathrm{D}} = \mu_{\mathrm{n}}\, C_{\mathrm{i}}\, \frac{Z}{L}\, (V_{\mathrm{G}} - V_{\mathrm{T}})\,. \tag{24.68b}$$

The saturation current (for constant mobility) is approximately

$$I_{\mathrm{D,sat}} \cong \mu_{\mathrm{n}}\, C_{\mathrm{i}}\, \frac{m\,Z}{L}\, (V_{\mathrm{G}} - V_{\mathrm{T}})^2\,, \tag{24.69}$$

where m depends on the doping concentration and is about 0.5 for low doping. For low p-doping of the substrate, the threshold voltage in (24.69) for the saturation regime is also given by (24.67). At higher doping, the threshold voltage becomes dependent on the gate voltage. C_{i} denotes the insulator capacitance

$$C_{\mathrm{i}} = \epsilon_{\mathrm{i}}/d_{\mathrm{i}}. \tag{24.70}$$

The forward transconductance in the saturation regime is

$$g_{\mathrm{m,sat}} = \mu_{\mathrm{n}}\, C_{\mathrm{i}}\, \frac{2\, m\, Z}{L}\, \left(V_{\mathrm{G}} - V_{\mathrm{T}}\right). \tag{24.71}$$

For constant drift velocity (Fig. 24.23b for field-dependent mobility), the saturation current is given by

$$I_{\mathrm{D,sat}} = Z\, C_{\mathrm{i}}\, v_{\mathrm{s}}\, \left(V_{\mathrm{G}} - V_{\mathrm{T}}\right), \tag{24.72}$$

and the forward transconductance in the saturation regime is

$$g_{\mathrm{m,sat}} = Z\, C_{\mathrm{i}}\, v_{\mathrm{s}}. \tag{24.73}$$

We note that the transistor properties depend on and can be separated into the geometry factor (Z/L) and the material properties ($\mu\, C_{\mathrm{i}} = \mu\, \epsilon_{\mathrm{i}}/d_{\mathrm{i}}$).

The threshold voltage can be changed by the substrate bias V_{BS} as ($\beta = e/kT$)

$$\varDelta V_{\mathrm{T}} = \frac{a}{\sqrt{\beta}} \left(\sqrt{2\Psi_{\mathrm{B}} + V_{\mathrm{BS}}} - \sqrt{2\Psi_{\mathrm{B}}} \right), \tag{24.74}$$

with (L_{D} being the Debye length (cf. 21.81b))

$$a = 2\, \frac{\epsilon_{\mathrm{s}}}{\epsilon_{\mathrm{i}}}\, \frac{d}{L_{\mathrm{D}}}. \tag{24.75}$$

Experimental data are shown in Fig. 24.24. For a Si/SiO$_2$ gate diode, $a = 1$ for, e.g. $d_{\mathrm{i}} = 10\,\mathrm{nm}$ and $N_{\mathrm{A}} = 10^{16}\,\mathrm{cm}^{-3}$. For gate voltages below V_{T}, the current is given by the diffusion current, similar to a npn transistor. This regime is important for low-voltage, low-power conditions. The related drain current is termed the *subthreshold* current and is given by

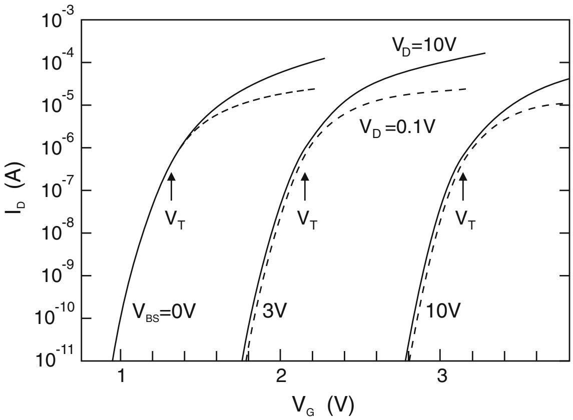


Fig. 24.24 Experimental subthreshold *I–V* characteristic of a MOSFET device with long channel (15.5 μm). *Solid lines* for $V_{\mathrm{D}} = 10\,\mathrm{V}$, *dashed lines* for $V_{\mathrm{D}} = 0.1\,\mathrm{V}$. Adapted from [1761]

$$I_{\mathrm{D}} = \mu_{\mathrm{n}} \frac{Z\, a\, C_{\mathrm{i}}\, n_{\mathrm{i}}^2}{2\, L\, \beta^2\, N_{\mathrm{A}}^2} \left[1 - \exp\left(-\beta V_{\mathrm{D}}\right)\right] \exp\left(-\beta \Psi_{\mathrm{s}}\right)\ \left(\beta \Psi_{\mathrm{s}}\right)^{-1/2} . \tag{24.76}$$

The drain current therefore increases exponentially with V_{G}, as shown in Fig. 24.24. V_{G} is ly proportional to Ψ_{B}:

$$\Psi_{\mathrm{s}} = (V_{\mathrm{G}} - V_{\mathrm{FB}}) - \frac{a^2}{2\beta} \left(\sqrt{1 + \frac{4}{a^2}\ (\beta V_{\mathrm{G}} - \beta V_{\mathrm{FB}} - 1)} - 1 \right), \tag{24.77}$$

where V_{FB} is the flat-band voltage of the gate MIS diode. The drain current is independent of V_{D} for $V_{\mathrm{D}} \gtrsim 3kT/e$.

24.5.3 MOSFET Types

MOSFETs can have an n-type channel (on a p-substrate) or a p-channel (on an n-type substrate). So far, we have discussed the normally off MOSFET. If there is a conductive channel even without a gate voltage, the MOSFET is normally on. Here, a negative gate voltage must be applied to close the channel. Therefore, similar to the JFET, a total of four different types of MOSFET exist, see Fig. 24.25.

24.5.4 Complementary MOS

Complementary metal–oxide–semiconductor technology (CMOS) is the dominating technology for highly integrated circuits. In such devices, MOSFETs with n-channel (NMOS) and p-channel (PMOS) are used on the same chip. The basic structure of logic circuits, the inverter, can be realized with a pair of NMOS and PMOS transistors, as shown in Fig. 24.26a with two normally off transistors. The load capacitor represents the capacitance of the following elements.

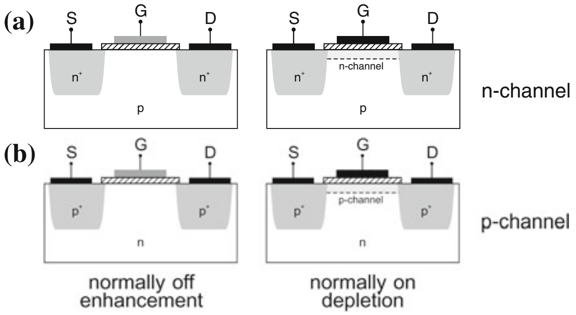


Fig. 24.25 The four MOSFET types. (**a**) Enhancement and (**b**) depletion type with n-channel (*top row*) and p-channel (*bottom row*)

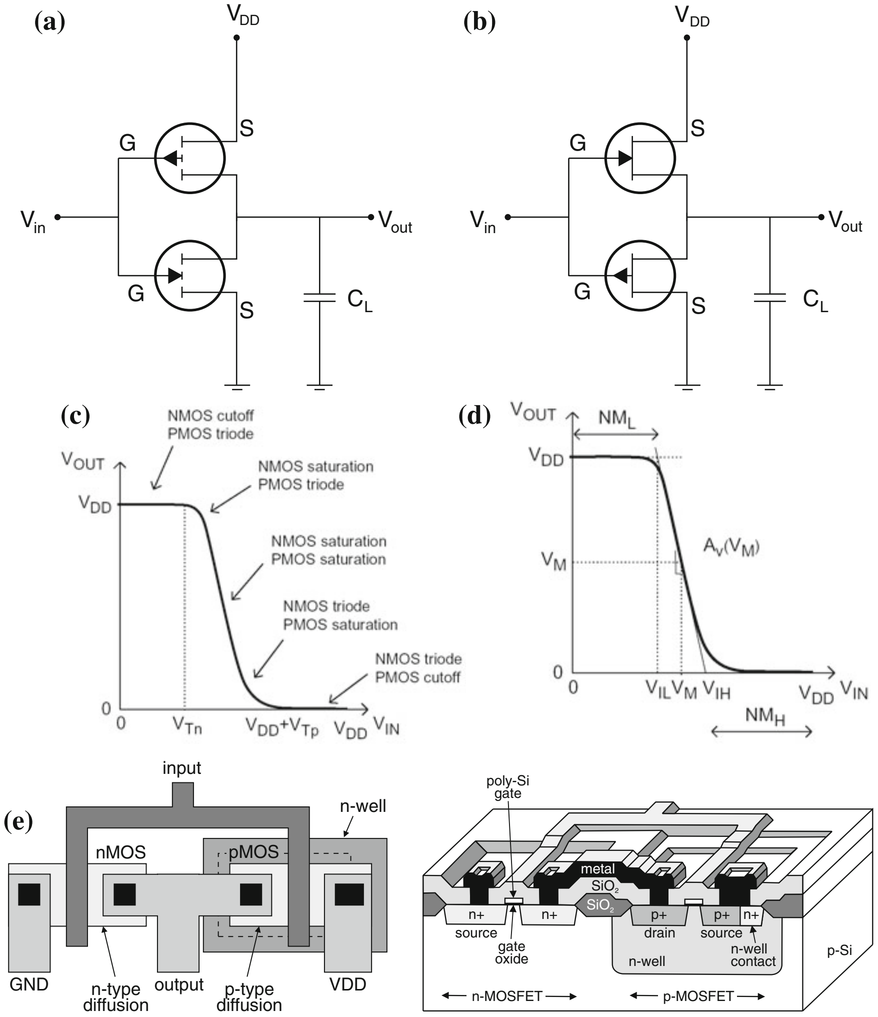


Fig. 24.26 Circuit diagram of (**a**) inverter with n-type (bottom) and p-type (normally off, enhancement mode) FETs and (**b**) inverter with p-type (bottom) and n-type (normally on, depletion mode) FETs. (**c**) Inverter characteristic with the transistor thresholds indicated, (**d**) inverter characteristic with middle voltage V_M indicated. $NM_{L,H}$ denotes the low- and high-noise margins, respectively, i.e. the voltage by which the input voltage can fluctuate without leading to switching. (**e**) Composite layout (*left panel*) and cross-sectional view (*right panel*) of CMOS inverter. Part (**e**) adapted from [1762]

If the input voltage is $V_{in} = 0$, the NMOS transistor is nonconductive ('off'). The (positive) voltage V_{DD} is at the PMOS transistor source, thus the gate is negative in relation to the source and the transistor is conductive ('on') since $-V_{DD} = V_{Gp} < V_{Tp} < 0$ (see Fig. 24.25). The current flows through the capacitor that becomes charged to $V_{out} = V_{DD}$. The current then subsides, since V_D at the PMOS becomes zero. If the input voltage is set to $V_{\rm DD}$, the NMOS transistor has a positive gate–source voltage larger than the threshold $V_{\rm Tn} < V_{\rm Gn} = V_{\rm DD}$ and becomes conductive. The charge from the capacitor flows over the NMOS to ground. The PMOS transistor has zero gate–source voltage and is in the 'off' state. In this case, the voltage $V_{\rm DD}$ drops entirely across the PMOS and the capacitor is uncharged with $V_{\rm out} = 0$.

In both its logic states, the CMOS inverter does not consume power. No current[7] flows in either of the two steady states since one of the two transistors is in both cases in the 'off' state. Current flows only during the switching operation. Therefore, the CMOS scheme allows for low power consumption.

The middle voltage for which $V_{\rm in} = V_{\rm out}$ can be calculated from the MOSFET characteristics. Both are, for this condition, in saturation and the currents are given by (cf. (24.69))

$$I_{\rm Dn} = \mu_{\rm n}\, C_{\rm ox}\, \frac{Z_{\rm n}}{2L_{\rm n}}\, (V_{\rm M} - V_{\rm Tn})^2 \tag{24.78a}$$

$$I_{\rm Dp} = \mu_{\rm p}\, C_{\rm ox}\, \frac{Z_{\rm p}}{2L_{\rm p}}\, \left(V_{\rm DD} - V_{\rm M} - V_{\rm Tp}\right)^2 . \tag{24.78b}$$

With $\gamma = \frac{Z_{\rm p}}{Z_{\rm n}}\, \frac{L_{\rm n}}{L_{\rm p}}\, \frac{\mu_{\rm p}}{(-\mu_{\rm n})}$, we find from $I_{\rm Dn} = -I_{\rm Dp}$,

$$V_{\rm M} = \frac{V_{\rm Tn} + \gamma\, \left(V_{\rm DD} + V_{\rm Tp}\right)}{1 + \gamma} . \tag{24.79}$$

As gate material, often polycrystalline silicon (poly-Si) is used (cf. Fig. 21.28). It is used instead of metals because its work function matches that of silicon closely. Also, poly-Si is more resistant to temperature. Despite its high doping, the resistance of poly-Si is two orders of magnitude larger than that of metals. Since it is easily oxidized, it cannot be used with high-k oxide dielectrics.[8]

For optimized ohmic contacts on the n- and p-Si, different metals are used to create a small barrier height (Fig. 21.23a) and low contact resistance (cf. Sect. 21.2.6). Figure 24.27 visualizes the band edges of silicon in relation to the work functions of various metals (see Table 21.1). For example, the work function of titanium matches the electron affinity of n-Si closely. However, a direct deposition of Ti on Si results in a Schottky barrier of 0.5 eV [1415]. A surface passivation with a group-VI element such as Se can help reduce this value to 0.19 eV [1763].

In the latest generation of CMOS ICs the PMOS (NMOS) device has a built-in compressive (tensile) channel strain for modifying the effective mass (cf. Sect. 6.10.2), both allowing higher drive current due to higher mobility. A detailed treatment can be found in [1764].

[7] Except for the subthreshold current and other leakage currents. These need to be reduced further since the dissipated power limits chip performance (speed and device density) and battery lifetime in handheld applications.

[8] The term 'high-k dielectric' means a dielectric material with large dielectric constant ϵ.

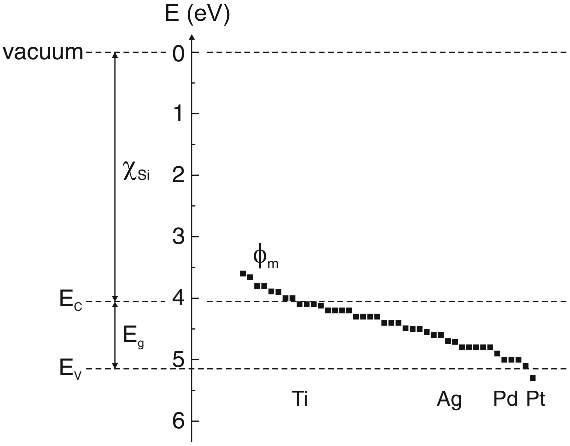


Fig. 24.27 Silicon band edges in relation to different metals and their work functions

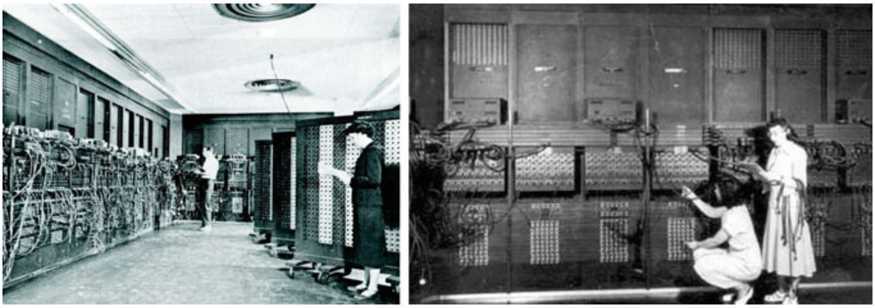


Fig. 24.28 ENIAC, the first electronic computer (J.P. Eckert, J.W. Mauchly, 1944/5). The images show only a small part of the 18 000 vacuum tubes

24.5.5 Large-Scale Integration

Historic Development

Compared to the first computers on the basis of vacuum tubes (triodes), e.g. ENIAC (Fig. 24.28), today's devices are extremely miniaturized and need many orders of magnitude less power per operation. ENIAC needed 174 kW of power. A comparable computing power was reached in 1971 with the few cm^2 large Intel 4004 microprocessor (Fig. 24.29b) consuming only several Watts with 2300 transistors. In 2004 about 42 million transistors were integrated in the Pentium 4 microprocessor (Fig. 24.30). Also, memory chips started to become highly integrated (Fig. 24.29a).

The development of electronic circuit integration is empirically described by Moore's 'law' [1765] that has been valid since the 1970s. According to this law,

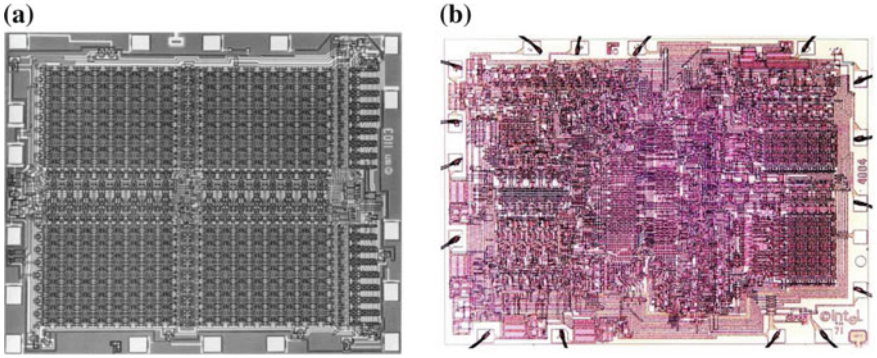


Fig. 24.29 (**a**) Intel™ 1103 1 KByte (1024 memory cells) dynamic random access memory (RAM), arranged in four grids with 32 rows and columns (1970), chip size: 2.9 × 3.5 mm^2. (**b**) Intel™ 4004 microprocessor (1971), chip size: 2.8 × 3.8 mm^2, circuit lines: 10 µm, 2, 300 MOS transistors, clock speed: 108 kHz

the number of transistors doubles every 20 months (Fig. 24.31a). At the same time, the performance has been improved by an increase of the clock speed (Fig. 24.31b).[9]

Interconnects

Moore's second law says that the cost of production also doubles for each new chip generation and is currently (2004) in the multi-billion US$ range. Most of the cost saved by integration is due to efficient *wiring* (interconnects) of the components, in 2004 (65 nm node) in eight layers above the active elements (transistors and capacitors) (Fig. 24.33), in 2008 (45 nm node) in eleven layers. Plane-view images of the first three layers of the interconnects are shown in Fig. 24.34. The Cu interconnects are fabricated with the so-called damascene process [1766–1768]. Barrier layers (e.g. TaN or TiN) are required to avoid out-diffusion of Cu into the silicon or other parts of the circuit. Three effects limit the conductivity: The interconnect metal line width and height approaches the mean free path of carriers ($d_{\mathrm{Cu}} \approx 40\,\mathrm{nm}$) [633, 1769], grain boundary scattering can limit mobility since grain size is reduced for thinner lines, and the (high resistivity) barrier reduces space for the conductive part of the metal line. In Fig. 24.32 the increase of the resistivity of copper with reduced dimension is shown as a function of film thickness t and for a 100 nm-film as a function of line width w. In a simplified approach, the line resistivity ρ_{line} is given as [1769]

$$\frac{\rho_{\mathrm{line}}}{\rho_0} = 1 + \frac{3}{8}\,(1-p)\left(\frac{d}{t} + \frac{d}{w}\right), \tag{24.80}$$

ρ_0 denoting the bulk resistivity (1.7 µΩ cm for Cu), d being the mean free path (8.7) and p being the electron scattering parameter ($p = 0$ for diffuse scattering).

[9] After year 2003 data for maximum clock rate are not for highest integration density processors.

Fig. 24.30 The Intel$^{\mathrm{TM}}$ Pentium 4 microprocessor (2000), circuit lines: 0.18 µm, 42 million transistors, clock speed: 1.5 GHz

In order to achieve the best high frequency performance the material between the metal interconnects should have low dielectric constant (‘low-k’ dielectric). Alternative materials to the standard SiO_2 ($\epsilon_r \approx 4.1$) are investigated such as SiOF ($\approx$3.8), SiCOH ($\approx$3.0), porous materials ($\approx$2.5) and air gaps [1770].

CMOS Scaling

Using planar technologies, LSI (large-scale integration), VLSI (very large-scale integration), ULSI (ultra large-scale integration) and further generations of devices have

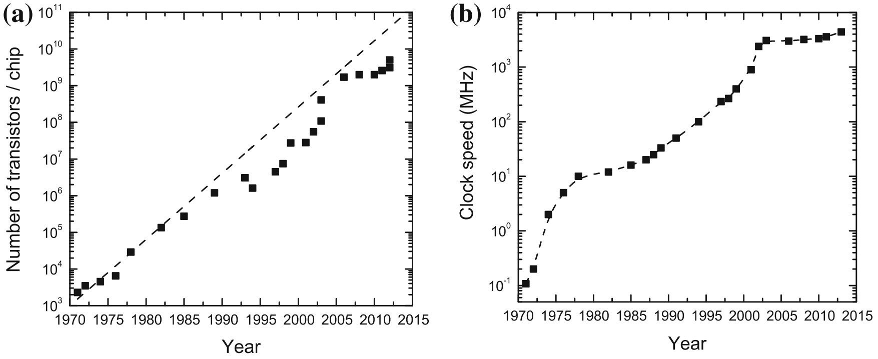
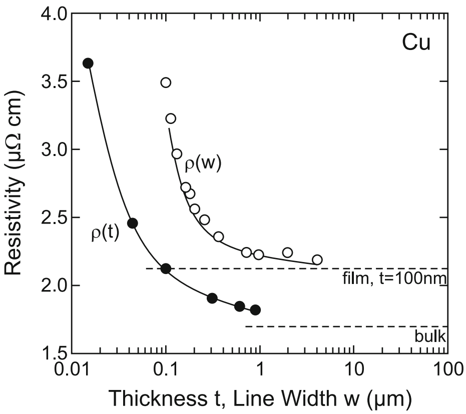


Fig. 24.31 (**a**) Moore's law on the exponential increase of transistors per chip (for Intel$^{\mathrm{TM}}$ processor chips). *Dashed line* corresponds to doubling in 20 months. (**b**) Historical increase of maximum clock speed, *dashed line* is guide to the eye. Note the almost constant rate of 10 MHz from the mid-1970s to the mid-1980s and another plateau developing after 2000

Fig. 24.32 Resistivity of copper at room temperature for various film thickness (*solid circles*), and for a 100 nm-film as a function of line width w (*empty circles*). *Solid lines* are theoretical dependence according to (24.80). The *dashed lines* indicate the limits for bulk material ($t \to \infty$) and for large line thickness ($d = 100$ nm, $w \to \infty$). Adapted from [1769]



been conceived, driven by high-density electronic memory devices. Subsequently also logic devices are produced with reduced device size.

The increase of the number of transistors per area requires the scaling of their geometrical properties. This impacts many other properties of the transistor and their scaling needs to be considered as well. From a general perspective, the physical properties scale while the thermal energy kT remains constant for room-temperature electronics.

If channel width Z and channel length L of a transistor are scaled down by a factor of $s > 1$, $Z' = Z/s$ and $L' = L/s$, the area obviously scales as $A' = A/s^2$. In subsequent transistor generations $s = \sqrt{2}$, i.e. doubling of the number of devices per area. In order to maintain the aspect ratio of the device also the oxide thickness (d_i) is scaled, $t'_\mathrm{ox} = t_\mathrm{ox}/s$ ('classical scaling').

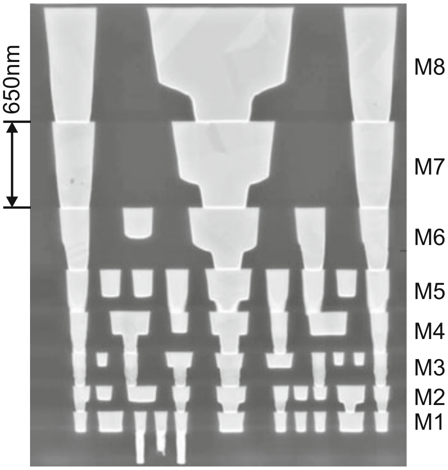


Fig. 24.33 Cross section through a logic chip (65 nm technology, 35 nm gate length) with eight layers of dual damascene Cu interconnects (M1–M8) with low-k carbon-doped oxide ($\epsilon_r = 2.9$) inter-level dielectric above the active elements. Adapted from [1771]

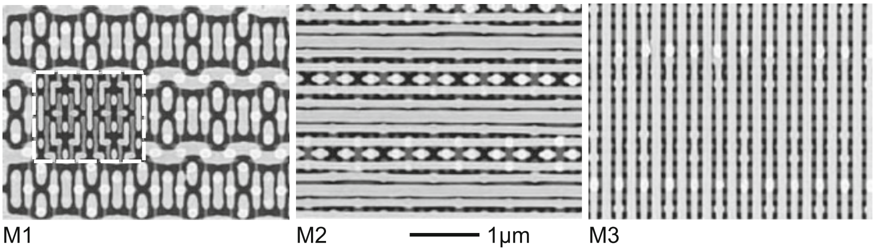


Fig. 24.34 Plane-view of the first three interconnect layers of a 45 nm node SRAM array (Intel® Xeon®). In the image of the M1 layer the gate layer metal connects are shown in the *inset*, framed with a *white dashed line*. Adapted from [1772]

The ultimate design criteria are maximum temperature and maximum power loss. The maximum temperature needs to be obeyed, the worst case usually taken as 100 °C. The power loss per area, e.g. heating, needs to stay constant at an appropriate maximum level around 200 kW/m^2 (Fig. 24.35b) unless higher (and more expensive) efforts on cooling are made. At the same time the device performance needs to be maintained if not improved, e.g. for lower power dissipation in battery operated devices. Very important is the reduction of operation voltage V_{DD} in order to keep electric fields and power consumption small enough (Fig. 24.35). The power consumption in stand-by mode P_{off} depends on V_{DD} and the subthreshold (off) current

$$P_{\mathrm{off}} = W_{\mathrm{tot}}\, V_{\mathrm{DD}}\, I_{\mathrm{off}}, \tag{24.81}$$

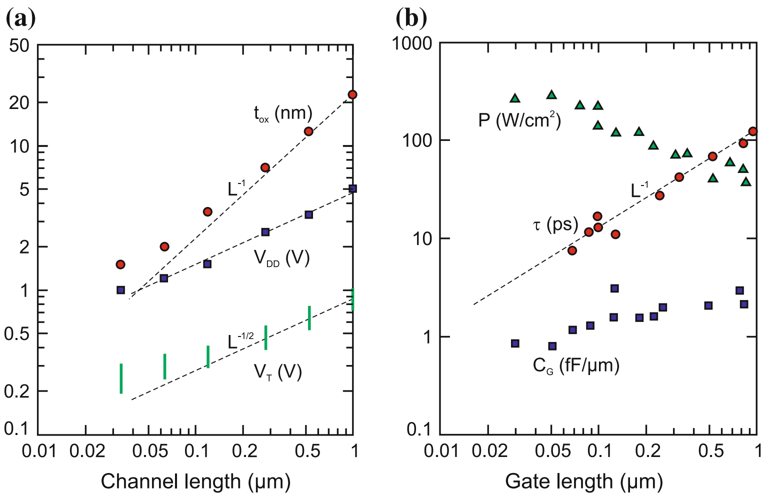


Fig. 24.35 Scaling of MOSFET parameters gate oxide thickness t_{ox}, power supply voltage V_{DD} (across source–drain), threshold voltage V_{T}, total power loss per area P, gate capacitance per channel width C_{G} and inverter delay τ, the time required to propagate a transition through a single inverter driving a second, identical inverter, commonly used as a means of gauging the speed of CMOS transistors. Data for (**a**) from [1773] and for (**b**) selected from [1774]

where W_{tot} is the total width of the turned-off devices and I_{off} is the average off-current per device per width. The latter increases exponentially with reduced threshold voltage V_{T},

$$I_{\mathrm{off}} = I_0 \exp\left(-\frac{e\,V_{\mathrm{T}}}{n\,kT}\right), \tag{24.82}$$

with ideality factor $n \approx 1.2$ and $I_0 \approx$1–10 μA/μm [1773]. A well-functioning MOSFET requires a ratio of $V_{\mathrm{T}}/V_{\mathrm{DD}}$ of < 0.3.

The power consumption in active mode P_{ac} depends also on the clock speed (frequency f) that increases with higher integration due to shorter gate length,

$$P_{\mathrm{ac}} = C_{\mathrm{sw}}\,V_{\mathrm{DD}}^2\,f, \tag{24.83}$$

where C_{sw} is the total node capacitance being charged and decharged in a clock cycle.

Historically the oxide thickness has been reduced less than the channel length [1773] (Fig. 24.35a) leading to increased local fields. The reduction of the physical gate oxide thickness is limited due to gate leakage through tunneling [1775]. While for a gate voltage of 1.5 V and oxide thickness $t_{\mathrm{ox}} = 3.6$ nm the leakage current is only about 10^{-8} A/cm^2, it is about 1 A/cm^2 for $t_{\mathrm{ox}} = 2.0$ nm and about 10^4 A/cm^2 for $t_{\mathrm{ox}} = 1.0$ nm. Obviously variations of oxide thickness are more harmful at small average thickness. 1.2 nm physical SiO$_2$ thickness has been used in the 90 nm (gate length) logic node.

The technological solution for further reduction of oxide thickness is the use of geometrically thicker layers, to suppress tunneling, with higher dielectric constant ('high-k dielectrics'), e.g. HfO_2 [1776], to maintain reasonable gate capacitance per gate width

$$C_G = \frac{\epsilon_{ox}}{t_{ox}} L, \tag{24.84}$$

(cmp. (24.70)) at a value of about 1.0–1.5 fF/μm (Fig. 24.35b). For the 45 nm technology node a 0.7-fold reduction in electrical oxide thickness was achieved while reducing gate leakage 1000× for the PMOS and 25× for the NMOS transistors [1777].

Materials

The electronics industry is based on silicon as the material for transistors. However, many other materials are incorporated in the technology. Traditionally silicon dioxide gate oxide is used, silicon nitride for insulation layers and polysilicon for gate contacts. For wiring aluminum has been used. Silicides were introduced as contact materials around 1986.

Progress was made with copper interconnects (IBM, 1997), replacing aluminum. The better electrical and heat conductivity could previously not be used since Cu is a deep level in Si (cf. Fig. 7.6). The key to success was an improved barrier technology based on amorphous TaN- or TiN-based barrier layers to prevent the diffusion of Cu into the silicon and dielectric layers. The first chip from series production, incorporating the Cu technology, was the PowerPC 750 (400 MHz) in 1998. Since 2000 high-k, i.e. large ϵ_r, Hf-containing gate dielectrics are used (Fig. 24.36). HfO_2 has a dielectric constant of 25–30. 45 nm node technology probably uses HfZrO, HfSiO or HfSiON [1778] gate dielectrics with $k \sim 12$ and an electrical thickness of $t_{ox}\, \epsilon_{SiO_2}/\epsilon_{ox} = 1.0\,$nm.

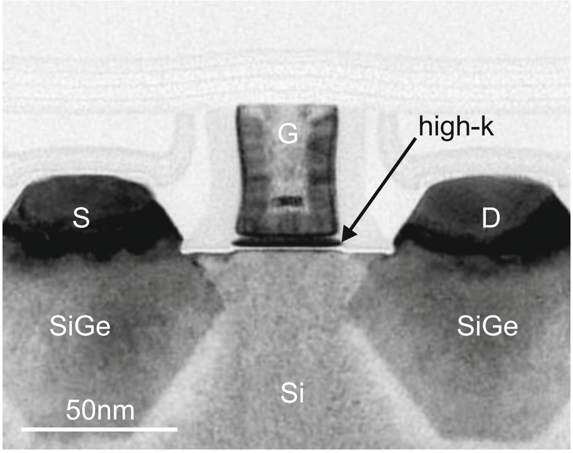


Fig. 24.36 Cross-section TEM image of 45 nm node PMOS transistor with high-k Hf-containing gate oxide (*dark*) above a thin SiO_2 layer (*white*). The role of the stressor SiGe pockets is explained in Fig. 24.35. Adapted from [1779]

Germanium is reintroduced into mainstream semiconductor technology via SiGe stressors in the source and drain for PMOS. Uniaxial compressive strain in the channel region leads to 30 % increased saturation current [1780] mostly due to reduced effective masses [658, 1781, 1782] (Sect. 8.3.11) for 90 nm transistors. Similarly, uniaxial tensile strain in NMOS, introduced by SiN caps or more recently tensile trench contacts [1777], allows for 10 % higher saturation current [1780] (Fig. 24.37). The enhanced electron mobility is due to strain-induced splitting of the X-valley and change of electron mass [1783]. Further improvement to 18 % (NMOS) and

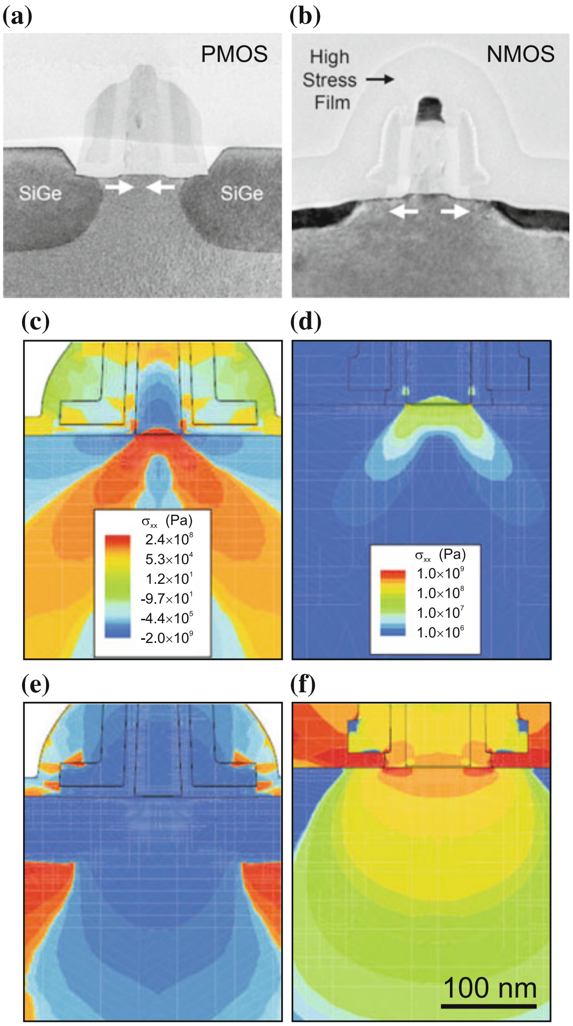


Fig. 24.37 Cross-section TEM images of strained (**a**) PMOS and (**b**) NMOS transistors. Adapted from [1784]. (**c**–**f**) Modelling of strain distribution: PMOS without (**c**) and with (**e**) $Si_{0.83}Ge_{0.17}$ pockets, NMOS without (**d**) and with (**f**) tensile cap layer. Adapted from [1785]

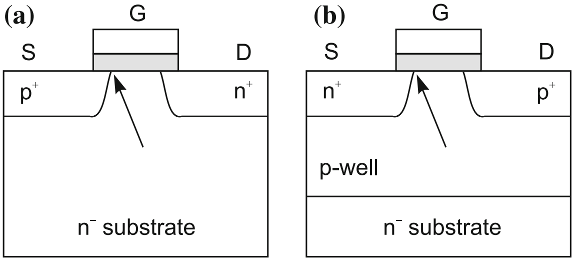


Fig. 24.38 Schematic of a (**a**) n-type and (**b**) p-type tunneling FET (TFET). D is reversely biased, i.e. positive for NTFET and negative for PTFET. The *grey areas* represent the gate oxide, the *arrows* denote the spatial position of tunneling (surface tunneling junction) for sufficient (NTFET: positive, PTFET: negative) gate voltage

50 % (PMOS) increase in $I_{\mathrm{D,sat}}$ compared to unstrained Si have been made in 65 nm transistors [1771].

The end of the miniaturization has been theoretically predicted many times and for various feature sizes. Today, only fundamental limits such as the size of an atom seems to limit circuit design.[10] Such limits (and the effects in nanostructures in the few-nm regime) will be reached beyond 2010, projected at about 2020. Up to then, it is probable that at least a few companies will follow the road map for further miniaturization, as laid out by the Semiconductor Industry Association[11] (SIA).

24.5.6 Tunneling FETs

A decisive parameter for FET performance is low leakage current. With shrinking device dimensions it increases rapidly for conventional FET design. A novel type of FET has thus been conceptualized, the tunneling FET (TFET) [1786]. It is a lateral p-i-n diode with a MOS gate (Fig. 24.38). The leakage current is minimized due to the reverse biased p-i-n structure. A low leakage current (per gate width) of less than 10^{-14} A/μm has been realized [1787, 1788]. The channel current is due to band-to-band tunneling as in an Esaki diode (Sect. 21.5.9) and can be controlled by the gate voltage [1789]. The surface tunneling junction is close to the source electrode. The use of germanium instead of silicon allows further performance enhancements [1790].

[10] Only commercial profit, rather than testing physical limits, drives the miniaturization. Insufficient economic advantages or low yield of further chip generations possibly can limit or slow down large-scale integration.

[11] www.semichips.org.

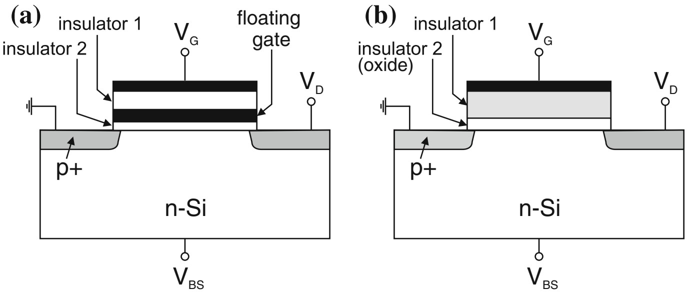


Fig. 24.39 MOSFET with (**a**) floating gate and (**b**) MIOS structure

24.5.7 Nonvolatile Memories

Floating Gate Memories

When the gate electrode of a MOSFET is modified in such a way that a (semi-)permanent charge can be stored in the gate, a nonvolatile electronic memory can be fabricated. In the floating-gate structure (Fig. 24.39a), an insulator–metal–insulator structure is used where charge is stored in the metal and cannot escape through the insulating barriers. The 'metal' is often realized by poly-Si. In the MIOS structure (Fig. 24.39b), the insulator–oxide interface is charged. The charge can be removed by UV light (EPROM, erasable programmable read-only memory) or by a sufficient voltage across the oxide at which the charge carriers can tunnel out (Fowler–Nordheim tunneling) (EEPROM, E^2PROM, electrically erasable programmable read-only memory).

Nowadays, a special type of EEPROM is used for the so-called *flash* memories. The stored gate charge causes a change in the MOSFET threshold voltage and is designed to switch between the on and off state. The storage time of the charge can be of the order of 100 years. Since tunneling limits the charge retention, the oxide must be sufficiently thick. In Fig. 24.40 a cross section of a 4 Gb, 73 nm SLC (single-level cell) flash memory is shown. The lower insulator (tunneling oxide at the channel) consist of 7.2 nm SiO_2, the upper insulator (insulator 1 in Fig. 24.39a) is a 18 nm thick oxide/nitride/oxide (ONO) stack. The floating gate has a $90 \times 90\,nm^2$ footprint, is about 86 nm high and consists of two polysilicon layers.

In a SLC memory the floating gate has two states, a certain charge value and the erased state. In a MLC (multi-level cell) the gate can store several charge states which can be sensed as different logic states, e.g. $2^2 = 4$ states. This increases the storage density, lowering cost per bit, but also increases the complexity. Typical endurance of SLC is at least 10^6 program–erase cycles. SLC cells so far have about ten times higher endurance (possible number of read–write cycles) and lower power consumption than MLC. Generally SLC flash memory is considered industrial grade and MLC flash is considered consumer grade. Recently also triple level cells (TLC),

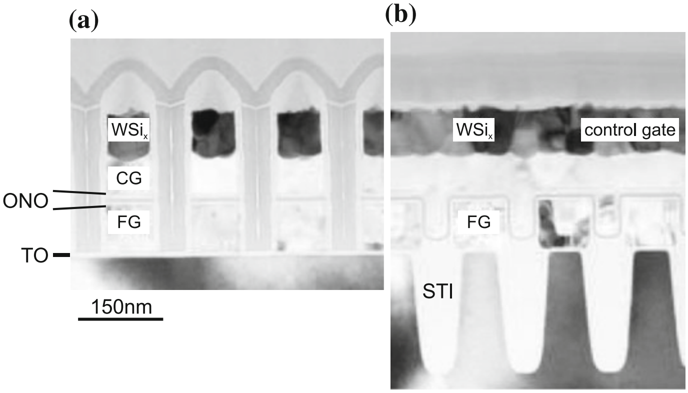


Fig. 24.40 Cross sections (**a**) perpendicular and (**b**) parallel to the control gate line of a 4 Gb, 73 nm SLC flash memory (Samsung K9F4G08U0M). 'CG' denotes the control gate, 'FG' the floating gate, 'TO' the tunneling oxide, 'ONO' the oxide/nitride/oxide insulator stack, and 'STI' the shallow trench insulation. Adapted from [1772]

storing 3 bit (8 states) are commercialized, however the increased storage density comes at high cost of reliability [1791].

The ultimate limit, explored currently, is to use a single electron charge to cause such an effect in the single-electron transistor (SET).

Future Concepts

Memory concepts beyond the storage of free charges include information storage via

- the static polarization in a ferroelectric material (either crystalline or polymer) (FeRAM [1792], Fig. 24.41a) which can be switched by an electric field.
- the phase change between amorphous and polycrystalline phases in a chalcogenide layer (typically GeSb [1793] or $Ge_2Sb_2Te_5$, GST [1794, 1795] with an $\alpha \leftrightarrow$ c transition, Fig. 24.42) upon local heating (similar to a rewritable DVD) and the related change is resistivity (PCM, phase change memory).
- the storage of magnetization direction (MRAM [1796, 1797]) and subsequent resistance change of a magneto-tunneling junction (MTJ) whose resistance depends on the relative magnetization (parallel or perpendicular) of two magnetic layers separated by a thin tunneling insulator (Fig. 24.41b). The largest TMR (tunnel-magnetoresistance) effect has been achieved with MgO as insulator [1797]. The magnetization of the bottom magnetic layer of the MTJ is fixed. The magnetization directions ±45° are written into the free layer with the magnetic fields of two perpendicular high current wires in two subsequent back-end interconnect layers sandwiching the MTJ.
- resistance change based on solid electrolytes (PMC, programmable metallization cell memory). The lowering of the resistance is attained by the reduction of ions in a fairly high resistivity electrolyte (e.g. from the system Cu,Ag–Ge–Se,S,O [1798, 1799] or oxides [1800]) to form a conducting bridge between the electrodes. The

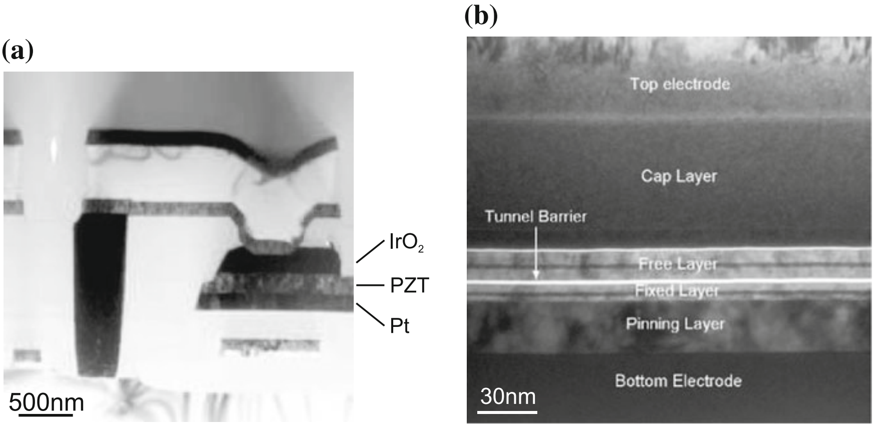


Fig. 24.41 (**a**) Cross-section TEM image of a cell from a Ramtron 4 Mb FeRAM. The information is stored in the electric polarization of a polycrystalline $Pb(Ti_xZr_{1-x})O_3$ (PZT) island, contacted on the bottom and top with platinum and iridium oxide, respectively. Adapted from [1772]. (**b**) Cross-section TEM of the magnetic tunneling junction from a Freescale 4.2 Mb MRAM, located between the M4 and M5 interconnect layers. The magnetization of the free layer can be switched, that of the fixed layer remains constant. Adapted from [1772]

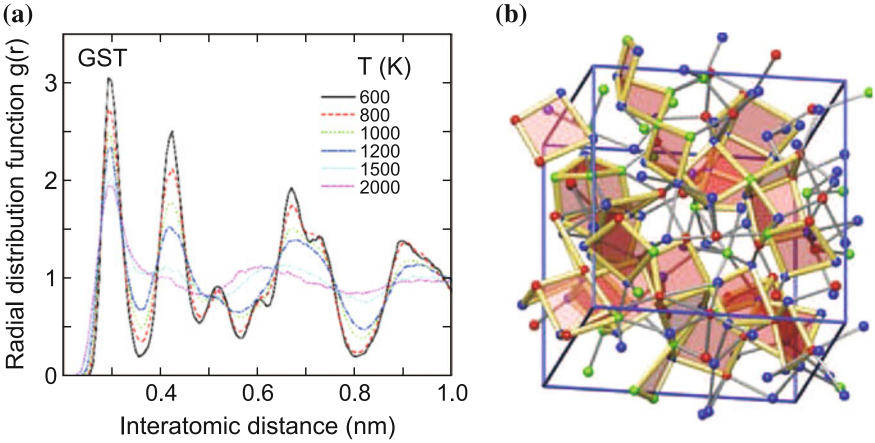


Fig. 24.42 (**a**) Radial distribution function of ions in $Ge_2Sb_2Te_5$ (GST) for various temperatures (cmp. Fig. 3.14b). Adapted from [756]. (**b**) Atom arrangement in the amorphous phase of GST with square units highlighted that nucleate crystallization. Adapted from [1795]

resistance is returned to the high value via the application of a reverse bias that results in the breaking of the conducting pathway.

- resistance change in transition metal oxides such as perovskites, e.g. $SrTiO_3$:Cr [1801, 1802] or NiO:Ti (RRAM). Electrical pulses of opposite polarity switch the resistance reversibly between a high- and a low-resistance state. Oxygen-vacancy drift modulates the valence of the mixed-valence transition-metal ion (e.g. Ti^{3+}–Ti^{4+}) and thus the conducting state [1803].

- a molecular configuration change (e.g. redox reaction) between crossed wire lines (molecular electronics [1804–1806]).

24.5.8 Heterojunction FETs

Several types of field-effect transistors have been devised that use heterojunctions (HJFET).

HIGFET

As conducting channel, the two-dimensional electron gas at an undoped heterointerface is used. Such a transistor is called a heterojunction insulating gate FET (HIGFET). With forward or backward gate voltage, an electron or hole gas can be created (channel enhancement mode), as visualized in Fig. 24.43. Thus, a complementary logic can be realized. However, the p-channel suffers from low hole mobility.

HEMT

If the top wide-bandgap layer is n-doped, a modulation-doped FET (MODFET) is made (see Sect. 12.3.4). This structure is also called a HEMT (high electron mobility transistor) or TEGFET (two-dimensional electron gas FET) (Fig. 24.44). A thin undoped AlGaAs spacer layer is introduced between the doped AlGaAs and the

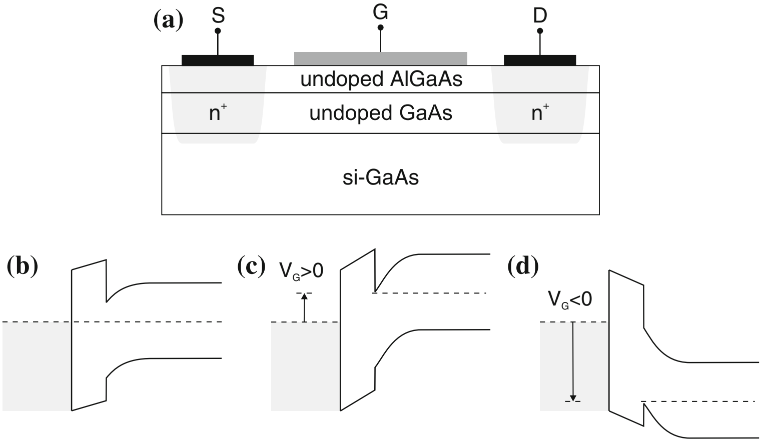


Fig. 24.43 (**a**) Scheme of a HIGFET structure with metal gate and undoped AlGaAs/GaAs heterointerface on semi-insulating GaAs. The source and drain contacts are n-doped such that this structure can be used as an n-HIGFET (see part (**c**)). (**b**) Band diagram for zero gate voltage. (**c**) Band diagram for positive gate voltage and n-channel, (**d**) for negative gate voltage and p-channel

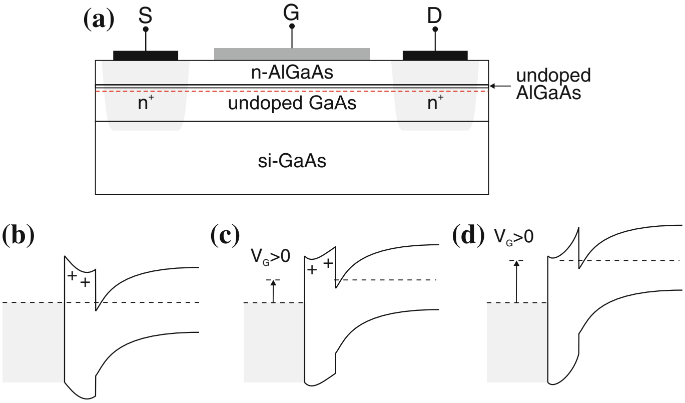


Fig. 24.44 (**a**) Scheme of a HEMT structure with n-AlGaAs/GaAs heterointerface on semi-insulating GaAs. The source and drain contacts are n-doped such that this structure can be used as an n-channel (normally-on) HEMT. The *horizontal dashed line* represents schematically the position of the 2DEG at the heterointerface on the GaAs side. (**b**) Band diagram at zero gate voltage. (**c**) Band diagram at positive gate voltage, increase of channel carrier concentration. (**d**) Band diagram at even larger positive gate voltage, formation of conducting channel in the AlGaAs layer

undoped GaAs to reduce impurity scattering from carriers that tunnel into the barrier. With increasing gate voltage, a parallel conduction channel in the AlGaAs is opened. The natural idea would be to increase the Al fraction in the AlGaAs to increase the quantum-well barrier height. Unfortunately, the barrier height is limited to 160 meV for an aluminum concentration of about 20 %. For Al content higher than about 22 %, the DX center (cf. Sect. 7.7.6) forms a deep level such that the apparent ionization energy increases drastically and no shallow donors can be used for modulation doping. An improvement for the barrier conduction problem is the use of δ-doping [1807], i.e. the introduction of a highly doped thin (mono-)layer (Fig. 24.45), which results in higher channel carrier concentration.

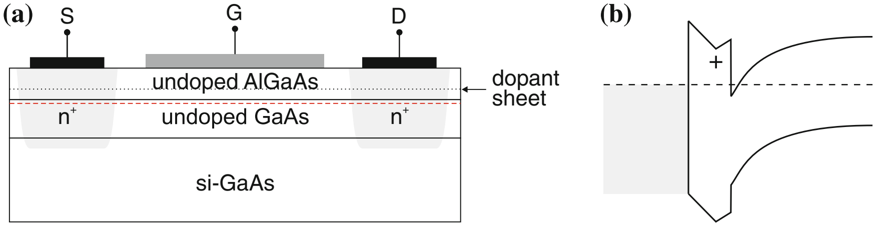


Fig. 24.45 (**a**) Scheme of a δ-doped HEMT structure with AlGaAs/GaAs heterointerface on semi-insulating GaAs. The source and drain contacts are n-doped such that this structure can be used as an n-channel HEMT. The *horizontal dashed line* represents schematically the position of the 2DEG in the GaAs layer. (**b**) Band diagram at zero gate voltage

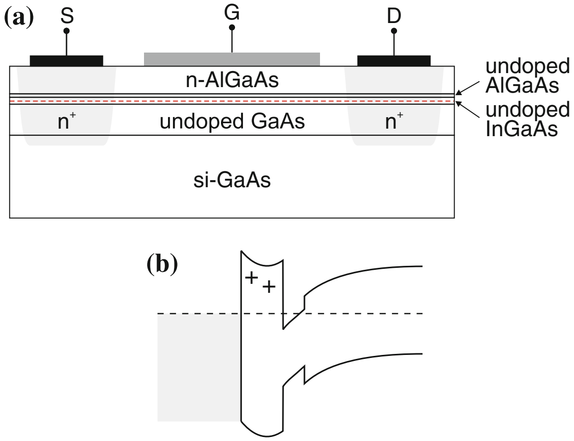


Fig. 24.46 (**a**) Scheme of a PHEMT structure with n-AlGaAs/InGaAs heterointerface on semi-insulating GaAs. The source and drain contacts are n-doped such that this structure can be used as an n-channel HEMT. The *horizontal dashed line* represents schematically the position of the 2DEG in the InGaAs layer. (**b**) Band diagram at zero gate voltage

Pseudomorphic HEMTs

Instead of increasing the height of the barrier, the depth of the well can be increased by using a low-bandgap material. On GaAs substrate, InGaAs is used (Fig. 24.46). However, strain is introduced in this case and the InGaAs layer thickness is limited by the onset of dislocation formation (cf. Sect. 5.4.1) (which reduces the channel mobility and the device reliability). For $In_{0.15}Ga_{0.85}As$ (thickness about 10–20 nm), a total barrier height of about 400 meV can be obtained. A barrier height of 500 meV can be reached with an InAlAs/InGaAs structure on InP (Fig. 24.47). The InAlAs does not suffer from the problem related to DX centers. The channel indium concentration is typically 50 %. The mobility increases with increasing indium concentration. This InP-based HEMT structure is widely used in satellite receivers for its excellent high-speed and low-noise properties in the 100–500 GHz range and beyond.

However, the InP technology is economically less favorable than GaAs due to smaller available substrate size and higher cost (2001: 4″ InP substrate: $1000, 6″ GaAs substrate: $450).

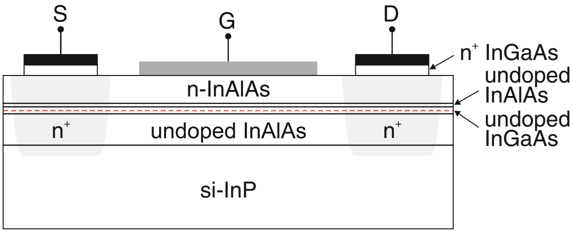


Fig. 24.47 Scheme of a PHEMT structure with n-AlInAs/InGaAs/InAlAs structure on semi-insulating InP. The source and drain contacts (with a highly doped InGaAs contact layer) are an n-doped such that this structure can be used as an n-channel HEMT. The *horizontal dashed line* represents schematically the position of the 2DEG in the InGaAs layer

Metamorphic HEMTs

A unification of the InAlAs/InGaAs structure with the best figure of merit and the GaAs substrate is achieved with the metamorphic HEMT (MHEMT). Here, a relaxed buffer is used to bring the in-plane lattice constant from that of GaAs to about that of InP. It is key that the defects occurring are confined to the relaxed buffer and do not enter the active device structure (see Fig. 24.48). The relaxed buffer is typically about 1 µm thick. It can be grown, e.g. with a graded $\mathrm{In}_x(\mathrm{Ga,Al})_{1-x}\mathrm{As}$ layer with $x = 0$–42 % or with a stepped structure with piecewise constant indium concentration in each layer. It is important that a smooth interface of the channel is achieved in order to avoid additional scattering mechanisms. For high-frequency operation, the fabrication of a small gate length is important, as shown in Fig. 24.49 for a 70-nm gate of a $f_\mathrm{T} = 293$ GHz, $f_\mathrm{max} = 337$ GHz transistor [1808]. SiGe channels, providing higher mobility than pure Si, can be fabricated using graded or stepped SiGe buffer layers on Si substrate. With such Si-based MHEMTs frequencies up to 100 GHz can be achieved.

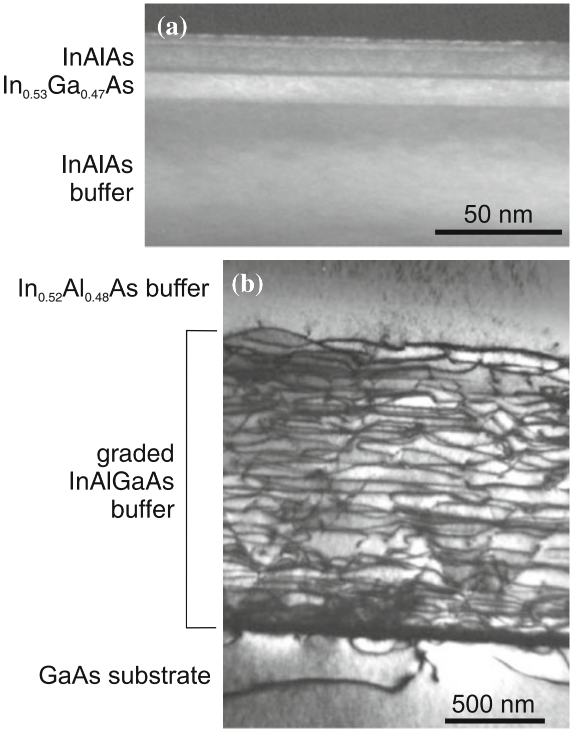

(a)
InAlAs
$\mathrm{In}_{0.53}\mathrm{Ga}_{0.47}\mathrm{As}$
InAlAs buffer
50 nm
(b)
$\mathrm{In}_{0.52}\mathrm{Al}_{0.48}\mathrm{As}$ buffer
graded InAlGaAs buffer
GaAs substrate
500 nm

Fig. 24.48 Cross-sectional TEM image of an InAlAs/InGaAs MHEMT: (**a**) Active layer with rms surface roughness of 2.0 nm (from AFM), (**b**) graded InGaAlAs buffer layer (1.5 µm) on GaAs substrate. Adapted from [1809]

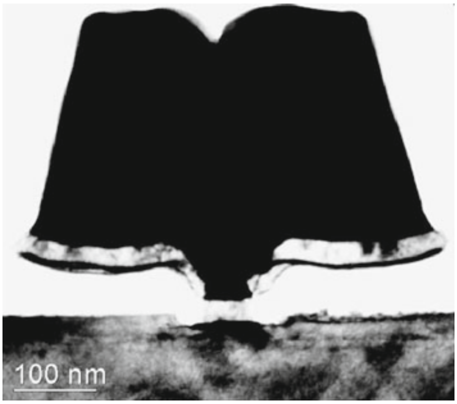


Fig. 24.49 Cross-sectional TEM image of the 70-nm gate of an InAlAs MHEMT on GaAs substrate. From [1808]

24.6 Thin-Film Transistors

Thin-film transistors (TFTs) are field-effect transistors with a channel formed as thin film on insulating substrate. A detailed treatment is available in [1810]. TFTs are typically fabricated as large-area arrays from thin layers of polycrystalline or amorphous silicon [1811] or organic semiconductors [1812–1815] on cheap substrates such as glass. Their most prominent use is driving pixels in active-matrix displays such as electroluminescence (EL) displays or twisted nematic liquid crystal displays (LCD) [1816]. Various gates and gate geometries have been reported as depicted in Fig. 24.50.

24.6.1 Annealing of Amorphous Silicon

Since the mobility in polycrystalline silicon is much higher (up to several hundred cm^2/Vs depending on grain size, see Sect. 8.3.8) than in amorphous silicon (<1 cm^2/Vs), such material is much more desirable as channel in TFTs. However, it

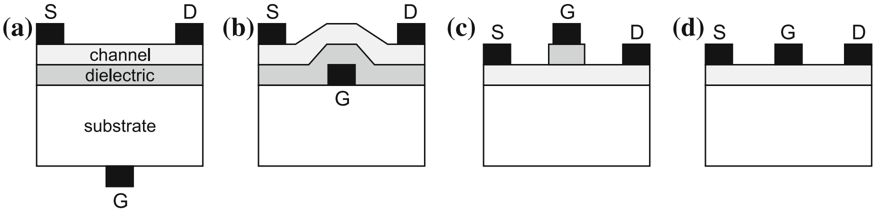


Fig. 24.50 Schematic geometries of TFTs: (**a**–**c**) MISFETs, (**d**) MESFET with (**a**, **b**) bottom gate and (**c**, **d**) top gate. Semiconductor channel layer (*light grey*), insulating dielectric (*dark grey*) and metals (*black*)

requires high deposition temperatures. In order to achieve polycrystalline silicon with large grain size from amorphous silicon films that can be deposited at low temperature (down to room temperature) several schemes have been developed, the most important being thermal annealing and (excimer) laser annealing (ELA). Crystallization occurs by thermally activated nucleation and growth processes [1817]. Polycrystalline layers will small grain size can be made amorphous with implantation of Si (self-implantation) and a subsequent optimized (re-)crystallization processes.

In laser annealing energy is locally introduced during short pulses (several 10 ns or even fs); subsequent material change occurs on a sub-µs time scale [1818]. Laser induced crystallization enables the use of inexpensive low-temperature substrates, such as plastic or glass, since it involves the ultrafast melting and resolidification of the near-surface region of the sample, and minimal heating of the substrate. Local processing is also possible using laser crystallization.

In Fig. 24.51 the effect of thermal annealing of amorphous silicon is shown. The annealing time necessary to convert the amorphous phase completely to polycrystalline, e.g. 10 h at 640 °C, depends largely on temperature as detailed in [1819, 1820] (Fig. 24.51a) with a large activation energy of 3.9 eV. Also the final grain size is temperature dependent (Fig. 24.51b).

The introduction of certain metals like Pd [1821], Al [1822], Au [1823] or Ni [1824] induces crystallization and allows for much lower annealing temperatures. Pd and Ni create silicides that play an important role for the grain nucleation or growth front. Au and Al are solved in the bulk but have a similar effect. For example, using Pd complete crystallization of a 150 nm thick a-Si film deposited at 480 °C can be achieved by thermal annealing after 10 h at only 500 °C [1825] (using metal-induced lateral crystallization, MILC).

24.6.2 TFT Devices

A schematic cross section of an amorphous silicon-based TFT is shown in Fig. 24.52a. Carriers in amorphous silicon have a low mobility typically less than 1 cm^2/Vs [1826, 1827]. As-grown polycrystalline silicon has a mobility of typically less than 10 cm^2/Vs. With the use of laser irradiation or thermal annealing, amorphous or small-grain polycrystalline silicon layers can be recrystallized, increasing the mobility up to several 100 cm^2/Vs, improving transistor performance [1826, 1828, 1829]. However, for display applications a mobility of 10 cm^2/Vs is large enough.

The main optimization criteria for thin-film transistors are high on-off ratio, long-term stability, good uniformity and reproducibility, and low cost. Recently, flexible (on polymer substrate) and transparent TFTs (TFET, transparent FET), e.g. with polycrystalline ZnO or GaInZnO (GIZO) channel (Fig. 24.52b), are being investigated for advanced applications such as all-transparent electronics and displays [1398, 1828–1832]. A compilation of recent results on transparent semiconducting oxide (TSO) channel FETs can be found in [1834]. In Fig. 24.53 performance data for

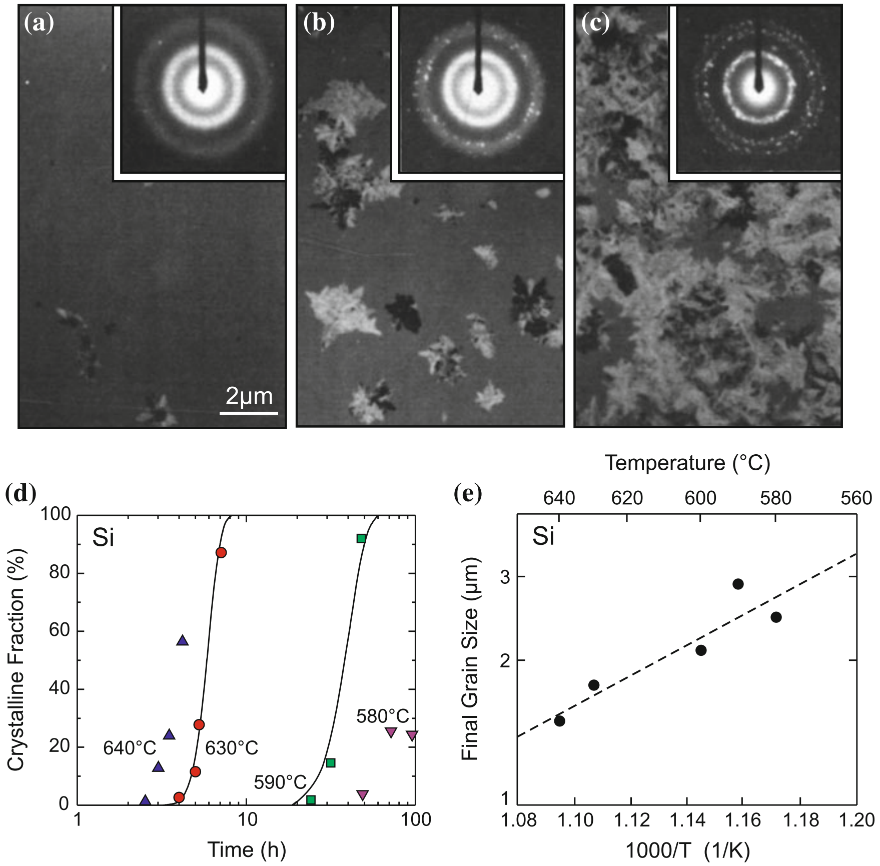


Fig. 24.51 Thermal annealing of 100 nm thick film of amorphous silicon (fabricated from LPCVD and amorphized by 100 keV Si^+ implantation with a dose of $5 \times 10^{15}\,cm^{-2}$). TEM images and diffraction patterns (*insets*) for amorphized Si after (**a**) 4 h, (**b**) 5.25 h and (**c**) 7.1 h annealing at $T = 630\,°C$. The crystalline fractions are 2, 28 and 87 %, respectively. (**d**) Crystalline fraction as a function of annealing time for various annealing temperatures as labeled. *Symbols* are experimental data, *solid lines* depict theory considering grain nucleation and growth. (**e**) Final grain size for various annealing temperature. *Dashed line* is exponential with a slope of 0.6 eV. Adapted from [1819]

various TSO channel FETs are visualized. In Fig. 24.54 a transparent inverter based on ZnO-MESFETs is depicted [1835].

24.6.3 OFETs

Organic field effect transistors (OFETs) [1837–1840] are transistors for which at least the channel consists of an organic material. Most work is done on thin film

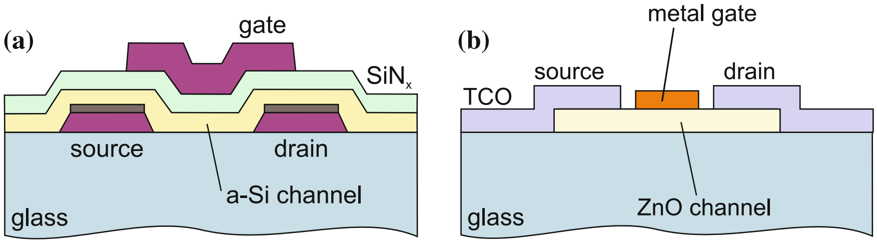


Fig. 24.52 (**a**) Schematic cross section of a top-gate amorphous silicon (a-Si) thin-film transistor (MISFET) on glass substrate. (**b**) Schematic cross section of a transparent ZnO thin-film transistor (MESFET)

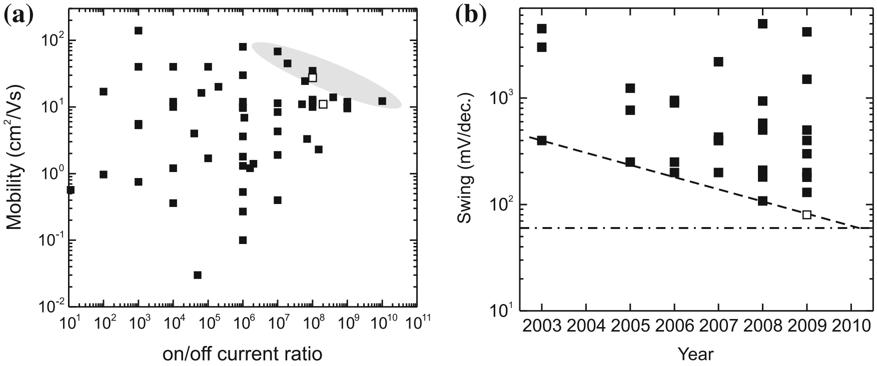


Fig. 24.53 (**a**) Field effect mobility and on/off current ratio for oxide channel transistors. *Filled squares* represent MISFET transistors; *open squares* are for MESFETs from [1833]. The *shaded area* indicates best performance. (**b**) Voltage swing for MISFETs (*filled squares*, subthreshold voltage swing) and MESFET (*empty square*, above turn-on voltage from [1833]) with TSO channels. The *dashed line* is guide to the eyes for the trend of best performance. The *dash-dotted line* indicates the thermodynamic limit of about 60 meV/decade for the swing [1836]. Adapted from [1834]

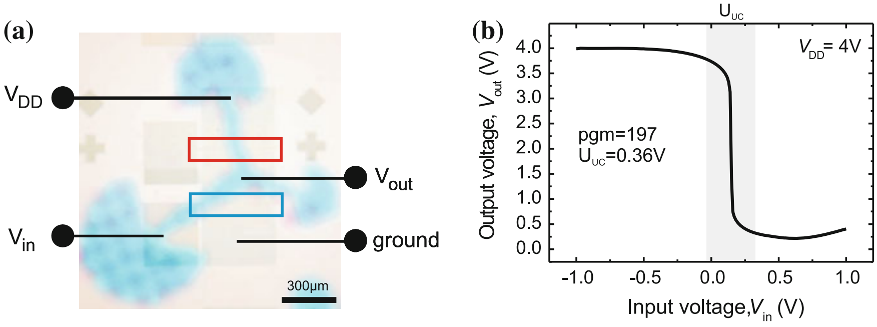


Fig. 24.54 (**a**) Optical image of transparent MESFET inverter based on ZnO. The *two rectangles* indicate the two gates. (**b**) Transfer characteristic for supply voltage $V_{DD} = 4\,V$

transistors, although some work on OFETs using bulk organic semiconductors has been reported [1841, 1842]. Organic materials are also used for the insulator and the contact materials. Often organic and flexible substrates are used. Applications are in low cost electronics, e.g. for driving display pixels or RFID tags (typically operating at 13.56 or 900 MHz). Processes like spin-on and printing can be used. Due to their larger chemical stability against oxidation, mostly p-type channel materials are used. The highest mobilities are reached for pentacene (6 cm^2/Vs) and sexithiophene (1 cm^2/Vs); n-type organic semiconductors exhibit field mobility below 0.1 cm^2/Vs [1838].