Plan

— **Bipolar junction transistor**

- homojunction and heterojunction
- Doping considerations
- Transport factor and current gain
- Frequency dependence of gain in a microwave transistor
- Base shrinkage and finite output conductance
- Ebers-Moll model and small-signal equivalent circuit

— **Elements of small-signal analysis**

- Two port; common terminal configurations
- Power gain, Mason theorem

— **Transistor Principles: PETs and FETs**

— **Field effect transistor**

- Two-dimensional channel; sheet conductance
- Conduction Laws by dimensional analysis
- Gated diode
- Short channel effects
- MOS structure: relevant energies
- Band bending; threshold evaluation
- Inversion layer structure
- Pinch-off effect in MOSFET
- Concept of quasi-Fermi level (imref)

— **Discussion**

- Why silicon ???

Junction Transistor

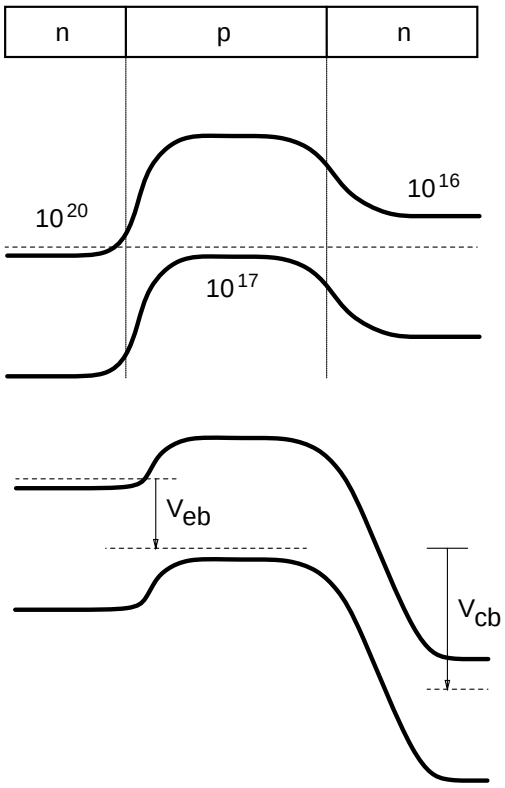


Figure 1: Schematic diagram of a *npn* transistor in equilibrium and under applied bias. By Kirchhoff law:

$$I_E = I_C + I_B \approx I_C$$

Neglecting recombination in the base and parasitic injection of holes into emitter,† the collector current flows through a much larger impedance than the emitter current, ***whence the power gain.***

---

† In a homojunction transistor, injection of holes into the emitter is suppressed compared to the useful injection of electrons into the base only by the factor $\frac{n_{p0}}{p_{n0}} = \frac{N_D(\text{emitter})}{N_A(\text{base})}$.
This means that the base must be lightly doped compared to the emitter and hence base resistance is a concern.
The fundamental trade-off: thicker base for lower base resistance, thinner base for faster diffusion across the base.

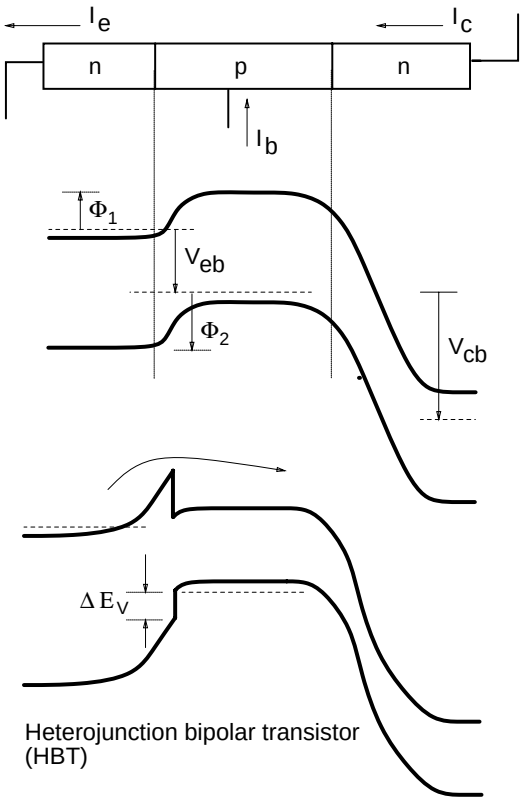


Figure 2: Homojunction and heterojunction *npn* transistor

In homojunction transistor at the base-emitter junction

$$J_E^{(e)} \propto e^{-\beta\Phi_1}$$

$$J_E^{(h)} \propto e^{-\beta\Phi_2}$$

$$\frac{J_E^{(h)}}{J_E^{(e)}} = \frac{N_A(\text{base})}{N_D(\text{emitter})}$$

Emitter efficiency $\eta \equiv \dfrac{J_E^{(e)}}{J_E^{(e)} + J_E^{(h)}} \approx 1 - \text{doping ratio}.$

In heterojunction transistor:

$$\frac{J_E^{(h)}}{J_E^{(e)}} = \frac{N_A(\text{base})}{N_D(\text{emitter})} \times e^{-\beta\Delta E_V}$$

Back to (homo)junction transistor.

We understand why the base doping must be much lower than emitter doping. Now why the collector doping must be much lower than emitter doping ?

$$N_{\mathrm{D}}(\text{emitter}) \gg N_{\mathrm{A}}(\text{base}) \gg N_{\mathrm{D}}(\text{collector})$$

Threefold answer:

— For W_{B} to have little dependence on V_{cb} (need high output impedance)

— to lower base-collector capacitance C_{cb}

— to lower the field in base-collector junction (increase breakdown voltage)

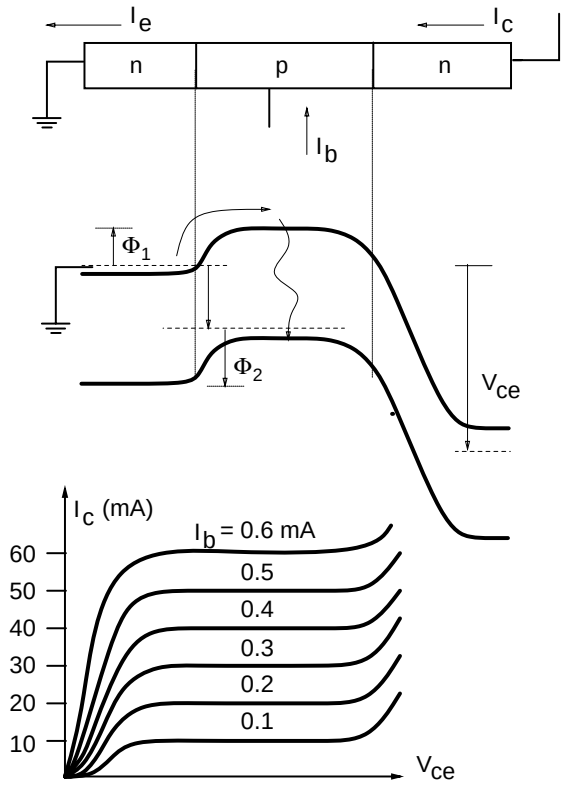


Figure 3: Common-emitter transistor characteristics. Base current is stepped up by increment of 0.1 mA and the emitter current increases by much larger amounts. Here current gain $\beta \approx 100$.

Base transport factor and current gain.

From the continuity equation:

$$n'' - \frac{n}{L_D^2} = 0$$

$$n(x) = A\, e^{\frac{x}{L_D}} + B\, e^{-\frac{x}{L_D}}$$

Boundary conditions at x and at W:

$$n(0) = \frac{n_i^2}{N_A} e^{\beta V_{eb}}$$

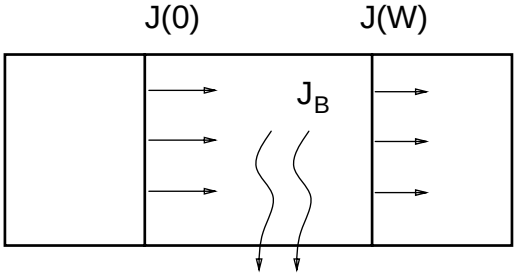
$$n(W) = \frac{n_i^2}{N_A} e^{\beta V_{cb}} \approx 0$$

$$\Rightarrow \quad n(x) = A \sinh\left[(W-x)/L_D\right] = \frac{n(0)\sinh\left[(W-x)/L_D\right]}{\sinh(W/L_D)}$$

$$\Rightarrow \quad J_n(x) = e D \frac{\partial n}{\partial x} = \frac{e D n(0)}{L_D} \frac{\cosh\left[(W-x)/L_D\right]}{\sinh(W/L_D)}$$

The base transport factor alpha:

$$\alpha \equiv \frac{J_n(W)}{J_n(0)} = \frac{1}{\cosh(W/L_D)} \approx 1 - \frac{W^2}{2L_D^2}$$



Current gain: $\beta = \frac{\partial I_C}{\partial I_B}$. By Kirchhoff's law, $\beta = \frac{\alpha}{1-\alpha}$

Combined with emitter efficiency η, we take instead of α the product $\alpha\eta$.

Frequency dependence of the current gain.

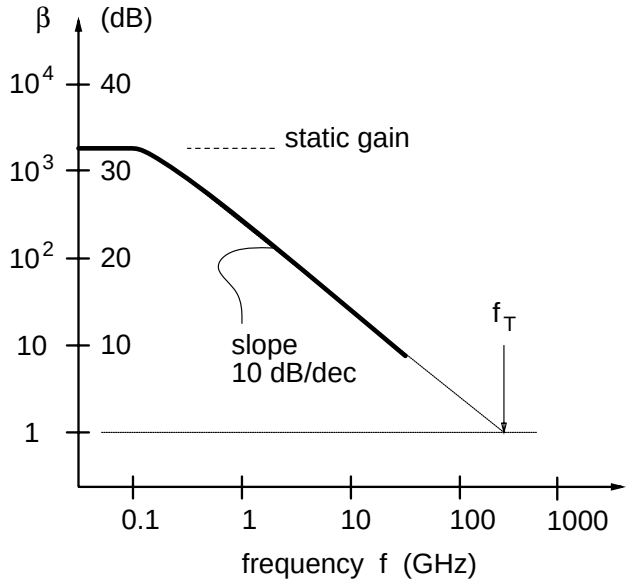


Figure 5: High-frequency gain. At high frequencies the current gain rolls-off at 10 dB/per decade (i.e. as $1/f$). This behavior is quite universal and has nothing to do with either recombination or emitter efficiency.

The characteristic cut-off frequency f_T is defined by the condition of unity current gain ($\beta \to 1$) and is mainly due to the propagation delay τ of minority carriers† through the base.

$$f_T = \frac{1}{2\pi\tau}$$

In general, $\tau = W/v$, where v is the average velocity of minority carrier propagation. For diffusive transport,

$$\tau_D = \frac{W^2}{2D} \ll \tau_C \ .$$

† Do not confuse this τ with the minority carrier lifetime (which is typically much longer). Let us denote the minority-carrier lifetime by τ_C ("capture" time).

Let us derive an expression for $\alpha(f)$. Begin with the continuity equation:

$$\frac{\partial n}{\partial t} = D \frac{\partial^2 n}{\partial x^2} - \frac{n}{\tau_C}$$

Take

$$n(t) = n_0 + \delta n \, e^{i\omega t}$$

where $n_0 = n_0(x)$ is the static solution and $\delta n = \delta n(x)$ is the harmonic variation amplitude at frequency $\omega \equiv 2\pi f$.

Both the static equation for n_0 and the dynamic equation for δn are of similar form:

$$\frac{\partial^2 n_0}{\partial x^2} - L_D^{-2} n_0 = 0$$
$$\frac{\partial^2 \delta n}{\partial x^2} - L^{-2} \delta n = 0$$

where

$$L_D^2 \equiv D \tau_C$$
$$L^{-2} = L_D^{-2}(1 + i\omega\tau_C) \underset{\omega\tau_C \gg 1}{\approx} \frac{i\omega}{D}$$

The solutions to both equations are similar too (in form):

$$n_0(x) = A \sinh\left[\frac{W - x}{L_D}\right]$$
$$\delta n(x, \omega) = B \sinh\left[\frac{W - x}{L}\right]$$

whence we find (in analogy to $\alpha_0 = \frac{1}{\cosh(W/L_D)}$

$$\alpha(\omega) \equiv \frac{\partial J(W, \omega)}{\partial J(0, \omega)} = \frac{1}{\cosh(W/L)}$$

At sufficiently high ω's (nothing "spectacular", just $\omega\tau_C \gg 1$) we have

$$\frac{W}{L} = \sqrt{2i\omega\tau_D} \quad \text{where} \quad \tau_D = \frac{W^2}{2D}$$

$$= \sqrt{2i\omega\tau_D} \quad \text{remember} \quad \sqrt{i} = e^{i\pi/4} = \frac{1+i}{\sqrt{2}}$$

$$\alpha(\omega) = \frac{1}{\cosh(W/L)} = \begin{cases} (1+i\omega\tau_D)^{-1} & \text{for} \quad \omega\tau_D \ll 1 \\ 2e^{-\omega\tau_D} e^{-i\omega\tau_D} & \text{for} \quad \omega\tau_D \gtrsim 1 \end{cases}$$

Im modern transistors, typically, $W \lesssim 1{,}000\,A$ and $D \sim 50\ \text{cm}^2\,\text{s}^{-1}$. Therefore, at frequencies easily accessible to the measurement (up to, say 20 GHz) one has, typically, $\tau_D \lesssim 10^{-12}$ s and $\omega\tau_D \ll 1$.

$$\alpha(\omega) \approx \frac{1}{1+i\omega\tau_D} \approx e^{-i\omega\tau_D}$$

$$\beta(\omega) = \frac{\alpha}{1-\alpha} = \frac{e^{-i\omega\tau_D}}{1-e^{-i\omega\tau_D}} = \frac{-i\,e^{-i\omega\tau_D/2}}{2\sin(\omega\tau_D/2)}$$

$$|\beta(\omega)| = \frac{1}{2\,|\sin(\omega\tau_D/2)|} \approx \frac{1}{\omega\tau_D}$$

Uisng heterostructures, it is possible to design an HBT such that $\alpha(\omega)$ does not spiral in significantly, even for large ω's – remaing close to the circle $\exp(i\omega\tau)$ for phase angles $\phi = \omega\tau$ as large as $\phi = 2\pi$. Such ***coherent transistors***,† are capable of "life after death", showing current gain above f_T and power gain above $f_{\max}$.

---

† A. A. Grinberg and S. Luryi, "Coherent transistor", *IEEE Trans. Electron Devices* **ED-40**, pp. 1512-1522 (1993).

S. Luryi, A. A. Grinberg, and V. B. Gorfinkel, "Heterostructure bipolar transistor with enhanced forward diffusion of minority carriers", *Appl. Phys. Lett.* **63**, pp. 1537-1539 (1993).

Base shrinkage and finite output conductance.

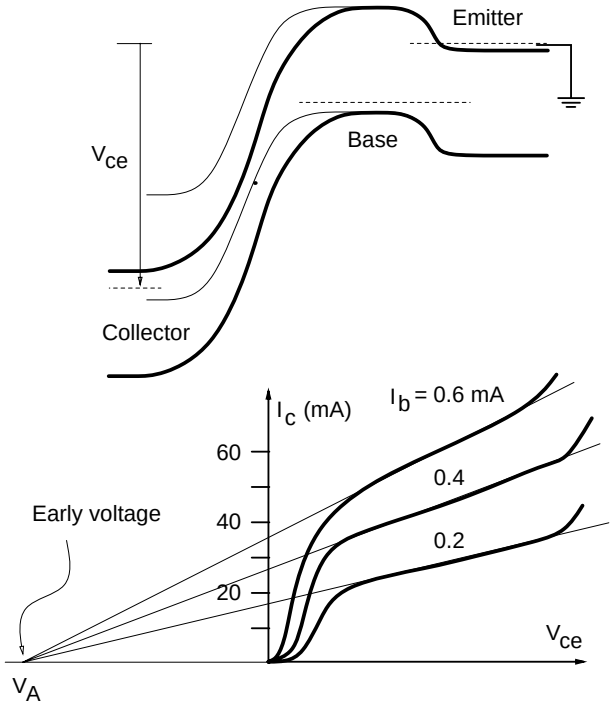


Figure: Base shrinkage and Early effect.

Neglecting recombination, the injected minority current must be constant in the base, $\vec{\nabla} \cdot \vec{J} = 0$. If the current is by diffusion only, then

$$\frac{\mathrm{d}n}{\mathrm{d}x} = \text{const} = \frac{n(0) - n(W)}{W} \approx \frac{n_{p0}\, \mathrm{e}^{\beta V_{\mathrm{eb}}}}{W}$$

W shrinks with the collector bias. Hence, at a fixed base current, one has an increasing collector current (Early effect)

$$I_{\mathrm{C}} \propto \left[1 + \frac{V_{\mathrm{cb}}}{V_{\mathrm{A}}}\right]$$

Finite output conductance is detrimental; we live better without it.

The Early voltage depends on the base width W and the Gummel number $n_{\mathrm{G}} = N_{\mathrm{A}} W$. For $W \gg$ than the BC junction depletion width, one has

$$V_{\mathrm{A}} \approx \frac{e\, N_{\mathrm{A}} W^2}{\varepsilon} \equiv \frac{e\, n_{\mathrm{G}} W}{\varepsilon}$$

Ebers-Moll model

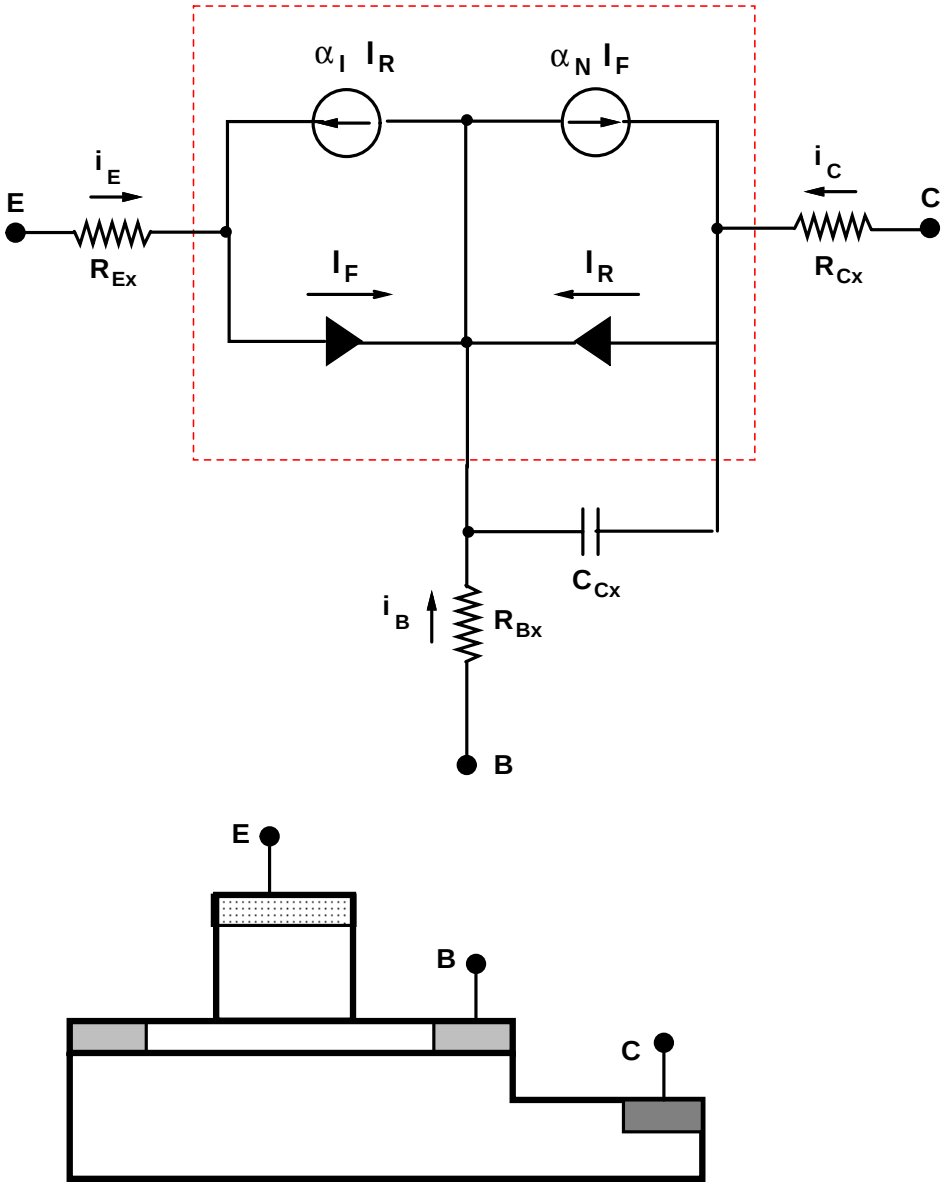


Figure 1: Dashed line indicates the "intrinsic" portion of the device, excluding "parasitic" extrinsic elements.

The model for the diode blocks can be further specified to include the internal junction capacitance.

Small-signal equivalent circuit of an abrupt junction HBT

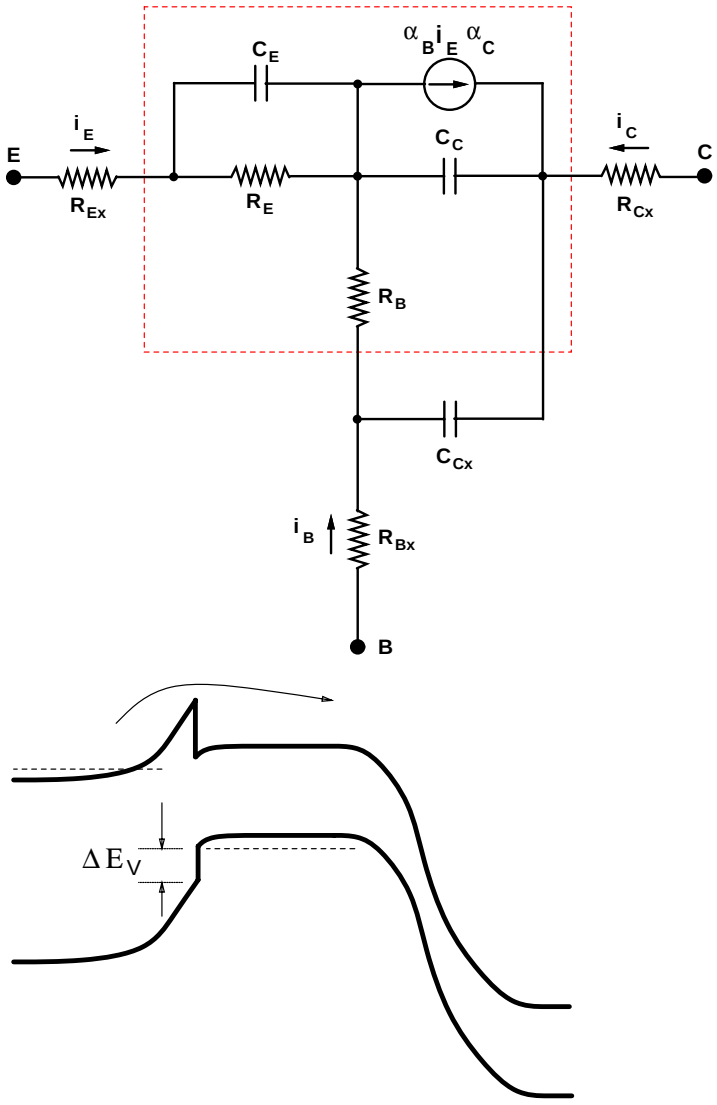


Figure 2: This model is good for "ballistic" propagation of carriers across the base. For diffusive propagation, the intrinsic portion must be adjusted,†

† A. A. Grinberg and S. Luryi, *IEEE Trans. Electron Devices* **ED-40**, pp. 1512-1522 (1993).

Elements of small-signal analysis

All variables $A(t)$ are considered varying ***harmonically*** in a small range about a dc point:

$$A(t) = A_0 + \delta A\, e^{i\omega t}$$

$$I(t) = I_0 + \delta I\, e^{i\omega t}$$

$$V(t) = V_0 + \delta V\, e^{i\omega t}$$

Many alternative notations, e.g., i and v instead of δI and δV.

The δA 's are ***complex*** quantitites, may be position-dependent fields, $\delta A(\vec{x})$.

The relationship between different δA 's, e.g. between δV and δI (generalized impedances or admittances) depend on the chosen dc point.

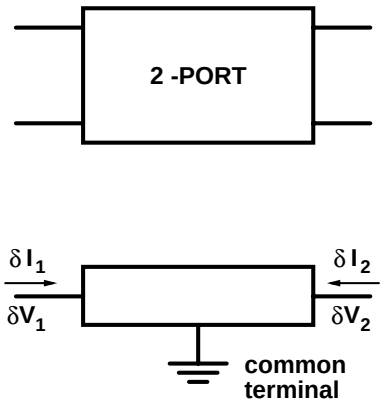


Figure 3: General two-port. Transformers are 2-ports ("passive"). From the small-signal point of view, transistors are two-port amplifiers.

Admittance matrix:

$$\begin{bmatrix} \delta I_1 \\ \delta I_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} \delta V_1 \\ \delta V_2 \end{bmatrix}$$

Admittance matrix:

$$\begin{bmatrix} \delta I_1 \\ \delta I_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} \delta V_1 \\ \delta V_2 \end{bmatrix}$$

For example:

$$y_{11} \equiv \left[\frac{\mathrm{d}I_1}{\mathrm{d}V_1} \right]_{V_2}, \text{ input admittance}$$

Impedance matrix:

$$\begin{bmatrix} \delta V_1 \\ \delta V_2 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} \begin{bmatrix} \delta I_1 \\ \delta I_2 \end{bmatrix}$$

For example:

$$z_{22} \equiv \left[\frac{\mathrm{d}V_2}{\mathrm{d}I_2} \right]_{I_1}, \text{ output impedance}$$

Hybrid matrix (h-parameters):

$$\begin{bmatrix} \delta V_1 \\ \delta I_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} \delta I_1 \\ \delta V_2 \end{bmatrix}$$

For example:

$$h_{21} \equiv \left[\frac{\mathrm{d}I_2}{\mathrm{d}I_1} \right]_{V_2}, \text{ forward current gain}$$

Definition of these parameters essentially involves specification of the **boundary condition** at one or another port. Thus

h_{22} is the output admittance for ***open-circuit*** input port.

y_{22} is the output admittance for ***short-circuit*** input port.

h_{21} is the forward current gain for ***short-circuit*** output port, etc.

Each set of parameters (z-parameters, y-parameters, h-parameters) is ***complete*** in the sense that it can be used to derive the other sets unambiguously.

Common terminal configurations

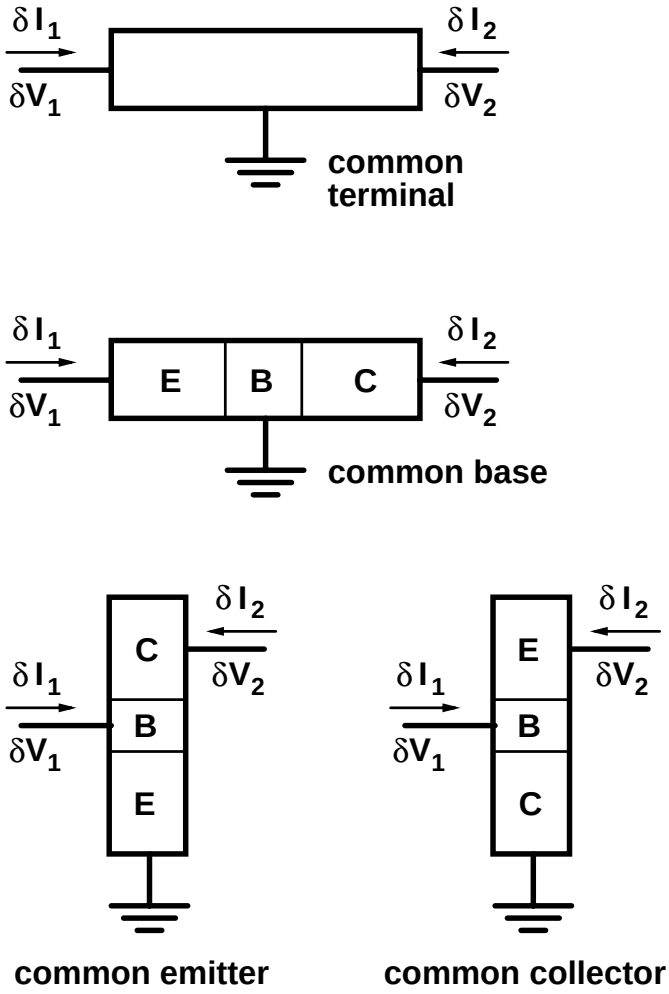


Figure 4: Different common-terminal configurations give rise to very different parameters. Thus, the short-circuit current gains are

$$h_{21}^{e} \equiv \beta \qquad h_{21}^{b} \equiv \alpha$$

and hence

$$h_{21}^{e} = \frac{1 - h_{21}^{b}}{h_{21}^{b}}$$

Indefinite parameters

All of the parameter sets corresponding to different configurations (common-base, common-emitter, common-collector) are derivable from one another.

A convenient trick (works best in y-parameter representation) is to disregard the common reference and treat the third terminal as an additional port:

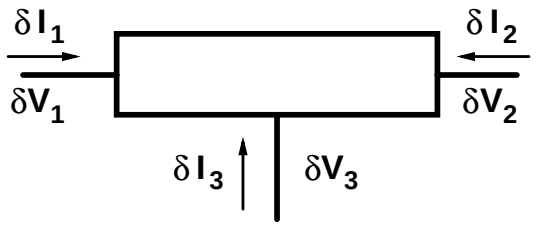


Figure 5: indefinite matrix:

$$\begin{bmatrix} \delta I_1 \\ \delta I_2 \\ \delta I_3 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{bmatrix} \begin{bmatrix} \delta V_1 \\ \delta V_2 \\ \delta V_3 \end{bmatrix}$$

From Kirhhoff's Law and the fact that the matrix should work for *arbitrary* set of $\{\delta V_i\}$ it follows that the sum of all columns (or rows) in the indefinite matrix is zero.

Thus, if we assume a short circuit at ports 1 and 2, the fact that the sum of all currents must be zero implies that the $\sum$ of y-parameters in the *third* column vanishes, and so on.

To prove that the $\sum$ must vanish in each *row*, we note that if all three $\{\delta V_i\}$ are equal no ac current can flow at any port.

It is exceedingly simple to transform from one common-terminal configuration to another. Thus, if we know the y-matrix in common base configuration, the corresponding common-emitter matrix is:

$$\begin{bmatrix} \mathbf{y}_{11}^{b} & \mathbf{y}_{12}^{b} & y_{13} \\ \mathbf{y}_{21}^{b} & \mathbf{y}_{22}^{b} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{bmatrix} \rightarrow \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{bmatrix} \rightarrow \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & \mathbf{y}_{11}^{e} & \mathbf{y}_{12}^{e} \\ y_{31} & \mathbf{y}_{21}^{e} & \mathbf{y}_{22}^{e} \end{bmatrix}$$

Power gain definitions:

- ***Power gain*** G is the ratio of power delivered to the load to power input into the network.

  It depends on both the input and the load circuits.

- ***Maximum available gain*** (MAG) is the maximum gain achievable from a particular transistor without external feedback.

  MAG equals the value of forward gain G which results when both the input and the output are simultaneously ***matched*** in an optimum way. For example, realization of MAG requires that the load resistance be matched to the output resistance $Re\,(z_{22})$.

- ***Unilateral gain*** U is the maximum available power gain of a device after it has been made unilateral by adding a lossless reciprocal feedback circuit. This means that the lossless network around the amplifier (inductances and capacitances) is adjusted so as to set the reverse power gain to zero.

Unilateral gain is *independent*† of common-lead configuration !

The unilateral gain U can be calculated from any of the following equivalent expressions:

$$U = \frac{|z_{21} - z_{12}|^2}{4[Re\,(z_{11})Re\,(z_{22}) - Re\,(z_{12})Re\,(z_{21})]};$$

$$= \frac{|y_{21} - y_{12}|^2}{4[Re\,(y_{11})Re\,(y_{22}) - Re\,(y_{12})Re\,(y_{21})]};$$

$$= \frac{|h_{21} + h_{12}|^2}{4[Re\,(h_{11})Re\,(h_{22}) + Im\,(h_{12})Im\,(h_{21})]},$$

where z_{ij}, y_{ij}, and h_{ij} are the impedance, the admittance, and the hybrid parameters of a transistor, respectively, ***for any configuration.***

---

† This remarkable result (Mason's theorem) is the main reason for the wide-spread use of U. See S. J. Mason, "Power gain in feedback amplifiers", *IRE Trans. Circuit Theory* **CT-1**, pp. 20-25 (1954).

Small-signal model of an abrupt junction HBT

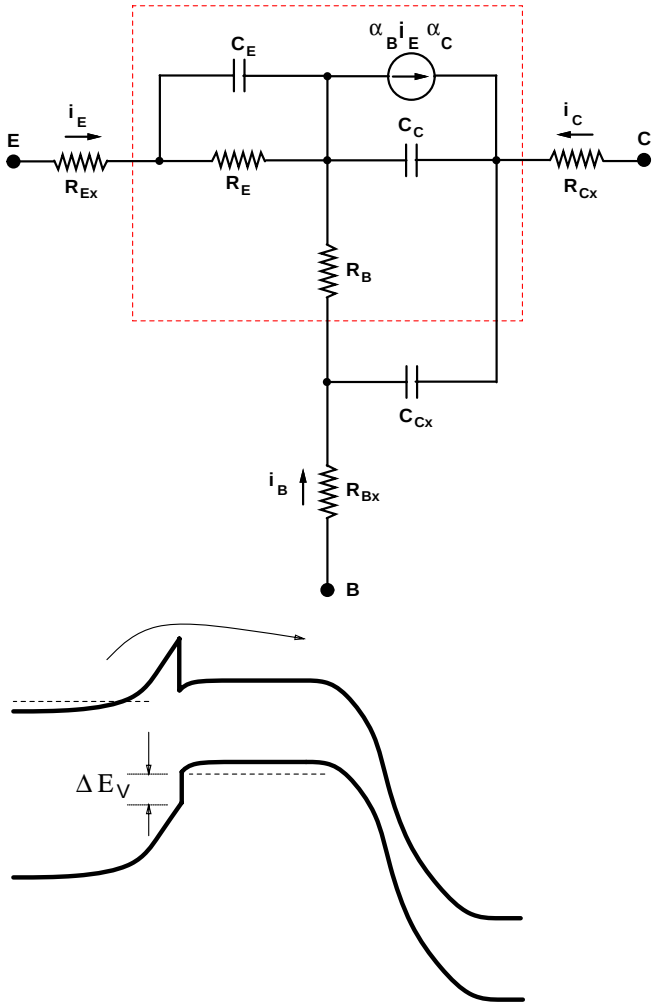


Figure 6: Small-signal analysis of this simple model, including frequency dependence of the power gain in both ballistic and diffusive regimes, has been carried out by Grinberg and Luryi (1993).†

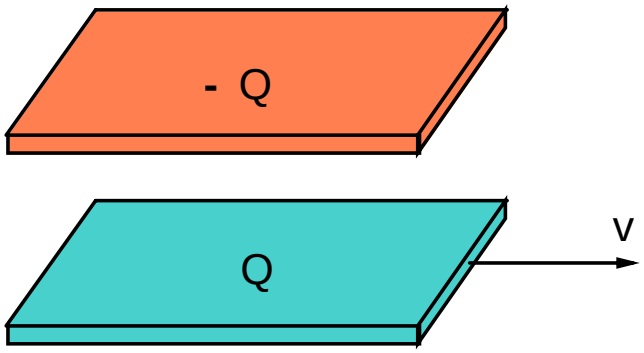
† A. A. Grinberg and S. Luryi, "Coherent transistor", *IEEE Trans. Electron Devices* **ED-40**, pp. 1512-1522 (1993).

A. A. Grinberg and S. Luryi, "Dynamic Early effect in heterojunction bipolar transistors", *IEEE Electron Device Lett.* **EDL-14**, pp. 292-294 (1993).

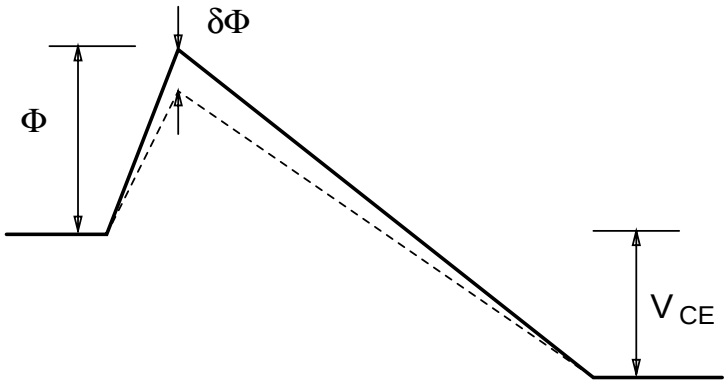
Quasi-static (Ebers-Moll-like) model of abrupt-junction HBT can be found in
A. A. Grinberg and S. Luryi, "On the thermionic-diffusion theory of minority transport in heterostructure bipolar transistors", *IEEE Trans. Electron Devices* **ED-40**, pp. 859-866 (1993).

TRANSISTOR PRINCIPLES : FETs & PETs

Field Effect: Screening

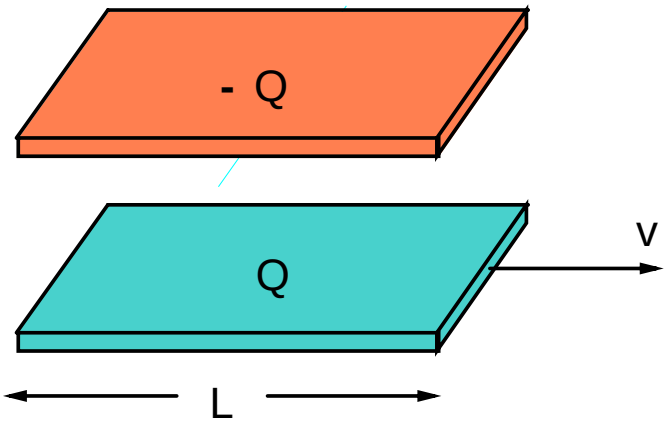


Potential Effect: Control of a cathode work function



FETs:

$$I = \frac{Q\,v}{L}$$



"Biblical" principle:

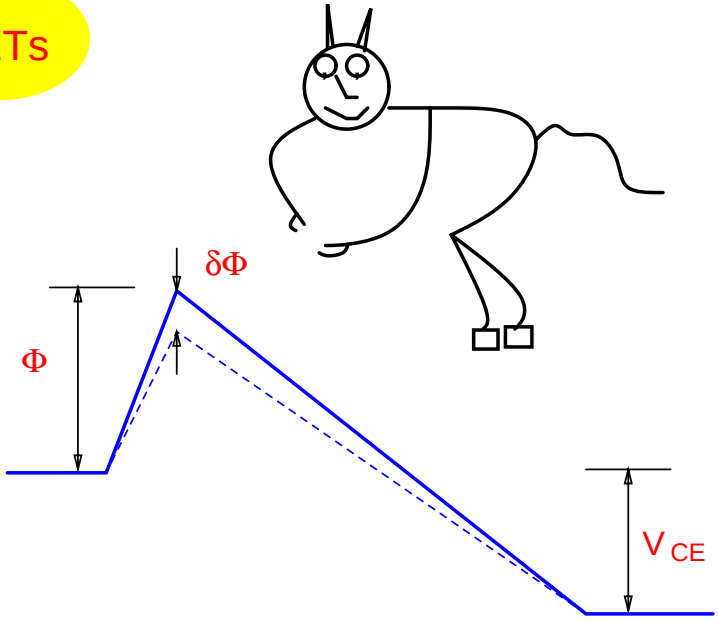
Q for Q

I for I

"Transit time" limitation :

$$\tau > \frac{Q_{in}}{I_{out}} = \frac{L}{v}$$

PETs



$$I \sim e^{-\Phi/kT}$$

$$\delta\Phi \sim \delta Q_{in}$$

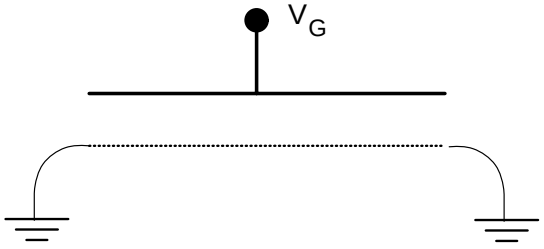
$$\tau \sim I^{-1}$$

Speed increases with current until exponential law fails at high currents

PET → FET (space-charge effect)

τ limited by transit time across

Two Dimensional Channel



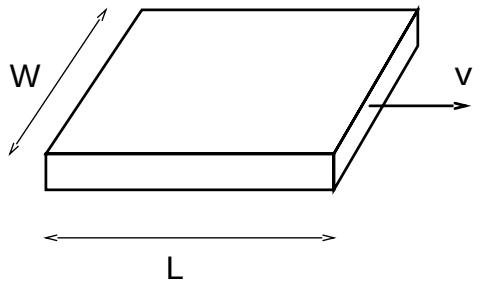
$$e\,n\ [\mathrm{cm}^{-2}] = \frac{C}{A}\,V_G$$

Conductance of 3D sample:

$$g \equiv \frac{\mathrm{d}I}{\mathrm{d}V} = [e\,N\,\mu]\,\frac{A}{L}\ , \qquad e\,N\,\mu\ \left[\frac{1}{\Omega\cdot\mathrm{cm}}\right]$$

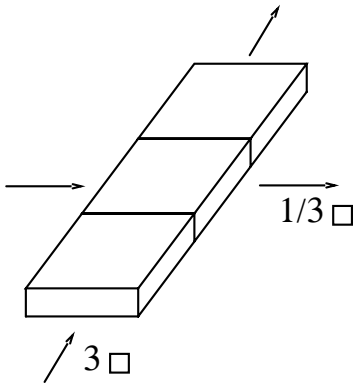
Conductance of 2D sample:

$$g \equiv \frac{\mathrm{d}I}{\mathrm{d}V} = [e\,n\,\mu]\,\frac{W}{L}\ , \qquad e\,n\,\mu\ \left[\frac{1}{\Omega}\right]$$



Note: resistance of a square independent of its size (contrast with a cube!)
☛ Ω/□

Knowing the "resistance per square" $\frac{1}{e\,n\,\mu}$ one can simply count squares:



Current density per unit width [A cm^{-1}]

$$J \equiv \frac{I}{W} = en\,v$$

$$= en\,\mu F = en\,\mu\,\frac{V}{L}$$

conductance per unit channel width ☛ $g = en\,\mu/L$ [mS/mm]

Transconductance (also per unit width)

$$g_m \equiv \left.\frac{\partial J}{\partial V_G}\right]_{V_D}$$

$$J = en\,v$$

$$g_m = \frac{\partial (en)}{\partial V_G}\,n = \frac{C}{A}\,v \quad [\text{mS/mm}]$$

Figure of Merit ("FOM"):

$$\frac{C}{g_m} = \frac{L}{v} \qquad (\text{delay time})$$

Conduction Laws by dimensional analysis§

In the CGS system the conductivity has the units of a velocity.† Taking this velocity to be an effective carrier velocity v, we can write a generic expression for the diode current in the form

$$I \propto \frac{\varepsilon}{4\pi} v\, V \frac{A}{L^2}\,, \qquad \text{(bulk)}$$
$$I \propto \frac{\varepsilon}{4\pi} v\, V \frac{W}{L}\,. \qquad \text{(film)}$$

where L is the length, $A = D \cdot W$ the cross-sectional area, and W the width of the diode. The relative permittivity $\varepsilon \equiv \varepsilon/\varepsilon_0$ of the material enters because it scales the space-charge potential in Poisson's equation.

The actual current-voltage dependence (up to a numerical coefficient) can be ''derived'' from the above equation – whenever the conduction process involves a dominant transport mechanism, which provides a unique scaling relationship between v and V.

Thus, for free electron motion, the velocity scales as $v^2 \propto (e/m)V$ and one obtains laws appropriate for ballistic transport, e.g. for the bulk case Child's law of vacuum electronics:

$$I = \zeta \frac{\varepsilon}{4\pi} \left[\frac{e}{m}\right]^{1/2} V^{3/2} \frac{A}{L^2} \qquad \text{where} \quad \zeta = \frac{4\sqrt{2}}{9} \quad [\text{Child}]$$

For the case when electron velocity is saturated, take $v = v_S$.

For the case of constant mobility μ, the velocity scales as $v \propto \mu V/L$, which leads to the following expressions:

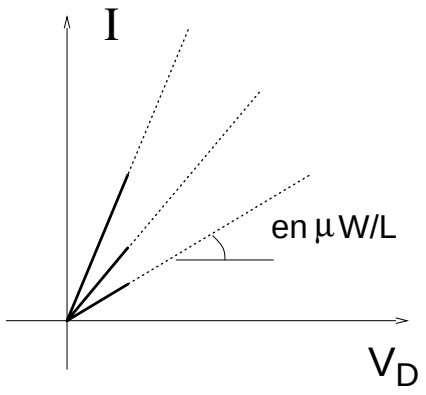
$$I = \zeta_3 \frac{\varepsilon}{4\pi} \frac{\mu V^2}{L^3} A\,, \qquad \text{(bulk)} \quad \text{where} \quad \zeta_3 = \frac{9}{8} \quad [\text{Mott–Gurney}]$$
$$I = \zeta_2 \frac{\varepsilon}{4\pi} \frac{\mu V^2}{L^2} W\,. \qquad \text{(film)}$$

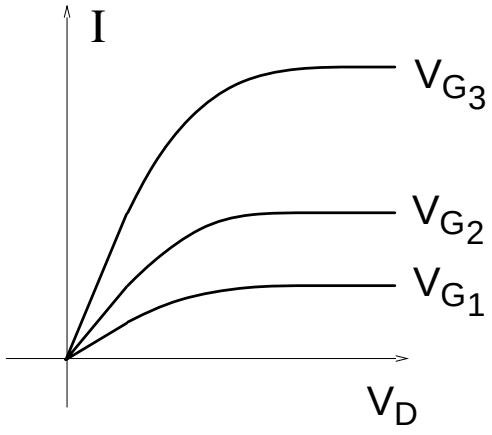
---

§ S. Luryi, "Device building blocks", Chap. 2 in *High-Speed Semiconductor Devices*, ed. by S. M. Sze, Wiley Interscience (1990) pp. 57-136.

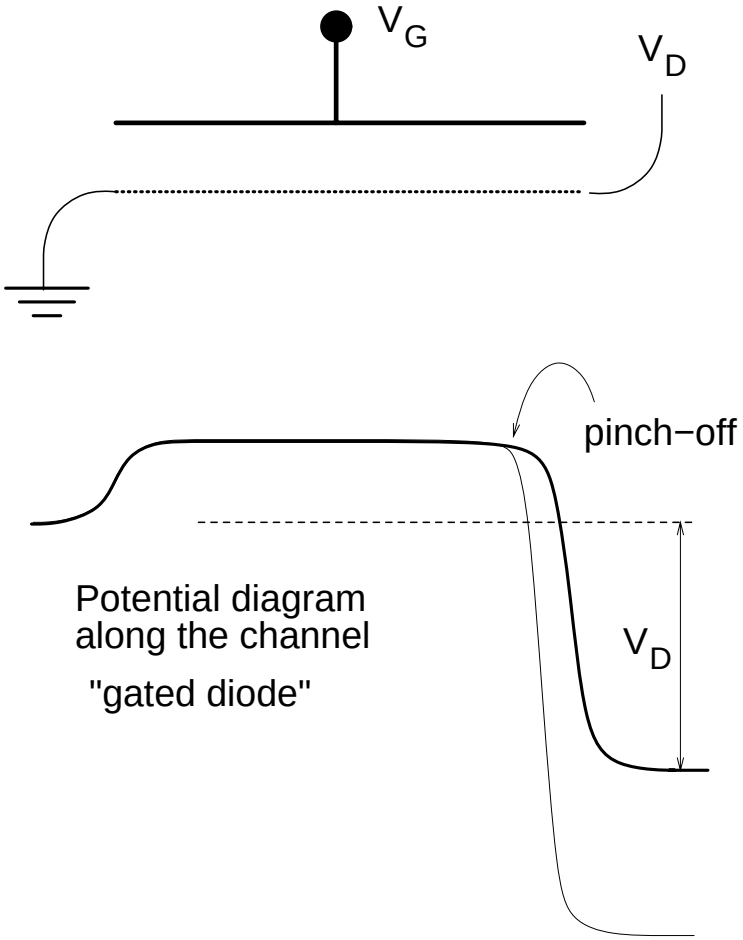
† Connection to the international units is obtained by replacing the dimensionless factor $\frac{\varepsilon}{4\pi}$ with ε in farads per meter.

Field-effect transistor characteristics

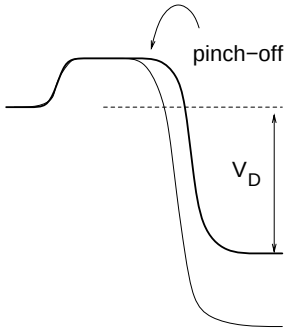




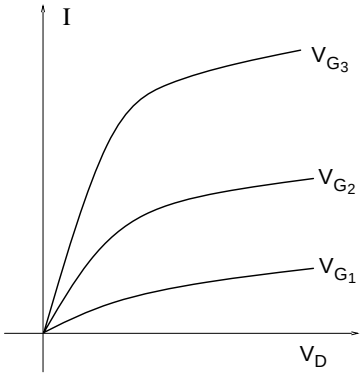
Transistor channel
versus thin film diode



Short channel effects



As the pinch off point moves left, the channel becomes shorter. The decreasing L leads to increasing current (finite output conductance) Recall Early effect in bipolars



For a given channel length, the closer the gate to the channel, the less important are the short-channel effects

MOS structure

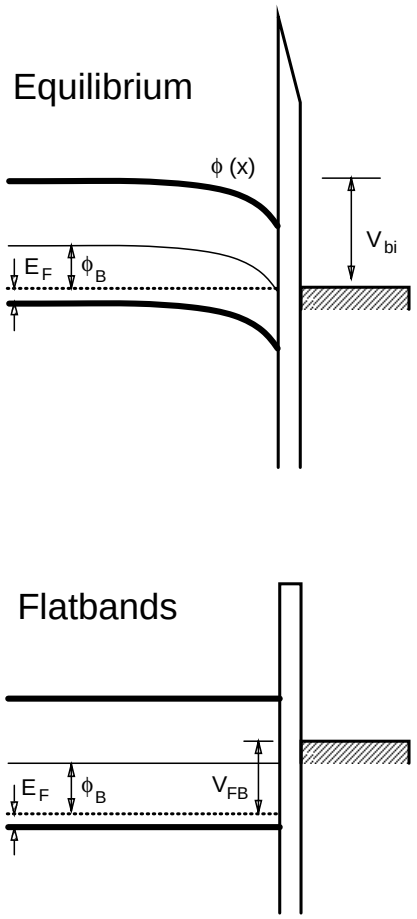


Figure 1: Built-in voltage $V_{\rm bi}$ and flatband voltage $V_{\rm FB}$.

These quantities are not identical.

$V_{\rm bi}$ is the electrostatic potential drop in equilibrium: it equals sum of the potential drops in the semiconductor and the oxide.

$V_{\rm FB}$ is the voltage that must be applied to the gate to flatten the bands. Since "voltage" means the Fermi level difference,† the $V_{\rm FB}$ equals the difference in the work functions of the semiconductor and the metal.

† Not the same thing as the electrostatic potential difference, which can exist even in equilibrium.

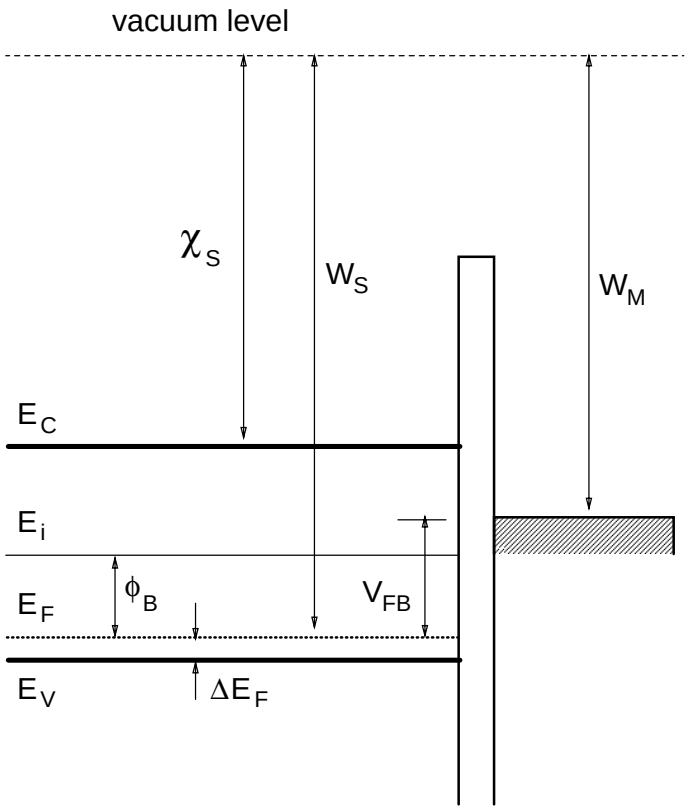


Figure 2: Relevant energies in the MOS system at flatbands.

- W_S: work function of the semiconductor
- W_M: work function of the metal
- $V_{FB} = W_S - W_M$: flatband gate voltage
- χ_S: electron affinity of the semiconductor
- $\phi_B \equiv E_i - E_F$: bulk doping characteristic, $\phi_B = kT \ln(N_A/n_i)$.

$$\begin{aligned} W_S &= \chi_S + E_G - \Delta E_F \\ &= \chi_S + (E_C - E_V) - (E_F - E_V) \\ &= \chi_S + E_C - E_F \end{aligned}$$

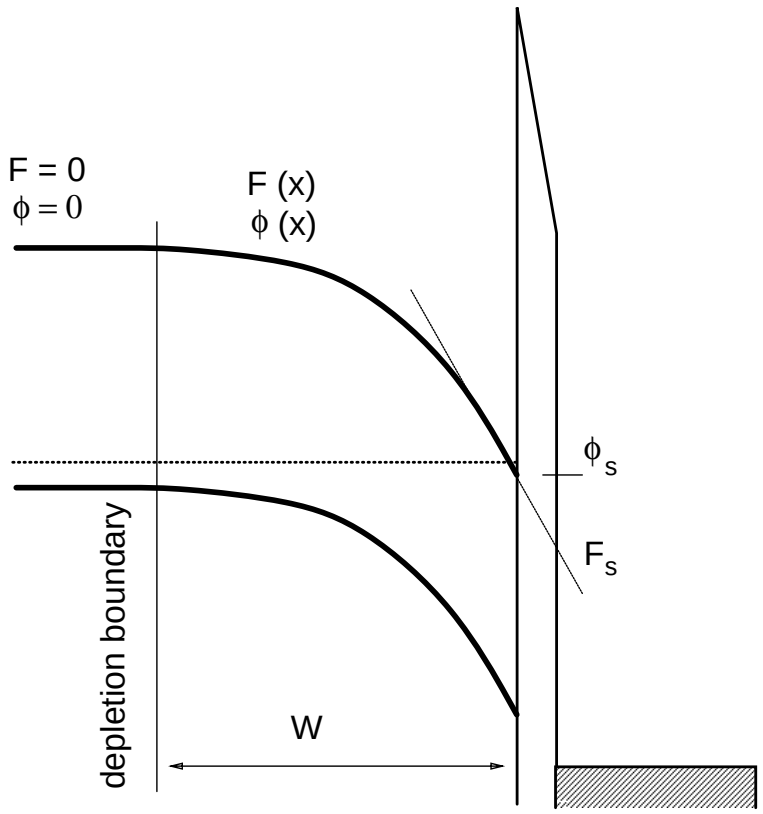


Figure 3: Evaluation of the band bending in MOS structure

Poisson's equation: $$\phi'' = -\frac{e(N_A - p)}{\varepsilon}$$

$p = N_A e^{-\beta\phi}$ ☛ $$\phi'' = -\frac{eN_A}{\varepsilon}\left[1 - e^{-\beta\phi}\right] \qquad (1)$$

Note a trick: $$\phi'' \equiv \frac{d\phi'}{dx} = \frac{d\phi'}{d\phi}\frac{d\phi}{dx} = \frac{dF}{d\phi}F = \frac{1}{2}\frac{dF^2}{d\phi}$$

☛ multiply Poisson's equation (1) by $d\phi$ and integrate

from	$x = -\infty$	$\phi = 0$	$F = 0$
to	$x =$ surface	$\phi = \phi_S$	$F = F_S$

using

$$\int_0^{\phi_S}\left[1 - e^{-\beta\phi}\right]d\phi = \frac{kT}{e}\left[\beta\phi_S + e^{-\beta\phi_S} - 1\right]$$

Exact:

$$\frac{1}{2} F_S^2 = \frac{kT\, N_A}{\varepsilon} \left[\beta\phi_S + e^{-\beta\phi_S} - 1 \right]$$

Approximate ($\beta\phi_S \gg 1$, i.e., $\phi_S \gg kT/e$):

$$F_S^2 = \frac{2kT\, N_A}{\varepsilon} \beta\phi_S = \frac{2e\, N_A}{\varepsilon} \phi_S$$

The same result is obtained in the depletion approximation:

$$\phi_S = \frac{eN_A W^2}{2\varepsilon} \qquad F_S = \frac{eN_A W}{\varepsilon}$$ where W is the depletion width

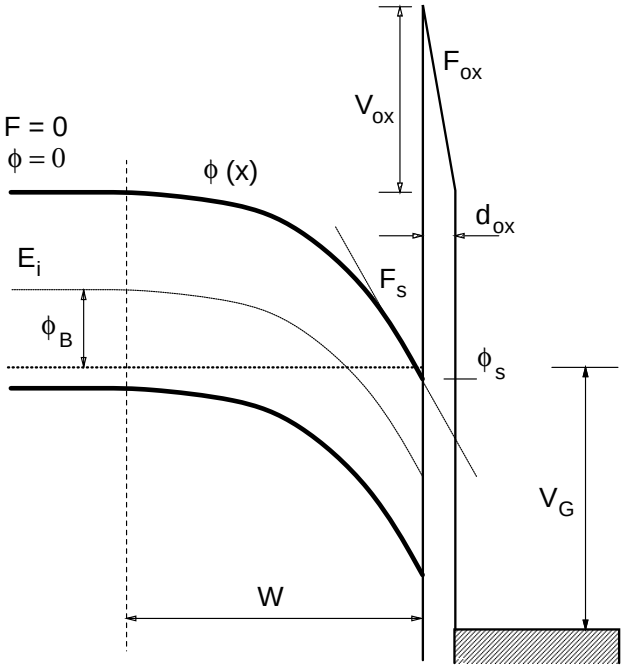


Figure 4: Threshold gate voltage evaluation:

Given bulk doping N_A ☛ $e\phi_B = kT \ln(N_A/n_i)$

at threshold $\phi_S = 2\phi_B$ ☛ $F_S = \sqrt{\frac{4e N_A \phi_B}{\varepsilon_S}}$

$\varepsilon_S F_S = \varepsilon_{OX} F_{OX}$ ☛ $F_{OX} = \frac{\varepsilon_S}{\varepsilon_{OX}} \sqrt{\frac{4e N_A \phi_B}{\varepsilon_S}}$

$V_{OX} \equiv d_{OX} F_{OX}$ ☛ $V_T = e\phi_S + V_{OX} + V_{FB}$

Above Treshold we need to include the charge of the inversion layer itself; this is a quantum mechanical problem, let us go back to

basic semiconductor physics

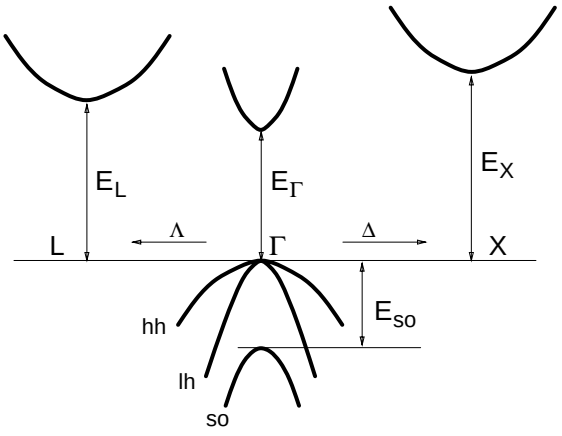


Figure. Important extremal points in the band structure of cubic semiconductors. The schematic picture (drawn not to scale) is appropriate for a direct-gap III-V semiconductor. For GaAs at room temperature the indicated energies are:
$E_\Gamma = 1.42\,eV$, $E_\mathrm{L} = 1.71\,eV$, $E_\mathrm{X} = 1.90\,eV$, $E_\mathrm{so} = 0.34\,eV$.
In silicon the lowest conduction band point is in the Δ direction, 85% of the way to X point. The indicated energies for silicon at 300 K are:
$E_\Gamma = 4.08\,eV$, $E_\mathrm{L} = 1.87\,eV$, $E_\Delta = 1.13\,eV$, $E_\mathrm{so} = 0.04\,eV$.
In Ge the lowest conduction band point is at L but the Γ point is not far away:
$E_\Gamma = 0.89\,eV$, $E_\mathrm{L} = 0.76\,eV$, $E_\Delta = 0.96\,eV$, $E_\mathrm{so} = 0.29\,eV$.

Bands: degenerate / nondegererate
isotropic / anisotropic
parabolic / nonparabolic

In the vicinity of a nondegenerate extremal point it is convenient to describe the dispersion relation $E_n\,(\mathbf{k})$ using *the effective mass tensor* $\mathbf{M}_n^{-1}$:

$$E_n\,(\mathbf{k}_0 + \mathbf{k}) - E_n\,(\mathbf{k}_0) = (\hbar^2/2)\,\mathbf{k}\cdot\mathbf{M}_n^{-1}\cdot\mathbf{k} \equiv \frac{\hbar^2}{2}\sum_{i,j=1}^{3}\left[\mathbf{M}_n^{-1}\right]_{ij} k_i\,k_j \quad ,$$

where the components of $\mathbf{M}_n^{-1}$ can be written down in terms of the free electron mass and parameters of the lattice potential (the periodic $V\,(x\,,y\,,z\,)$ characteristic of the point $\mathbf{k}_0$.

In the vicinity of a nondegenerate band extremum, the surfaces of equal energy are ellipsoids, as evident from Eq. (1). The symmetric tensor $\mathbf{M}_n^{-1}$ has, most generally, six independent components. The coordinate axes can always be chosen so as to diagonalize this tensor, i.e. along the ellipsoid's principal directions:

$$\mathbf{M}_n^{-1} = \begin{bmatrix} 1/m_1 & 0 & 0 \\ 0 & 1/m_2 & 0 \\ 0 & 0 & 1/m_3 \end{bmatrix} .$$

In general, the energy ellipsoid is determined by six independent parameters: three diagonal values of $\mathbf{M}_n^{-1}$ and three directions of the principal axes. However, the number of parameters can be often reduced by symmetry considerations. The ellipsoid symmetry is determined uniquely by the symmetry of the extremal point $\mathbf{k}_0$. For an extremum located on a crystal symmetry axis, one of the principal directions of the energy ellipsoid coincides with the symmetry axis. If the latter is an axis of 3-fold, 4-fold, or 6-fold symmetry, then the ellipsoid is an ellipsoid of revolution ($m_1 = m_2 \equiv m_t$, $m_3 \equiv m_l$). If more than one such axis intersects at $\mathbf{k}_0$, then the ellipsoid anisotropy disappears and energy surfaces become spherical ($m_1 = m_2 = m_3 \equiv m$). Such is the situation in the conduction band at the Γ point of cubic semiconductors:

$$E(k) = \frac{\hbar^2 k^2}{2m} .$$

In silicon the conduction band minima are on the 4-fold rotation axes and the low-energy isoenergetic surfaces are ellipsoids of revolution, their long axes being along <100> directions. There are six symmetry related minima. Similar local minima exist in the conduction band of germanium, but the true conduction band minima in Ge are located at L points.

There are only four symmetry-related ellipsoids of constant energy in the vicinity of the conduction band edge of Ge. It is convenient to picture these ellipsoids as eight half-ellipsoids joined together on opposite faces by translations through suitable reciprocal lattice vectors. In each ellipsoid, the band curvature is least in the direction along the rotation axis and highest in the transverse directions. This means that the longitudinal mass is heavier than the transverse mass. The anisotropy is particularly high in Ge, $m_l/m_t = 20$, but it is also considerable in Si, where $m_l/m_t \approx 5$.

In most cases, it is a reasonable approach to describe the electronic motion in a uniform electric field by the effective mass equation.

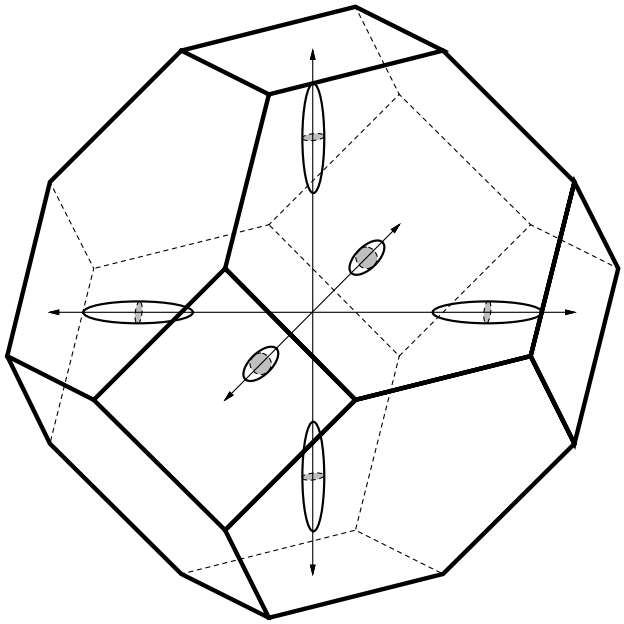


Figure . Surface of constant energy in the vicinity of the conduction band edge in silicon represents six ellipsoids of revolution, extended along <100> directions. The band curvature is least in the longitudinal direction (heavy mass) and highest in the transverse direction (light mass) The effective mass ratio $m_l/m_t \approx 5$.

If the band edge in Si were at the zone boundary rather than at a general point in the Δ direction ($\approx 85\%$ toward X), then there would be only three ellipsoids. Location away from the zone boundary of the conduction band edge in Si and the local Δ minimum in Ge is related to the inversion symmetry of the diamond structure.

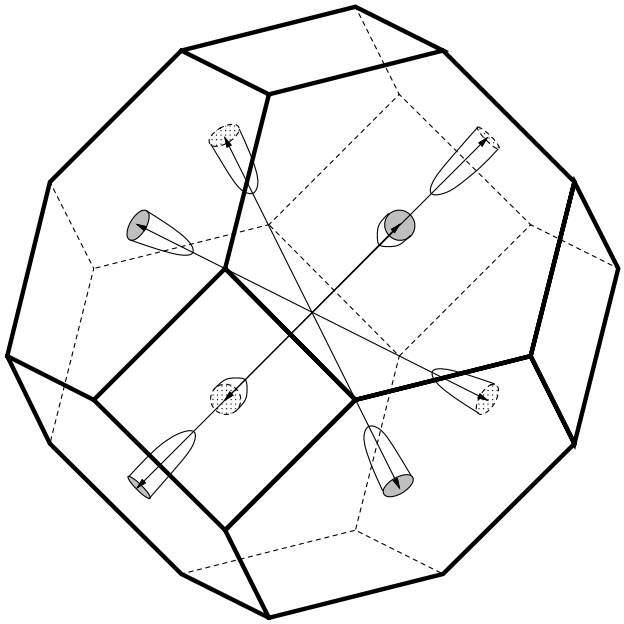


Figure: Surface of constant energy in the vicinity of the conduction band edge in germanium represents four ellipsoids of revolution, extended along <111> directions. The effective mass ratio is very large, $m_l/m_t = 20$, so that each ellipsoid really looks like a sausage.

In order to exhibit a full ellipsoid we would have to chose a primitive cell for which some L point would be internal. In the Brillouin zone picture, each ellipsoid is cut in two by the boundary and the equivalent half is shifted to the opposite face by a reciprocal lattice vector.

Inversion layer in a silicon MOS structure

The schematic cross-section of a silicon MOS structure is illustrated on the next page along with the band-bending near the Si/SiO_2 interface under a sufficiently large positive gate bias. Let us look in more detail at the band structure of the 2DEG in an inversion layer on the {100} surface.

In a (roughly triangular) quantum well formed near the Si/SiO_2 interface under a positive gate bias, ellipsoids oriented differently with respect to the surface give rise to a quite different subband structure. Let us specify the actual {100} Si surface as a (100) crystal plane. Electrons in the two ellipsoids elongated in [100] direction possess the heavy mass m_l in z direction and an isotropic light mass m_t in any direction lying in the (100) plane. These electrons give rise to the subbands whose bottom-edge energies are denoted by E_n. The other four ellipsoids, whose longitudinal axes lie in the (100) plane, correspond to the light mass m_t in the [100] direction, and their subbands are denoted by E_n'.

Because the quantum-well energy levels scale with $1/m$, one has $E_0 < E_0'$, and so the inversion-layer electrons in their ground subband have an isotropic light mass. The order of the higher-lying subbands can be established only on the basis of self-consistent numerical calculations. The subband energies depend not only on the field and the background doping but also on the temperature, which affects the relative population of higher-lying subbands and, hence, the self-consistent field.

---

‡ Notation: equivalent crystallographic planes, e.g., (100), $(01\bar{0})$, etc., are collectively denoted by {100}. Similarly, equivalent (symmetry-related) directions in the reciprocal lattice, e.g., [100], [010], etc., are collectively referred to as the <100> direction.

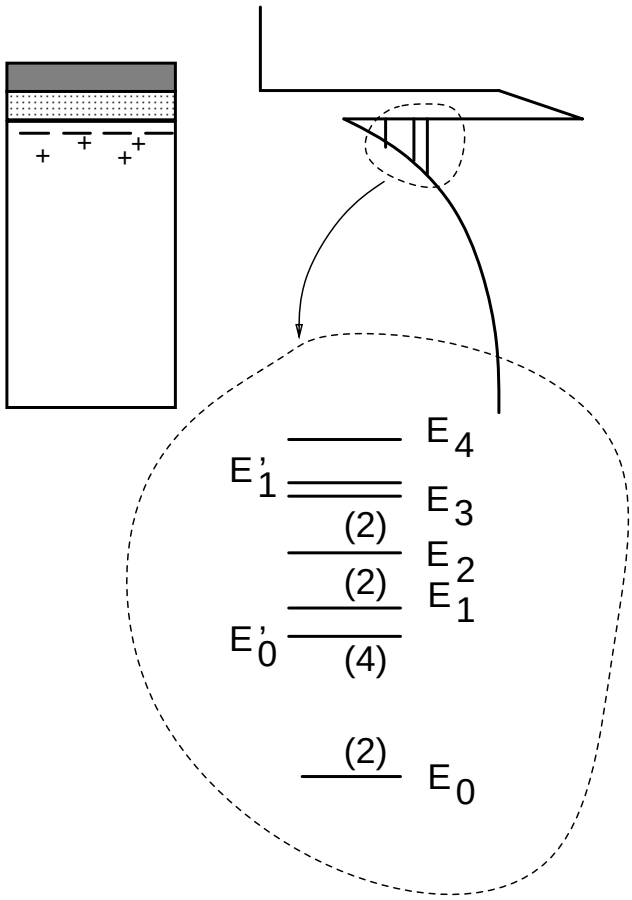


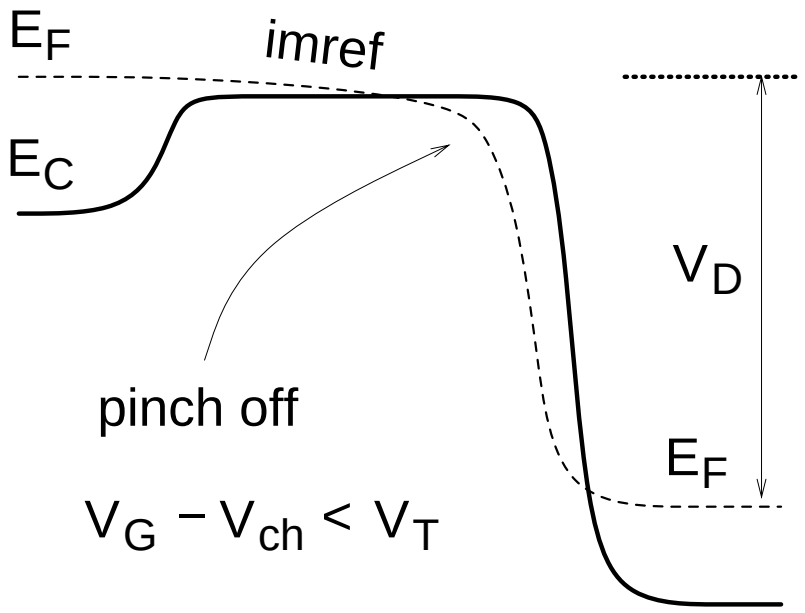
Figure: The order of the seven lowest subbands, calculated for an inversion layer at a Si-{100} surface with $n_S = 10^{12}\,\mathrm{cm}^{-2}$ in a lightly-doped ($N_A = 10^{15}\,\mathrm{cm}^{-3}$) p-type material at room temperature. The levels E_0' and E_1 are close in energy and, in fact, change their order at lower T and/or n_S.

Pinch off in a MOSFET

Recall the gradual channel approximation: Treat the potential diagram ***locally*** at any channel cross-section (x) by ignoring the voltage difference between the source and the drain, but taking the channel to be not not at ground voltage but at $V = V_{\rm ch}(x)$, i.e., by replacing

☛ $V_{\rm G} \rightarrow V_{\rm G} - V_{\rm ch}(x)$

The $V_{\rm ch}(x)$ is the imref (the quasi-Fermi level) which varies monotonically from $E_{\rm F} = 0$ in the source to $E_{\rm F} = V_{\rm D}$ in the drain.



Imref

The term Fermi level in semiconductor physics is synonymous with ‘‘chemical potential’’; in equilibrium it is defined by

$$n = \int_{E_0}^{\infty} dE\ g(E)\ f(E - E_F) \ \simeq\ N_C\, e^{(E_F - E_C)/kT}\ ,$$

where $f(E) = [\exp(E/kT) + 1]^{-1}$ is the Fermi function and the approximation is valid for nondegenerate semiconductors; here we may have

$E_C = E_C(x)$ and $n = n(x)$, but the chemical potential $E_F = \text{const}$

In the presence of a current flow the concept of Fermi level is not defined.

However, let us ***define***

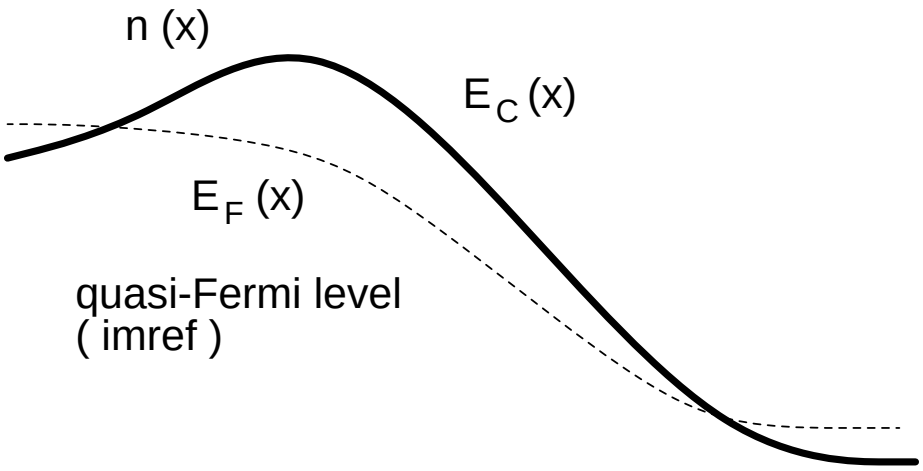
$$n \ = \ N_C\, e^{(E_F - E_C)/kT}$$

where

$n = n(x)$
$E_C = E_C(x)$
☛ $E_F = E_F(x)$

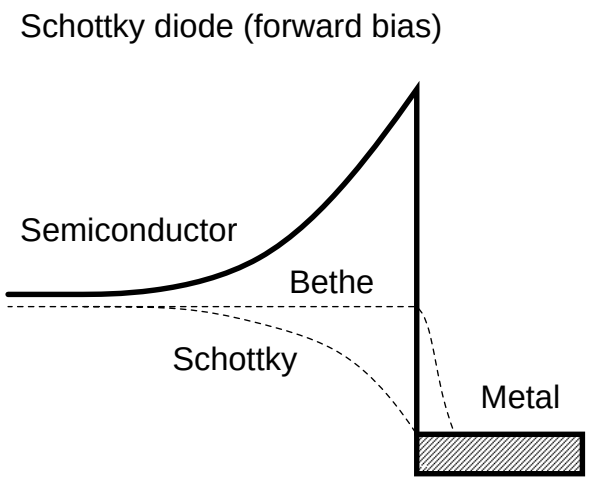
The drift diffusion equation then reduces to

$$J \ = \ e\,(n\,\mu\mathbf{F} \ + \ D\,\nabla n\,) \ \equiv \ e\,n\,\mu\,\nabla E_F$$



Example: use of imref to distinguish "pictorially" different transport mechanisms.

Thermionic (Bethe) versus diffusion (Schottky) mechanism of conduction in Schottky diodes.



☞ Imref varies little where the net current is much smaller than either the diffusion or the drift components. These regions are "approximately in equilibrium".