In equilibrium, electron concentration \( n_{p0} \) and hole concentration \( p_{p0} \) are
\[
\begin{align*}
    n_{p0} &= n_t^2/n_A \\
    p_{p0} &= n_A 
\end{align*}
\]

Energy Band Diagram of a P-type Semiconductor

In the diagram, the following are noted:
- \( E_f \): Fermi level
- \( E_c \): Conduction band edge
- \( E_v \): Valence band edge
- \( E_F^p \): Intrinsic Fermi level
- \( E_F^p \): Mobility edge
- \( E_F^p \): Fermi level

MOS Structure

\( V_G \) (Gate voltage)
\( V_D \) (Drain voltage)
\( V_S \) (Source voltage)
\( V_B \) (Substrate voltage)

Maxwell Law
\[
V_F = kT \ln \left( \frac{N_t}{n_t} \right) < 0 \quad (3.4) \\
V_F = kT \ln \left( \frac{N_A}{n_A} \right) > 0 \quad (3.5) 
\]

\( \Phi^* \): electron affinity - potential difference between the conduction band and the vacuum (true space)

\( \Phi^* \): work function - the energy required for an electron to move from the Fermi level to true space

\[
\Phi^* = \Phi^* - \Phi + E_x - E_F 
\]
Energy Band Diagram of individual components of MOS

Energy Band Diagram of MOS system

MOS system under External Bias

Band Diagram of MOS system with V_d > 0

Energy Band Diagram of MOS system for V_F > 0 (Large)

Cross-section of MOSFET
\[ V_{DS} = V_{GS} - V_{TO} \]

\[ V_{AD} = V_{DS} - V_{TO} \]

\[ Q_e(y) = \text{total mobile electron charge in the surface inversion layer} \]

\[ Q_e(y) = -C_{ox} \left[ V_{GS} - V_c(y) - V_{TO} \right] \]

\[ dR = \text{incremental resistance of the differential channel segment} \]

\[ dR = \frac{\mu_n \cdot R(x)}{W \cdot \mu_n \cdot R(x)} \frac{dy}{dV_c} \]

\[ \frac{\mu_n \cdot R(x)}{W \cdot \mu_n \cdot R(x)} \frac{dy}{dV_c} = \frac{dy}{dV_c} \]

\[ \int_0^y \mu_n \cdot R(x) dx = \int_0^y dV_c \]

\[ \Rightarrow \int_0^y = -W \cdot \mu_n \int_0^{V_{DS}} -C_{ox}(V_{GS} - V_c(y) - V_{TO}) dV_c \]

When \( V_{DS} = V_{GS} - V_{TO} \)

\[ \Rightarrow \int_0^y = \int_0^{V_{GS} - V_{TO}} \]

Slide 2.9: Simplicity comes at a cost. Comparing the I-V curves produced by the model to those of the actual device CMOS SPICE model, a large discrepancy can be observed for saturation values of transistors. When using the model for the derivation of propagation delays (behavior of a CMOS gate), accuracy in this section of the overall operation region is not that crucial. What is most important is that the values of current at the highest values of \( V_{DS} \) and \( V_{GS} \) are predicted correctly – as these predominantly determine the charge and discharge times of the output capacitor. Hence, the propagation delay error is only a couple of percent, which is only a small penalty for a major reduction in model complexity.
Thresholds and Sub-Threshold Current

\[ I_D = - \frac{W}{L} \frac{V_{DD}}{2} \left( \frac{V_{GS} - V_T}{V_{DD}} \right)^2 \]

For long-channel NMOS transistors,

\[ V_{GS} = 0 \quad \text{for} \quad V_{GS} < V_T \quad \text{(threshold voltage)} \]

\[ V_0 = V_{To} + \frac{1}{2} \left( \sqrt{2q \rho_n E_s} + V_B \right) \left( 1 + \frac{1}{2} \right) \]

\[ L = L + \left( \frac{L}{2} \right) \]

\[ \rho_n = q \mu_n (kT) \]

\[ \mu_n = \frac{e^2}{2\pi \hbar^2} \]

\[ E_s = \text{oxide thickness} \]

\[ C_{ox} = \frac{q^2}{2\pi \epsilon_0 \hbar^2} \]

\[ C_{ox} \approx \frac{q^2}{2\pi \epsilon_0 \hbar^2} \]

\[ g = n \times 10^3 \text{ cm}^{-2} \]

\[ V_{To} = -\frac{\sqrt{2q \rho_n E_s}}{C_{ox}} - \frac{V_B}{C_{ox}} \]

\[ A_{sub} = \frac{q \mu_n}{2} \]

\[ A_{sub} = \frac{kT}{2} \ln \left( \frac{V_{DD}}{V_T} \right) \]

\[ kT \approx 2.3 \times 10^{-19} \text{ J} \]

\[ 1.6 \times 10^{-19} \text{ J} \]

\[ \phi_{tr} = \frac{q \mu_n}{2} \ln \left( \frac{V_{DD}}{V_T} \right) \]

\[ \phi_{tr} = 0.026 \left( \frac{V}{11.43} \right) = 0.0143 \left( \frac{V}{11.43} \right) \]

\[ A_{sub} = \frac{q \mu_n}{2} \ln \left( \frac{V_{DD}}{V_T} \right) \]

\[ A_{sub} = 0.026 \left( \frac{V}{11.43} \right) \]

\[ A_{sub} = 0.026 \left( \frac{V}{11.43} \right) \]

\[ A_{sub} = 0.026 \left( \frac{V}{11.43} \right) \]

\[ V_{To} = -\frac{\sqrt{2q \rho_n E_s}}{C_{ox}} - \frac{V_B}{C_{ox}} \]

\[ V_{To} = 1.1 \times 10^3 \text{ cm}^{-2} \]

\[ V_{To} = 2.2 \times 10^3 \text{ cm}^{-2} \]

\[ 2.2 \times 10^3 \text{ cm}^{-2} \]

\[ V_{To} = -1.1 \left( \frac{V}{2.05} \right) - 1.1 \times 10^{-1} \left( \frac{V}{2.05} \right) \]

\[ 2.2 \times 10^3 \text{ cm}^{-2} \]

\[ 2.2 \times 10^3 \text{ cm}^{-2} \]
\[ V_c = V_B + \sqrt{\frac{1}{2} \mu \frac{W}{L} (V_{GS} - V_T) - \frac{V_{GS}^2}{2}} \]  

\[ V_T = \frac{1}{\mu} \frac{W}{L} \frac{V}{V_{GS}} \]

\[ Y = \frac{1}{2} \frac{\mu W}{L} \frac{V}{V_{GS}} \]

\[ \frac{V_T}{V_B} = \frac{1}{\mu} \frac{W}{L} \frac{V}{V_{GS}} \]

\[ V_{GS} = V_B + \sqrt{\frac{1}{2} \mu \frac{W}{L} (V_{GS} - V_T) - \frac{V_{GS}^2}{2}} \]

Forward and Reverse Body Bias

Threshold values can be adjusted through the fourth terminal, the transistor body.

For long channel NMOSFETs via gradual channel approx.

\[ I_{DS} = \mu_n C_{ox} \frac{W}{L} \left[ \frac{1}{2} (V_{GS} - V_T) V_{DS}^2 - V_{GS} V_{DS} \right] \]  

\[ (3.24) \]

Channel Length Modulation parameter \( \lambda \)

\[ \lambda = l - a \]

\[ \lambda = \text{empirical parameter} \]

With channel length modulation

\[ I_{DS} = \begin{cases} 0 & \text{if } V_{GS} < V_T \\ \frac{\mu_n C_{ox} W}{L} (V_{GS} - V_T) V_{DS}^2 - V_{DS} V_{GS} & \text{if } V_{GS} > V_T \end{cases} \]

\[ (3.29) \]

\[ I_{DS, \text{lin}} = \begin{cases} \frac{\mu_n C_{ox} W}{L} (V_{GS} - V_T) V_{DS}^2 - V_{DS} V_{GS} & \text{if } \lambda = 0 \\ \frac{\mu_n C_{ox} W}{L} (V_{GS} - V_T) (1 + \lambda V_{DS}) & \text{if } \lambda \neq 0 \end{cases} \]

\[ (3.29) \]
Similarly for PMOS: $V_{ds} = V_{ds, sat} + \frac{1}{2} \frac{M_o}{L} \left( \frac{V_{gs}-V_T}{2} \right)^2$ (3.58)

which is same as

$\frac{M_o}{L} \left( \frac{V_{gs}-V_T}{2} \right) \left( \frac{1 + \lambda}{\lambda} \right)$

$I_{d, sat} = \frac{M_o}{L} \left( \frac{V_{gs}-V_T}{2} \right) \left( \frac{1 + \lambda}{\lambda} \right)$

$I_{d} \text{ linear} = \frac{M_o}{L} \left[ \frac{1}{2} \left( \frac{V_{gs}-V_T}{2} \right) \right]^2$ (3.59)

For NMOS:

$I_{d} \text{ linear} = \frac{M_o}{L} \left( \frac{1}{2} \left( \frac{V_{gs}-V_T}{2} \right) \right)^2$

$V_{ds} > V_{gs} - V_T$

$V_{ds} < \left( \frac{V_{gs}-V_T}{2} \right)$

where $E_c = \text{channel electric field}$

Threshold voltage of small geometry devices:

$V_t = V_{t0} + K_v \left( \sqrt{V_{ds}} + V_{th} \right) - \sqrt{V_{ds}}$ (3.67)

$K_v = \frac{V_{th}}{2}$

$\Delta V = \Delta V_{th} + \Delta V_{T,MOS}$

$\Delta V_{th} = \frac{V_{th}}{2}$

$\Delta V_{T,MOS} = \frac{V_{th}}{2}$

$I_{d, subthreshold} = \frac{2 e}{L} \frac{W}{L} \frac{1}{e^2} \left( \frac{V_{ds}-V_{th}}{V_{ds}} \right)$ (3.115)

MOS Transistor Capacitances

Diffusion Capacitances in MOST ($C_{sb}, C_{ab}$)

Fig 3.49
Depletion capacitance of a reverse biased "abrupt" p-n junction:

\[ C_{j2}(v) = \frac{A}{\sqrt{C_{j0}} \sqrt{N_A + N_D}} \left( \frac{1}{\phi_0 - v} \right) \] (3.127)

More generally, with junction grading:

\[ C_{j2}(v) = \frac{A}{\sqrt{C_{j0}} \sqrt{N_A + N_D}} \left( \frac{1}{\phi_0 - v} \right)^m \] (3.128)

where \( m \) = grading coefficient

- \( m = \frac{1}{2} \) abrupt
- \( m = \frac{1}{3} \) linearly graded profile

\[ C_{j0} = C_{j2}(v) \bigg|_{v = 0 \text{ in (3.128)}} \]

where

\[ \phi_0 = \frac{1}{2} \left( \sqrt{\phi_1^2 + \phi_2^2} - \sqrt{\phi_1^2 - \phi_2^2} \right) \]

and \( \phi_0 < \phi_2 \) where

\[ C_{j0} = \frac{2}{V_2 - V_1} \left( \frac{1}{\sqrt{\phi_0^2 - \phi_1^2}} - \frac{1}{\sqrt{\phi_0^2 - \phi_2^2}} \right) \]

For \( m = \frac{1}{2} \) (abrupt junction):

\[ C_{j2} = \frac{A}{V_2 - V_1} \left( \frac{1}{\sqrt{\phi_0^2 - \phi_1^2}} - \frac{1}{\sqrt{\phi_0^2 - \phi_2^2}} \right) \] (3.132)
\[ j_0 = \sqrt{ \frac{q \cdot \frac{N}{2} \cdot \frac{4}{3} \cdot 10^{-15} \cdot 1 \cdot 6 \times 10^{-15} \cdot 9 \cdot 1.7 \times 10^{-15}}{2 \times 1.3 \times 10^{-7}}} \]
\[ j_0 = 2.21 \times 10^7 \text{ A/cm}^2 \]

\[ V_{th} = 2.7 \left( \frac{V_{ds} - V_{gs}}{V_{ds} - V_{th}} \right) \]
\[ \frac{1}{2.7} \left( \frac{1.97 - 0.7}{1.97 - 0.7} \right) - \frac{1}{2.7} \left( \frac{1.97 - 0.7}{1.97 - 0.7} \right) \]
\[ 0.82 \]

\[ \eta = 2.33 \times 10^{-7} \text{ cm}^2 / \text{A} \]

**Slide 2.19**

**MOS Transistor Leakage Components**

- Variability
- Leakage
- Sub-threshold leakage
- Gate leakage
- Body leakage

**Slide 2.20**

The topic of variability rounds out the discussion of the nanometer transistor and its properties. It has always been the case that transistor parameters such as the geometric dimensions or the threshold voltage are not deterministic. When sampled between wafers, within a wafer, or even over a die, each of these parameters exhibits a statistical nature. In the past, the projection of the parameter distributions onto the performance space yielded quite a narrow distribution. This is costly and undesirable. When the supply voltage is 3V and the threshold is at 0.5V, a 25% variation in the threshold has only a small impact on the performance and leakage of the digital module. However, when the supply voltage is at 1V and the threshold at 0.3V, the same variation has a much larger impact.

So, in past generations processors it was sufficient to evaluate a design over its worst-case corners (FF, SS, FF, SS) in addition to the nominal operation point to determine the yield distributions. Today, this is not sufficient, as the performance distributions have become much wider, and a pure worst-case analysis leads to wasteful design and does not give a good yield perspective either.

**Slide 2.21**

While variations influence the high-performance design regime, their impact is far more pronounced in the low-power design arena. First of all, the prediction of leakage currents becomes hard. The sub-threshold current is an exponential function of the threshold voltage, and each variation in the latter is amplified in a major way in leakage fluctuations. This is illustrated very well in the performance leakage distribution plot for the 130nm technology. When sampled over a large number of dies (and wafers), gate performance varies over 30%, while the leakage current fluctuates by a factor of 3. Observe that the leakiest designs are also the ones with the highest performance (this should be no surprise).

**Slide 2.22**

The number of transistors in the introduction, we have alluded to the increasing effects of "leakage" currents in the nanometer MOS transistor. An ideal MOS transistor (at least from a digital perspective) should not have any currents flowing into the bulk (or well) should not conduct any current between drain and source when off, and should have an infinite gate resistance. As indicated in the accompanying slide, a number of effects are causing the contemporary devices to deviate from this ideal model. Leakage currents flowing through the reverse-biased source-bulk and drain-bulk junctions, have always been present. Yet, the levels are so small that their effects could generally be ignored, except in circuits that relies on charge storage such as DRAMs and dynamic logic.

The scaling of the minimum feature sizes has introduced some other leakage effects that are far more influential and exceed junction leakage currents by 1-2 orders of magnitude. Most important are the sub-threshold drain-source and the gate leakage effects, which we will discuss in more detail.
Sub-threshold Current

- Sub-threshold behavior can be modeled physically as:

\[ I_{\text{sub}} = 2e\mu d C_V \left( \frac{W}{L} \right)^{0.5} \left( \frac{1}{1 - \alpha} \right)^{0.5} \left( \frac{1}{1 - \beta} \right)^{0.5} \]

where \( \alpha \) is the slope factor (\( \approx 1 \)) and \( \beta = 2e\mu d C_V \left( \frac{W}{L} \right)^{0.5} \)

- Very often expressed in base 10 as:

\[ I_{\text{sub}} = I_0 \left( \frac{1}{1 - \alpha} \right) \left( \frac{1}{1 - \beta} \right) \]

where \( I_0 = 10^{-3} \), the sub-threshold swing, ranging between 0 mV and 100 mV

such as the minimum mean square (MMS) are used. Be aware that these are not yield unique solutions and that in practice the modeler finds the ones with the best fit.

Dotted to its simplicity, the alpha model is at the heart of the optimization framework discussed in later chapters.

Slide 2.9

Even simpler is the alpha model, introduced by Sokolov and Newton in 1990, which does not even attempt to approximate the actual 1V curves. The values of \( \alpha \) and \( \beta \) are purely empirical, chosen such that the propagation delay of a digital gate, approximated by the alpha model, best matches the propagation delay curves obtained from simulation. Typically, curve-fitting techniques such as the minimum mean square (MMS) are used. Be aware that these do not yield unique solutions and that in practice the modeler finds the ones with the best fit.

Dotted to its simplicity, the alpha model is at the heart of the optimization framework discussed in later chapters.

Slide 2.10

The MOS transistor current is proportional to the power impressed on parameter \( V = \mu d C_V \left( \frac{W}{L} \right)^{0.5} \). To increase \( I \) through scaling, one must either find a way to increase the mobility of the carriers or increase the gate capacitance per unit area. The former requires a fundamental change in the device structure to be discussed later. With the traditional way of increasing the gate capacitance (i.e., scaling \( W \)) running out of steam, the only remaining option is to look for gate dielectrics with a higher permittivity \( \varepsilon_r \) — the so-called high-k.
Slide 2.46: This increased two-dimensional control can be exploited in a number of ways. In the dual-gate device, the fact that the gate is controlling the channel from both sides as well as the top leads to increased process tolerances. Another option is to use an end gate to eliminate the top part of the gate, leading to the back-gate transistor. In this structure, one of the gates acts as the standard control gate, whereas the other is used to manipulate the threshold voltage. In a sense, this device offers similar functionality as the best case for FD-SOI transistor discussed earlier. Controlling the work functions of the two gates through the selection of appropriate types and quantities of dopants helps to maximize the range and sensitivity of the control knobs.
Slide 2.47
The fact that the FinFET and its cousins are dramatically different devices compared to your standard bulk MOSFET is best illustrated with these pictures from Berkeley and Intel. The process steps that set and control the physical dimensions are entirely different. Although this creates new opportunities, it also brings challenges, as the process steps involved are vastly different. The ultimate success of the FinFET depends greatly upon how these changes can be translated into a scalable, low-cost and high-yield process - some formidable questions, indeed! Also unclear at this time is how the adoption of such a different structure impacts variability, as critical dimensions and device parameters are dependent upon entirely different process steps.