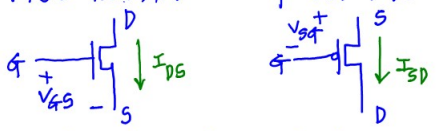
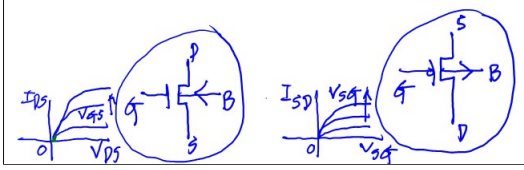
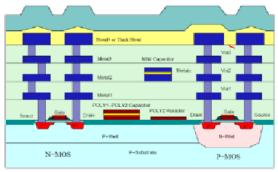
ECE222 Lecture 2 Oct 1, 2019

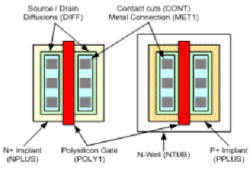
MOS Transistors chap 3 (pp 51-78; 79-104)

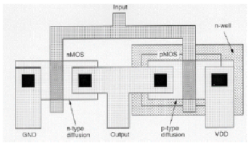


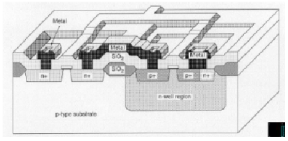
n-channel (NMOST) p-channel (pMOST)

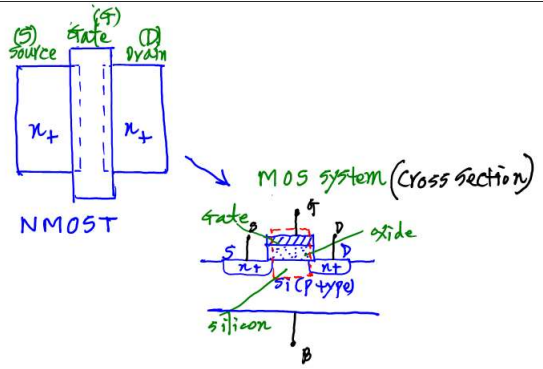


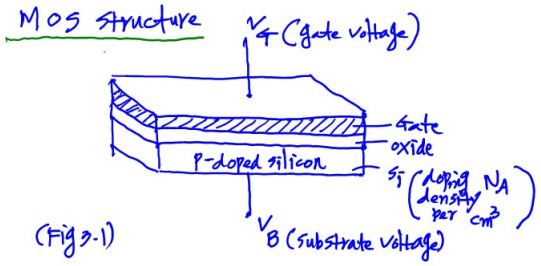












(Fig 3.1)

Mass action law $n \cdot p = n_i^2$

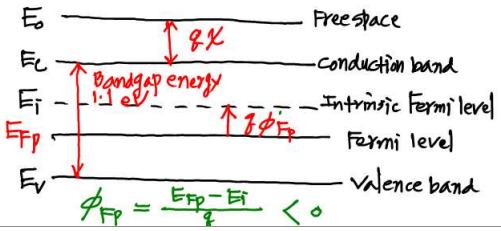
$n_i = 1.45 \times 10^{10}/cm^3$ at $T = 300^\circ K$

In equilibrium, electron concentration (n_{p0}) and hole concentration (p_{p0}) are

$$n_{p0} \cong n_i^2 / N_A$$

$$p_{p0} \cong N_A$$

Energy Band Diagram of a P-type Semiconductor



$\phi_{Fp} = \frac{E_{Fp} - E_i}{q} < 0$

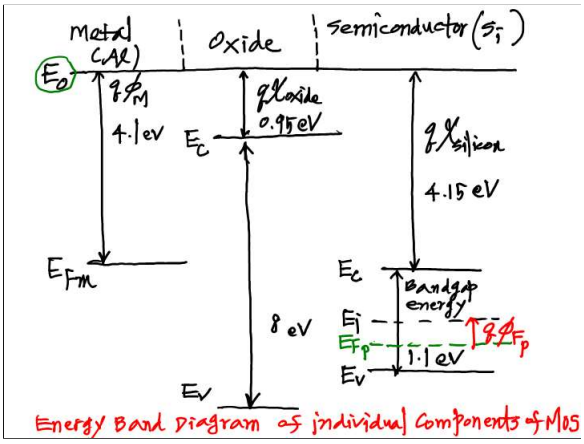
$$\phi_{Fp} = \frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right) < 0 \quad (3.4)$$

$$\phi_{Fn} = \frac{kT}{q} \ln\left(\frac{N_D}{n_i}\right) > 0 \quad (3.5)$$

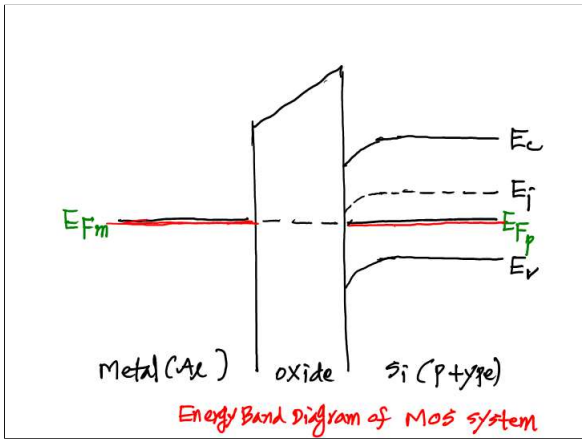
$q\chi$ = electron affinity = potential difference between the conduction band and the vacuum (free space)

$q\phi_s$ = work function = the energy required for an electron to move from the Fermi level to free space

$$q\phi_s = q\chi + E_C - E_F$$

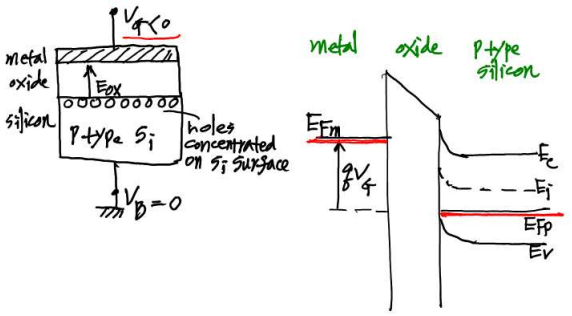


Energy Band diagram of individual components of MOS

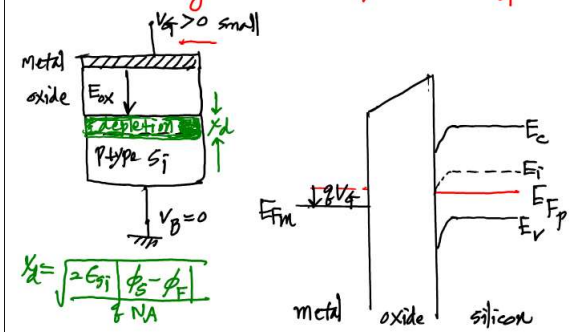


Energy Band Diagram of MOS system

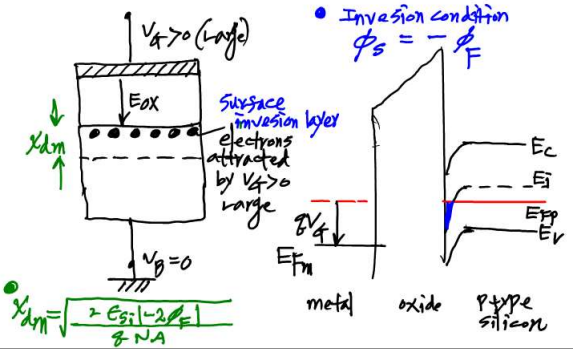
MOS system under External Bias

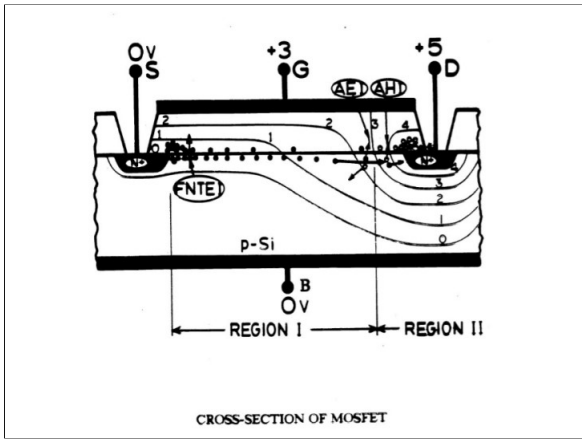


Band Diagram of MOS system with $V_G>0$

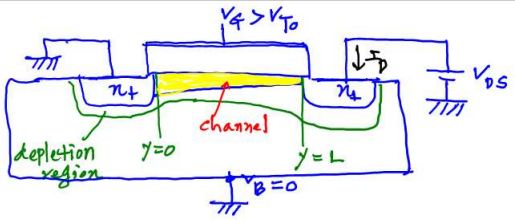


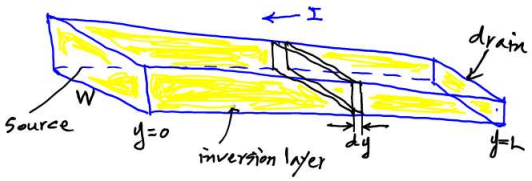
Energy Band Diagram of MOS system for $V_G>0$ (Large)





CROSS-SECTION OF MOSFET





$V_{GS} = V_G - V_S (= 0) \geq V_{T0}$

$V_{GD} = V_{GS} - V_{DS} \geq V_{T0}$

$Q_I(y)$ = total mobile electron charge in the surface inversion layer

$$Q_I(y) = -C_{ox}[V_{GS} - V_C(y) - V_{T0}] \quad (3.29)$$

dR = incremental resistance of the differential channel segment

$$dR = \frac{dy}{W \cdot \mu_n \cdot Q_I(y)} \quad (3.28)$$

where μ_n = surface electron mobility

W = channel width (transistor size)

$$-dV_C = I_D\, dR = \frac{I_D}{W \cdot \mu_n \cdot Q_I(y)}\, dy$$

$$-W \cdot \mu_n \cdot Q_I(y)\, dV_C = I_D\, dy$$

$$-W \cdot \mu_n \int_0^{V_{DS}} Q_I(y)\, dV_C = \int_0^L I_D\, dy$$

$$\Rightarrow I_D L = -W\mu_n \int_0^{V_{DS}} -C_{ox}(V_{GS} - V_C(y) - V_{T0})\, dV_C$$

$$= \mu_n C_{ox} W \int_0^{V_{DS}} (V_{GS} - V_C(y) - V_{T0})\, dV_C$$

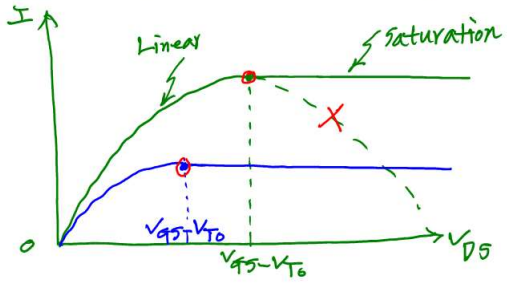
$$= \mu_n C_{ox} W \left[(V_{GS} - V_{T0})V_C - \frac{1}{2}V_C^2\right]\Big|_0^{V_{DS}}$$

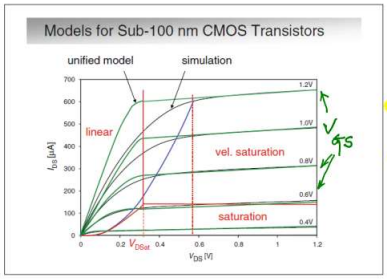
$$= \mu_n C_{ox} W \left[(V_{GS} - V_{T0})V_{DS} - \frac{1}{2}V_{DS}^2\right]$$

$$\Rightarrow \boxed{I_D = \mu_n C_{ox} \frac{W}{L}\left[(V_{GS} - V_{T0})V_{DS} - \frac{1}{2}V_{DS}^2\right]} \quad (3.32)$$

when $V_{DS} = V_{GS} - V_{T0}$ (saturation)

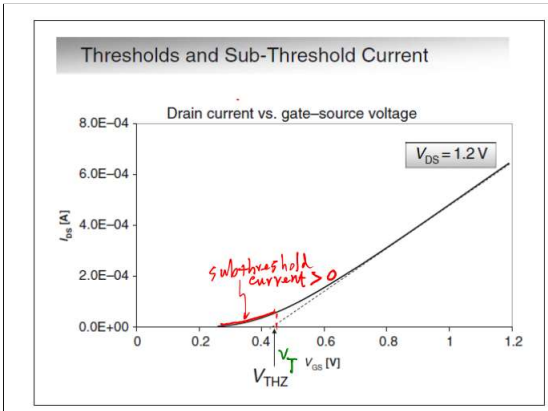
$$(3.32) \Rightarrow I = \mu_n C_{ox} \frac{W}{L} \frac{1}{2}(V_{GS} - V_{T0})^2$$

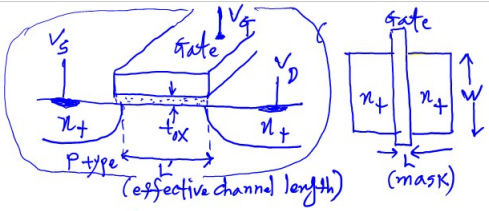




Slide 2.8
Simplicity comes at a cost however. Comparing the I–V curves produced by the model to those of the actual devices (BSIM-4 SPICE model), a large discrepancy can be observed for intermediate values of V_{DS} (around V_{DSat}). When using the model for the derivation of propagation delays (performance) 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 percents, which is only a small penalty for a major reduction in model complexity.

$$\text{delay} \propto \frac{\Delta Q}{I_{avg}}$$





$L' = L - \Delta L$

For long-channel NMOSTs,

$I_{DS} = 0$ for $V_{GS} < V_T$ (threshold voltage)

and $V_T = V_{T0} + \gamma\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right)$ (3.23)

$\gamma = \sqrt{2q N_A \epsilon_{Si}}/C_{ox}$, $C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$ per unit area

	NMOST	PMOST
ϕ_F substrate Fermi potential	−	+
γ substrate bias coefficient	+	−
V_{SB} substrate bias voltage	+	−

and

$V_{T0} = \phi_{GC} - 2\phi_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}}$ Eq. (3.19)

$Q_{B0} = -\sqrt{2q N_A \epsilon_{Si}|-2\phi_F|}$

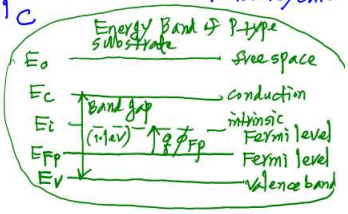
$Q_B = -\sqrt{2q N_A \epsilon_{Si}|-2\phi_F + V_{SB}|}$

$Q_{ox} = q N_{ox}$ (= oxide interface fixed charge density)

- $\epsilon_{Si} = 11.7\epsilon_0 = 11.7 \times (8.854 \times 10^{-14})$
- N_A = acceptor concentration (typically Boron)
- hole concentration in the p-type substrate (body)

  $p_{p0} \approx N_A$ (typically $10^{15} \sim 10^{16}/cm^3$, but can be much higher)
- electron concentration

  $n_{p0} \approx \frac{n_i^2}{N_A}$, n_i = intrinsic carrier concentration in Si (at T = 300°K, $1.45 \times 10^{10}/cm^3$)
- $q = 1.602 \times 10^{-19}$ C
- $\phi_F = \frac{E_F - E_i}{q}$ ($\phi_{Fp} < 0$)



(e.g.) for p type substrate $N_A = 4 \times 10^{18}/cm^3$

polysilicon gate doping $N_D = 2 \times 10^{20}/cm^3$

$\phi_F(\text{substrate}) = \frac{kT}{q} \ln\left(\frac{n_i}{N_A}\right)$

$= \frac{1.38 \times 10^{-23}\,[J/K] \times 300\,[K]}{1.6 \times 10^{-19}\,[C]} \ln\left(\frac{1.45 \times 10^{10}}{4 \times 10^{18}}\right)$

($T = 300°K$ at room temp)

$= 0.026\,[V] \times (-19.435) = -0.505\,[V]$

$\phi_F(\text{gate}) = \frac{kT}{q} \ln\left(\frac{N_D}{n_i}\right) = 0.026[V] \ln\left(\frac{2 \times 10^{20}}{1.45 \times 10^{10}}\right)$

(polysilicon)

$= 0.026[V] \times (+23.347) = +0.607\,[V]$

work function difference between gate and channel,

$\phi_{GC} = \phi_F(\text{substrate}) - \phi_F(\text{gate}) = -0.505 - 0.607$

$= -1.11\,[V]$

Depletion region charge density at $V_{SB} = 0$

$Q_{B0} = -\sqrt{2q N_A \epsilon_{Si}|-2\phi_F(\text{substrate})|}$

$= -\sqrt{2(1.6 \times 10^{-19})(4 \times 10^{18})(11.7 \times 8.85 \times 10^{-14})|2(-0.505)|}$

$= -1.16 \times 10^{-6}\,[C]/cm^2$

oxide interface charge

$Q_{ox} = q N_{ox} = 1.6 \times 10^{-19}[C] \times \underbrace{4 \times 10^{12}/cm^2}_{\text{specified}}$

$= 6.4 \times 10^{-9}\,[C]/cm^2$

$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{3.99 \times 8.85 \times 10^{-14}\,F/cm}{\underbrace{16 \times 10^{-8}\,cm}_{t_{ox} = 16\,Å}} = 2.20\,\mu F/cm^2$

$V_{T0} = \phi_{GC} - 2\phi_F(\text{substrate}) - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}}$

$= -1.11[V] - 2(-0.505) - \frac{-1.16 \times 10^{-6}\,[C]}{2.20 \times 10^{-8}\,[F]} - \frac{6.4 \times 10^{-9}\,[C]}{2.2 \times 10^{-8}\,[F]}$

$= -1.11 + 1.01 + 0.527 - 2.91\times10^{-3}$

$= 0.427\,[V] = V_{T0}$

$\gamma = \sqrt{2q N_A \varepsilon_{Si}}\,/\,C_{ox}$

$= \dfrac{\sqrt{2\times1.6\times10^{-19}[C]\cdot4\times10^{18}/cm^3\cdot11.7\times8.85\times10^{-14}\ F/cm}}{2.20\times10^{-6}\ F/cm^2}$

$= \dfrac{\sqrt{1325.376\times10^{-15}\ C\cdot F/cm^4}}{2.2\times10^{-6}\ F/cm^2} = \dfrac{11.51\times10^{-7}\ F/cm^2\sqrt{V}}{2.2\times10^{-6}\ F/cm^2}$

$= 0.523\,[\sqrt{V}]$

$V_T = V_{T0} + \gamma\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right)$ (3.23)

$= 0.427\,[V] + 0.523\,[\sqrt{V}]\left(\sqrt{|-2(-0.503)+1|\,[V]} - \sqrt{|2(-0.503)|\,[V]}\right)$

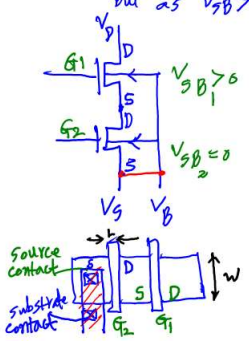
$\left(\phi_F(\text{substrate}) = -0.503\right)$ $= 0.427\,[V] + 0.523\,[\sqrt{V}]\left(\sqrt{2.01\,V} - \sqrt{1.01\,V}\right)$

$V_{SB} = 1\,V$ $= (0.427 + 0.523\times0.413)[V] = 0.64\,[V] = V_T$

From $V_T = V_{T0} + \gamma\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}\right)$

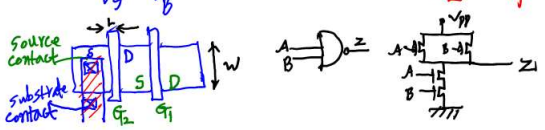
when $V_S = V_B$ $V_T = V_{T0}$

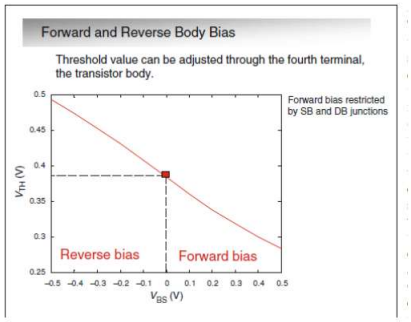
But as $V_{SB} > 0$ increase V_T ↑

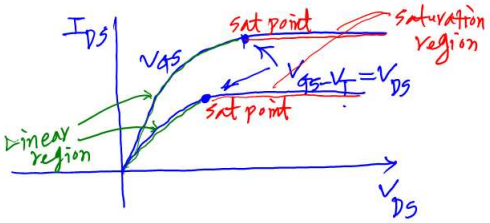


$V_{T_2} > V_{T_1} = V_{T0}$

Note that due to a Body Effect $V_{T_2} > V_{T_1}$







For Long channel NMOSTs via Gradual channel approx.

Linear region $I_{DS} = \mu_n C_{ox}\dfrac{W}{L}\left[2(V_{GS}-V_T)V_{DS} - V_{DS}^2\right]$ (3.34)

electron mobility — for $V_{GS} > V_T > 0$

Saturation region $I_{DS} = \mu_n C_{ox}\dfrac{W}{L}\left[2(V_{GS}-V_T)(V_{GS}-V_T) - (V_{GS}-V_T)^2\right]$

$(V_{GS}-V_T = V_{DS})$

$= \mu_n C_{ox}\dfrac{W}{L}(V_{GS}-V_T)^2$ (3.38)

Channel Length Modulation parameter λ

$L \leftarrow L' = L - \Delta L = L\left(1 - \dfrac{\Delta L}{L}\right)$

$\approx L(1 - \lambda V_{DS})$

λ = emperical parameter

With channel length modulation

(3.38) ← $I_{DS_{sat}} = \mu_n C_{ox}\dfrac{W}{L(1-\lambda V_{DS})}(V_{GS}-V_T)^2$

$= \mu_n C_{ox}(V_{GS}-V_T)^2(1+\lambda V_{DS})$ (3.49)

$\left(\dfrac{1}{1-\epsilon} = 1+\epsilon\right)$

In summary for NMOSTs

$I_{DS\,linear} = \dfrac{\mu_n C_{ox}}{2}\dfrac{W}{L}\left[2(V_{GS}-V_T)V_{DS} - V_{DS}^2\right]$ (3.51)

$I_{DS\,sat} = \dfrac{\mu_n C_{ox}}{2}\dfrac{W}{L}(V_{GS}-V_T)^2(1+\lambda V_{DS})$ (3.52)

$\times(1+\lambda V_{DS})$

Similarly for PMOSTa not μ_n (typo)

$$I_{SD\,linear} = \frac{\mu_p C_{ox}}{2}\frac{W}{L}\left[2(V_{GS}-V_T)V_{DS}-V_{DS}^2\right] \quad (3.58)$$

for $V_{GS}<V_T<0$ & $V_{DS}>V_{GS}-V_T$

which is same as

$$\frac{\mu_p C_{ox}}{2}\frac{W}{L}\left[2(V_{SG}+V_T)V_{SD}-V_{SD}^2\right]$$

$$I_{SD\,sat} = \frac{\mu_p C_{ox}}{2}\frac{W}{L}(V_{GS}-V_T)^2(1+\lambda V_{SD}) \quad (3.59)$$

$1-\lambda V_{DS}$, $V_{DS}<0$ (error in the book)

I-V Equations for short channel MOSTa

$$\mu_{n(eff)} = \frac{\mu_{no}}{1+\gamma(V_{GS}-V_T)} \quad (3.69)$$

γ = empirical coefficient

μ_{no} = low-field electron mobility

For NMOSTa

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

for $V_{GS}>V_T>0$ &

$$V_{DS} < \frac{(V_{GS}-V_T)E_c L}{(V_{GS}-V_T)+E_c L} \quad (3.85)$$

where E_c = channel electric field

With V_{sat} (saturated drift velocity of electrons)

$$I_{DS\,sat} = W V_{sat} C_{ox}\frac{(V_{GS}-V_T)^2}{(V_{GS}-V_T)+E_c L}(1+\lambda V_{DS}) \quad (3.86)$$

for $V_{GS} \gtrsim V_T$, $V_{DS} \gtrsim \frac{(V_{GS}-V_T)E_c L}{(V_{GS}-V_T)+E_c L}$

Similarly for pMOSTa

(3.87) & (3.88)

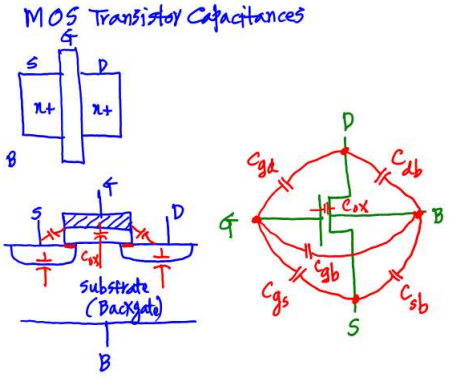
Threshold voltage of small Geometry Devices

$$V_T = V_{T0} + K_1\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}\right) + K_2 V_{SB} - \Delta T_{SCE} + \Delta T_{NWE} - \Delta V_{T\,DIBL} + \Delta V_{T,RSCE} - \Delta V_{T,DITS} \quad (3.116)$$

ΔT_{SCE}: short channel effect; ΔT_{NWE}: narrow width effect; $\Delta V_{T\,DIBL}$: drain induced Barrier Lowering; $\Delta V_{T,RSCE}$: (reverse SCE); $\Delta V_{T,DITS}$: Drain induced threshold shift

$$I_{DS\,subthreshold} \cong \frac{q D_n W x_c n_o}{L_B} e^{\frac{q\phi_t}{kT}} e^{\frac{q}{kT}(AV_{GS}+BV_{DS})} \quad (3.115)$$

MOS Transistor Capacitances



Diffusion Capacitances in MOST (C_{sb}, C_{db}) (p-n junctions)

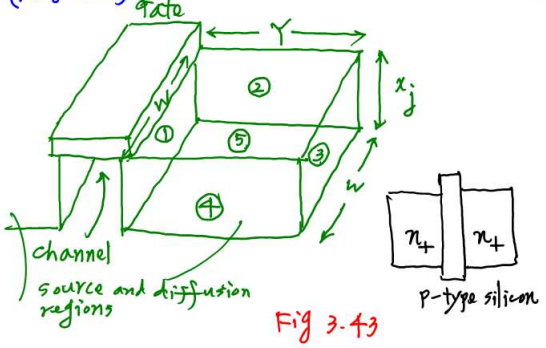
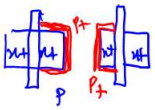


Fig 3.43

p-n Junction	Area (A)	Type
①	$W \cdot x_j$	n^+/p
②	$Y \cdot x_j$	$n^+/p \leftarrow p_+$
③	$W \cdot x_j$	$n^+/p \leftarrow p_+$
④	$Y \cdot x_j$	$n^+/p \leftarrow p_+$
⑤	$W \cdot Y$	n^+/p

channel stop



Depletion capacitance of a reversely biased "abrupt" p-n junction : (approximation)

$$x_d = \sqrt{\frac{2\epsilon_{si}}{q}\frac{N_A+N_D}{N_A \cdot N_D}(\phi_0 - V)} \quad (3.123)$$

where $\phi_0 = \frac{kT}{q}\ln\frac{N_A N_D}{n_i^2}$

and $V < 0$

$$Q_j = A \cdot q \cdot \frac{N_A N_D}{N_A+N_D} x_d \quad (3.124)$$

$$= A\sqrt{2\epsilon_{si} q \frac{N_A N_D}{N_A+N_D}(\phi_0 - V)} \quad (3.125)$$

$$C_j(V) = \left|\frac{dQ_j}{dV}\right|$$

$$= A\sqrt{\frac{\epsilon_{si} q}{2}\frac{N_A N_D}{N_A+N_D}}\frac{1}{\sqrt{\phi_0 - V}} \quad (3.127)$$

More generally, with junction grading

$$C_j(V) = \frac{A\, C_{j0}}{\left(1 - \frac{V}{\phi_0}\right)^m} \quad (3.128)$$

where m = grading coefficient
= $\frac{1}{2}$ abrupt
$\frac{1}{3}$ linearly graded profile

$C_{j0} = C_j(V)\big|_{V=0}$ in (3.127)

$$C_{eq} = \frac{\Delta Q}{\Delta V} = \frac{Q_j(V_2) - Q_j(V_1)}{V_2 - V_1}$$

$$= \frac{1}{V_2 - V_1}\int_{V_1}^{V_2} C_j(V)\, dV$$

$$= \frac{A\, C_{j0}\, \phi_0}{(V_2 - V_1)(1-m)}\left[\left(1 - \frac{V_2}{\phi_0}\right)^{1-m} - \left(1 - \frac{V_1}{\phi_0}\right)^{1-m}\right] \quad (3.131)$$

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

$$C_{eq} = \frac{2A\, C_{j0}\, \phi_0}{V_2 - V_1}\left[\sqrt{1 - \frac{V_2}{\phi_0}} - \sqrt{\frac{1 - V_1}{\phi_0}}\right] \quad (3.132)$$

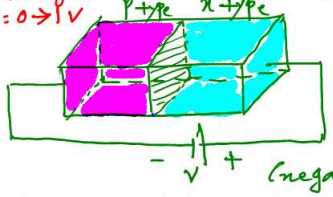
$$= A\, C_{j0}\, K_{eq}$$

where

$$K_{eq} = \frac{2\sqrt{\phi_0}}{V_2 - V_1}\left(\sqrt{\phi_0 - V_2} - \sqrt{\phi_0 - V_1}\right)$$

and $0 < K_{eq} < 1$

Find the equivalent p-n junction capacitance for $V: 0 \rightarrow -1\,V$



$N_D = 2.2 \times 10^{18}/cm^3$
$N_A = 1.6 \times 10^{18}/cm^3$
Junction Area
$A = 10\mu m \times 10\mu m$
(negative bias)

$$\phi_0 = \frac{kT}{q}\ln\frac{N_A N_D}{n_i^2}$$

$$= 0.026 \ln\left(\frac{1.6\times10^{18} \cdot 2.2\times10^{18}}{(1.45\times10^{10})^2}\right)$$

$$= 0.026 \times 37.43 = \underline{0.97\ [V]}$$

$$C_{j0} = \sqrt{\frac{\epsilon_{Si} q}{2}\left(\frac{N_A N_D}{N_A + N_D}\right)\frac{1}{\phi_0}}$$

$$= \sqrt{\frac{11.7 \times 8.85\times10^{-14}\,[F/cm] \times 1.6\times10^{-19}\,[C]}{2 \times 0.97}\left(\frac{1.6\times10^{18}\cdot 2.2\times10^{18}}{1.6\times10^{18} + 2.2\times10^{18}}\right)\left[\frac{1}{cm^3}\right]}$$

$$= 2.81 \times 10^{-7}\ F/cm^2$$

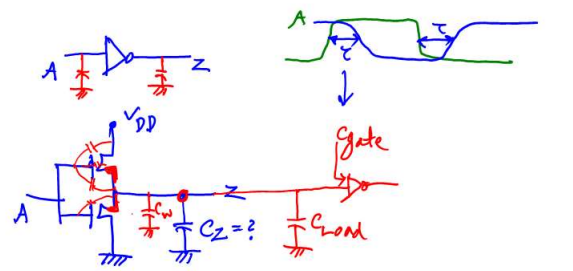
K_{eq} for V switching from 0 to 1 V

$$K_{eq} = \frac{2\sqrt{\phi_0}}{V_2 - V_1}\left(\sqrt{\phi_0 - V_2} - \sqrt{\phi_0 - V_1}\right) \quad (3.19)$$

$$= \frac{2\sqrt{0.97}}{-1-0}\left(\sqrt{0.97-(-1)} - \sqrt{0.97-0}\right)$$

$$= 0.82$$

$$C_{eq} = A\,C_{j0}\,K_{eq} = 10^{-3}cm^2 \times 2.81\times10^{-7}F/cm^2 \times 0.82 = 230\,fF$$



CAD tool extract capacitance from layout

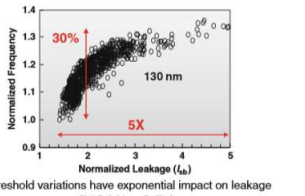
Variability

- Scaled device dimensions leading to increased impact of variations
  - Device physics
  - Manufacturing
  - Temporal and environmental
- Impacts performance, power (mostly leakage) and manufacturing yield
- More pronounced in low-power design due to reduced supply/threshold voltage ratios

Slide 2.32
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 easily understandable. When the supply voltage is 3 V and the threshold is at 0.5 V, a 25 mV 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 1 V and the threshold at 0.3 V, the same variation has a much larger impact.

So, in past generation processors it was sufficient to evaluate a design over its worst-case corners (FF, SS, FS, SF) 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.

Variability Impacts Leakage

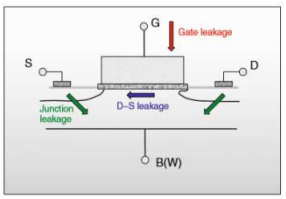


Threshold variations have exponential impact on leakage

[Ref: P. Gelsinger, DAC'04]

Slide 2.33
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 130 nm 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 5. Observe that the leakiest designs are also the ones with the highest performance (this should be no surprise).

MOS Transistor Leakage Components



Slide 2.17
Quite a number of times 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 digress from this ideal model.

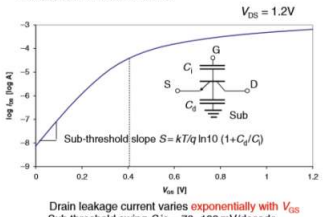
Leakage currents, flowing through the reverse-biased source–bulk and drain–bulk *pn* junctions, have always been present. Yet, the levels are so small that their effects could generally be ignored, except in circuitry that relies on charge storage such as DRAMs and dynamic logic.

The scaling of the minimum feature sizes has introduced some other leakage effects that are far more influential and exceed junction leakage currents by 3–5 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 Leakage

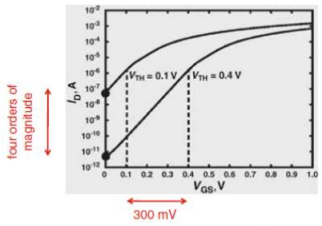
The transistor in "weak inversion"



Drain leakage current varies exponentially with V_{GS}
Sub-threshold swing S is ~ 70–100 mV/decade

Slide 2.18
In earlier slides, we have alluded to a relationship between the value of the threshold voltage V_{TH} and (sub-threshold) leakage. When the gate voltage of a transistor is lowered below the threshold voltage, the transistor does not turn off instantaneously. In fact, the transistor enters the so-called "sub-threshold regime" (or weak inversion). In this operation mode, the drain–source current becomes an exponential function of V_{GS}. This is clearly observed from the I_D–V_{GS} curves, if the current is plotted on a logarithmic scale.

Impact of Reduced Threshold Voltages on Leakage



Leakage: sub-threshold current for $V_{GS} = 0$

Sub-threshold Current

- Sub-threshold behavior can be modeled physically

$$I_{DS} = 2n\mu C_{ox}\frac{W}{L}\left(\frac{kT}{q}\right)^2 e^{\frac{V_{GS}-V_{TH}}{nkT/q}}\left(1-e^{\frac{-V_{DS}}{kT/q}}\right) = I_S e^{\frac{V_{GS}-V_{TH}}{nkT/q}}\left(1-e^{\frac{-V_{DS}}{kT/q}}\right)$$

where n is the *slope factor* (≥1, typically around 1.5) and $I_S = 2n\mu C_{ox}\frac{W}{L}\left(\frac{kT}{q}\right)^2$

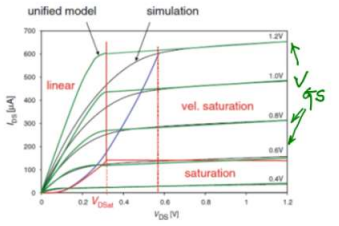
- Very often expressed in base 10

$$I_{DS} = I_S 10^{\frac{V_{GS}-V_{TH}}{S}}\left(1-10^{\frac{-nV_{DS}}{S}}\right)$$

≈1 for V_{DS} > 100 mV

where $S = n\left(\frac{kT}{q}\right)\ln(10)$, the *sub-threshold swing*, ranging between 60 mV and 100 mV

Models for Sub-100 nm CMOS Transistors



Slide 2.8
Simplicity comes at a cost however. Comparing the I–V curves produced by the model to those of the actual devices (BSIM-4 SPICE model), a large discrepancy can be observed for intermediate values of V_{DS} (around V_{DSat}). When using the model for the derivation of propagation delays (performance) 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 percents, which is only a small penalty for a major reduction in model complexity.

Alpha Power Law Model

- Alternate approach, useful for hand analysis of propagation delay

$$I_{DS} = \frac{W}{2L}\mu C_{ox}(V_{GS}-V_{TH})^{\alpha}$$

- Parameter α is between 1 and 2.
- In 65–180 nm CMOS technology α ~ 1.2–1.3

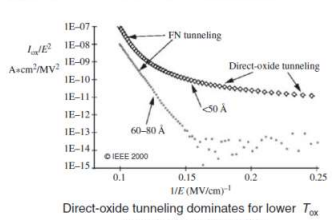
- This is not a physical model
- Simply empirical:
  - Can fit (in minimum mean squares sense) to a variety of α's, V_{TH}
  - Need to find one with minimum square error – fitted V_{TH} can be different from physical

[Ref: Sakurai, JSSC'90]

Slide 2.9
Even simpler is the *alpha model*, introduced by Sakurai and Newton in 1990, which does not even attempt to approximate the actual I–V curves. The values of α and V_{TH} are purely empirical, chosen such that the propagation delay of a digital gate, approximated by $t_p = \frac{kV_{DD}}{(V_{DD}-V_{TH})^{\alpha}}$, best resembles 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 it is up to the modeler to find the ones with the best fit.

Owing to its simplicity, the alpha model is the corner stone of the optimization framework discussed in later chapters.

Gate-Leakage Mechanisms

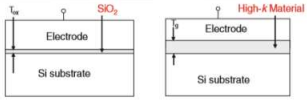


Direct-oxide tunneling dominates for lower T_{ox}

[Ref: Chandrakasan-Bowhill, Ch3, '00]

Slide 2.26
Gate leakage finds its source in two different mechanisms: Fowler–Nordheim (FN) tunneling, and direct-oxide tunneling. FN tunneling is an effect that has been effectively used in the design of non-volatile memories, and is already quite substantial for oxide thickness larger than 6 nm. Its onset requires high electric-field strengths, though. With reducing oxide thicknesses, tunneling starts to occur at far lower field strengths. The dominant effect under these conditions is direct-oxide tunneling.

High-*k* Gate Dielectric



- Equivalent Oxide Thickness = EOT = $T_{ox} = T_g^* (3.9/\varepsilon_g)$, where 3.9 is relative permittivity of SiO_2 and ε_g is relative permittivity of high-*k* material
- Currently SiO_2/Ni; Candidate materials: HfO_2 (ε_{eff}~15–30); $HfSiO_x$ (ε_{eff}~12–16)
  - Often combined with metal gate

Reduced Gate Leakage for Similar Drive Current

Slide 2.28
The MOS transistor current is proportional to the *process transconductance* parameter $k' = \mu C_g = \mu\varepsilon/t_g$. To increase k' 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 T_g) running out of steam, the only remaining option is to look for gate dielectrics with a higher permittivity ϵ – the so-called high-*k*

This was the (First) Era of Geometrical (classical) Scaling. This type of scaling was the foundation of the National Technology Roadmap for Semiconductors (NTRS) initiated in 1991.

The ITRS laid out the foundations of the (Second) Era: Equivalent Scaling (e.g., strained silicon, high-K/metal gate, Multigate transistors and use of non-silicon semiconductors in general) between 1998 and 2000.

Fig. 3.1 The Ideal MOS Transistor

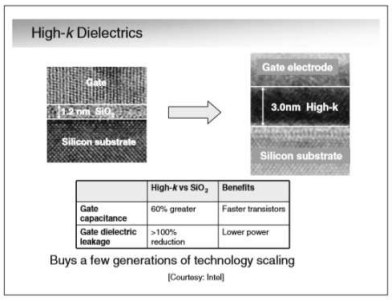
The implementation of these technologies successfully supported the growth of the semiconductor industry in the past decade and it will continue to do so until the end of the present decade and beyond.

In the next decade ITRS 2.0 predicts that the advent of the third phase of scaling ***"3D Power Scaling"*** will become the driver of the rejuvenated semiconductor industry and this ***answers the question*** posed before about the future of the semiconductor industry: ***"Yes the semiconductor industry will continue to be a key enabler of the IoE"***

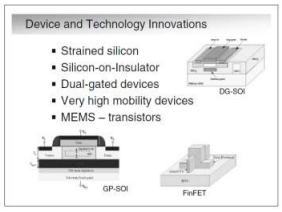
Moore's Law is now entering into a third phase characterized by vertical integration and performance specifications driven towards reduction of power in either the active or the stand-by modes.

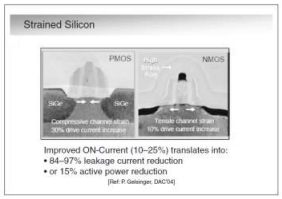
Electrical Properties of Transistors

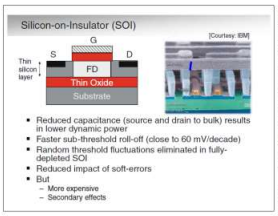
YEAR OF PRODUCTION	2015	2017	2019	2021	2024	2027	2030
Logic device technology naming	P70M56	P48M36	P42M24	P32M20	P24M12G1	P24M12G2	P24M12G3
Logic industry "Node Range" Labeling (nm)	"16/14"	"11/10"	"8/7"	"6/5"	"4/3"	"3/2.5"	"2/1.5"
Logic device structure options	FinFET FDSOI	FinFET FDSOI	FinFET LGAA	FinFET LGAA VGAA	VGAA, M3D	VGAA, M3D	VGAA, M3D
DEVICE ELECTRICAL SPECS							
Power Supply Voltage - Vdd (V)	0.80	0.75	0.70	0.65	0.55	0.45	0.40
Subthreshold slope - [mV/dec]	75	70	68	65	40	25	25
Inversion layer thickness - [nm]	1.10	1.00	0.90	0.85	0.80	0.80	0.80
Vt sat (mV) at Ioff=100nA/μm - HP Logic	129	129	133	136	84	52	52
Vt sat (mV) at Ioff=100pA/μm - LP Logic	351	336	333	326	201	125	125
Effective mobility (cm2/V-s)	200	150	120	100	100	100	100
Rext (Ohms.μm) - HP Logic	280	238	202	172	146	124	106
Ballisticity, Injection velocity (cm/s)	1.20E-07	1.32E-07	1.45E-07	1.60E-07	1.76E-07	1.93E-07	2.13E-07
Vdsat (V) - HP Logic	0.115	0.127	0.136	0.128	0.141	0.155	0.170
Vdsat (V) - LP Logic	0.125	0.141	0.155	0.153	0.169	0.186	0.204
Ion (μA/um) at Ioff=100nA/μm - HP logic w/ Rext=0	2311	2541	2782	2917	3001	2670	2408
Ion (μA/μm) at Ioff=100nA/μm - HP logic, after Rext	1177	1287	1397	1476	1546	1456	1391
Ion (μA/μm) at Ioff=100pA/μm - LP logic w/ Rext=0	1455	1567	1614	1603	2008	1933	1582
Ion (μA/μm) at Ioff=100pA/μm - LP logic, after Rext	596	637	637	629	890	956	821
Cch, total (fF/μm2) - HP/LP Logic	31.38	34.52	38.35	40.61	43.14	43.14	43.14
Cgate, total (fF/μm) - HP Logic	1.81	1.49	1.29	0.97	1.04	1.04	1.04
Cgate, total (fF/μm) - LP Logic	1.96	1.66	1.47	1.17	1.24	1.24	1.24
CV/I (ps) - FO3 load, HP Logic	3.69	2.61	1.94	1.29	1.11	0.96	0.89
it(CV) /(Ips) - FO3 load, HP Logic	0.27	0.38	0.52	0.78	0.90	1.04	1.12
Energy per switching [CV2] (fJ/switching) - FO3 load, HP Logic	3.47	2.52	1.89	1.24	0.94	0.63	0.50



Slide 2.29
The advantages offered by high-*k* gate dielectrics are quite clear: faster transistors and/or reduced gate leakage.



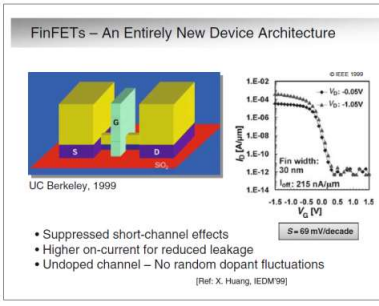




Slide 2.43
Silicon-on-Insulator (SOI) is a technology that has been "on the horizon" for quite a long time, yet it never managed to really break ground, though with some exceptions here and there. An SOI MOS transistor differs from a "bulk" device in that the channel is formed in a thin layer of silicon deposited above an electrical insulator, typically silicon dioxide.

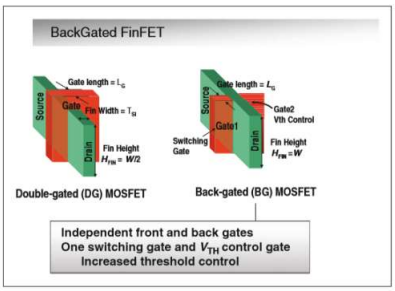
Doing so offers some attractive features. First, as drain and source diffusions extend all the way down to the insulator layer, their junction capacitances are substantially reduced, which translates directly into power savings. Another advantage is the higher sub-threshold slope factor (approaching the ideal 60 mV/decade), reducing leakage. Finally, the sensitivity to soft errors is reduced owing to the smaller collection efficiency, leading to a more reliable transistor. There are some important negatives as well. The addition of the SiO_2 layer and the thin silicon layer increases the cost of the substrate material, and may impact the yield as well. In addition, some secondary effects should be noted. The SOI transistor is essentially a three-terminal device without a bulk (or body) contact, and a "body" that is floating. This effectively eliminates body biasing as a threshold-control technique. The floating transistor body also introduces some interesting (ironically speaking, …) features such as hysteresis and state-dependency.

Device engineers differentiate between two types of SOI transistors: partially-depleted (PD-SOI) and fully-depleted (FD-SOI). In the latter, the silicon layer is so thin that it is completely depleted under nominal transistor operation, which means that the depletion/inversion layer under the gate extends all the way to the insulator. This has the advantage of suppressing some of the floating-body effects, and an ideal sub-threshold slope is theoretically achievable. From a variation perspective, the threshold voltage becomes independent of the doping in the channel, effectively eliminating a source of random variations (as discussed in Slide 2.37). FD-SOI requires the depositing of extremely thin silicon layers (3–5 times thinner than the gate length!).



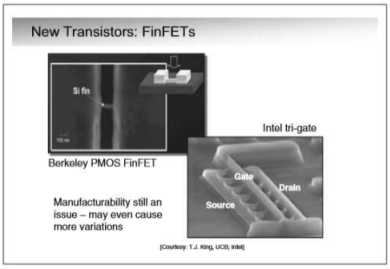
Slide 2.45
The FinFET (called a tri-gate transistor by Intel) is an entirely different transistor structure that actually offers some properties similar to the ones offered by the device presented in the previous slide. The term FinFET was coined by researchers at the University of California at Berkeley to describe a non-planar, double-gated transistor built on an SOI substrate. The distinguishing characteristic of the FinFET is that the controlling gate is wrapped around a thin silicon "fin", which forms the body of the device. The dimensions of the fin determine the effective channel length of the device. The device structure has shown the potential to scale the channel length to values that are hard, if not impossible, to accomplish in traditional planar devices. In fact, operational transistors with channel lengths down to 7 nm have been demonstrated.

In addition to a suppression of deep submicron effects, a crucial advantage of the device is again increased control, as the gate wraps (almost) completely around the channel.



Slide 2.46

This increased two-dimensional control can be exploited in a number of ways. In the dual-gated device, the fact that the gate is controlling the channel from both sides (as well as the top) leads to increased process transconductance. Another option is to remove the top part of the gate, leading to the back-gated 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 buried-gate FD-SOI transistor discussed earlier. Controlling the work functions of the two gates through the selection of appropriate type and quantity of the 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 MOS transistor 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 question, 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.