EE222 Lecture #2 January 11, 2018

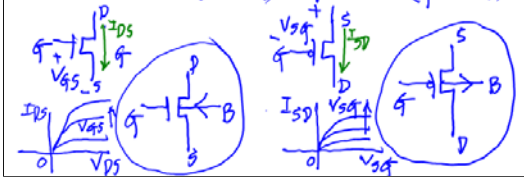
Instructor Steve Kang (http://nis1.soe.ucsc.edu)

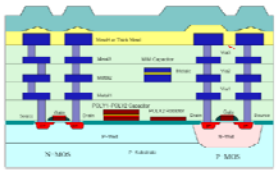
Website https://ee222-winter18-01.courses.soe.ucsc.edu

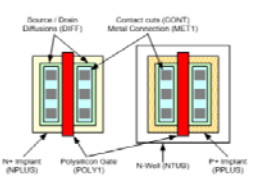
Webcast https://webcast.ucsc.edu (to be available)

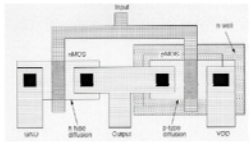
MOS Transistors Chap 3 (pp 5-78; 79-104)

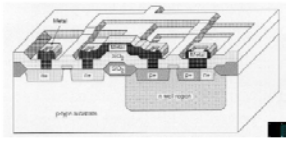
n-channel (NMOST) p-channel (pMOST)

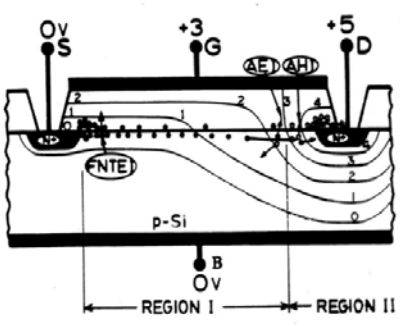




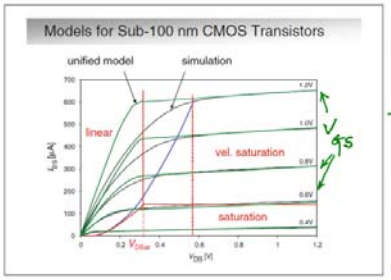






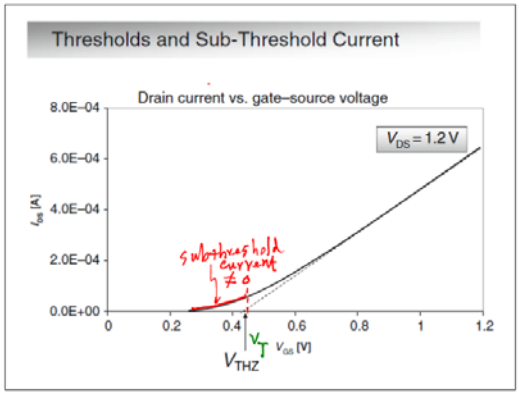


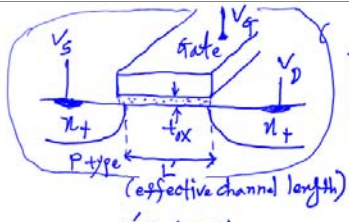
CROSS-SECTION OF MOSFET



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.





$L' = L - \Delta L$

For long-channel NMOSTs,

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

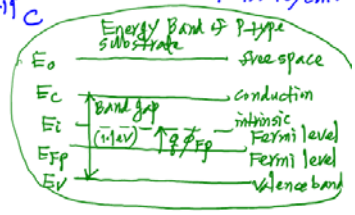
and $V_T = V_{TO} + \gamma\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}\right)$ (3.23)

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

- $\varepsilon_{Si} = 11.7\,\varepsilon_0 = 11.7 \times (8.854 \times 10^{-14})$
- N_A = acceptor concentration (typically Boron)
- hole concentration in the p-type substrate (body)

  $p_{p_0} \approx N_A$ (typically $10^{15} \sim 10^{16}/cm^3$, but can be much higher)
- electron concentration

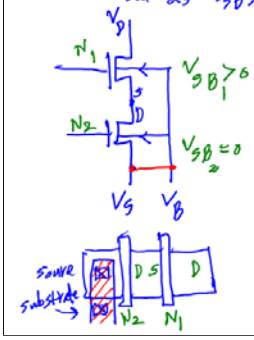
  $n_{p_0} \approx \frac{n_i^2}{N_A}$, n_i = intrinsic carrier concentration in Si (At $T = 300^\circ K$, $1.45 \times 10^{10}/cm^3$)
- $q = 1.602 \times 10^{-19}\,C$
- $\phi_F = \frac{E_F - E_i}{q}$ $(\phi_{Fp} < 0)$



From $V_T = V_{T_0} + \gamma\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}\right)$

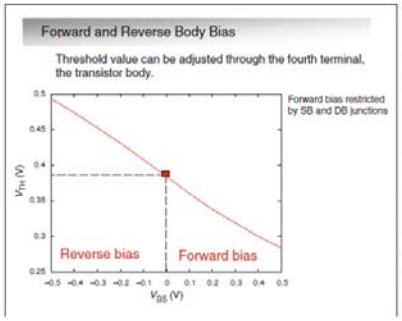
when $V_S = V_B$ $V_T = V_{T_0}$

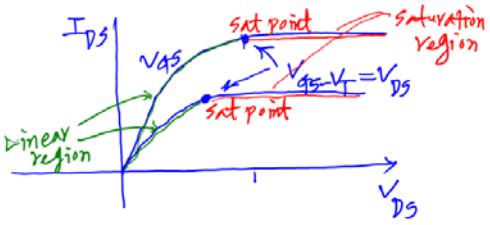
But as $V_{SB} > 0$ increase $V_T \uparrow$



$V_{T_2} > V_{T_1} = V_{T_0}$

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





For Long channel NMOSTs,

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

(μ_n = electron mobility; for $V_{GS} > V_T > 0$)

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

$V_{DS\,sat} = V_{GS} - V_T$

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

Channel Length Modulation parameter λ

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

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

λ = emperical parameter

With channel length modulation

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

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

$\left(\frac{1}{1-\epsilon} \approx 1 + \epsilon\right)$

In summary for NMOSTs

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

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

Similarly for PMOSTa — not μ_n (typo)

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

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

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

$-\Delta V_{T,SCE}$: short channel effect

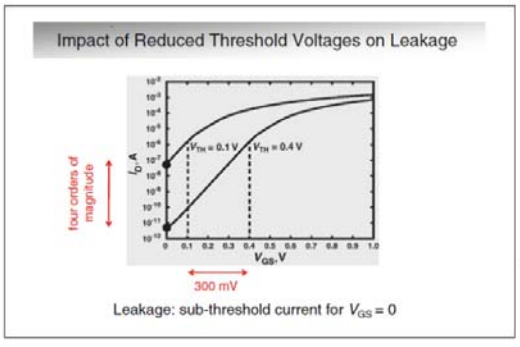
$\Delta V_{T,NWE}$: narrow width effect

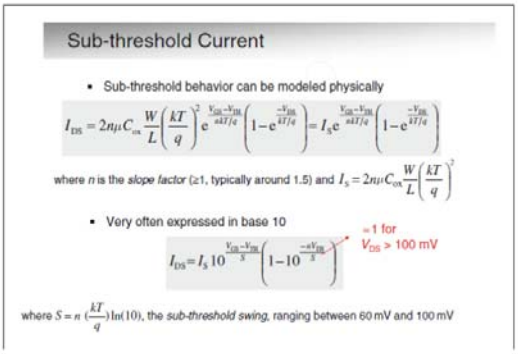
$-\Delta V_{T,DIBL}$: drain induced barrier lowering

$\Delta V_{T,RSCE}$: (reverse SCE)

$-\Delta V_{T,DITS}$: Drain induced threshold shift

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





Alpha Power Law Model

- Alternate approach, useful for hand analysis of propagation delay

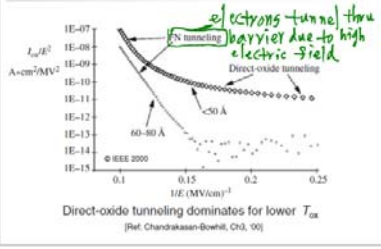
$$I_{DS} = \frac{W}{2L}\mu C_{ox}(V_{GS} - V_{TH})^{\alpha}$$

- Parameter α is between 1 and 2.
- In 65–180 nm CMOS technology $\alpha \sim 1.2$–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]

Gate-Leakage Mechanisms



Direct-oxide tunneling dominates for lower T_{ox}

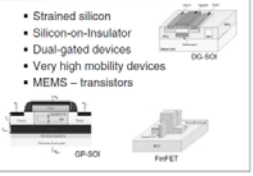
[Ref: Chandrakasan-Bowhill, Ch3, '00]

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.

Device and Technology Innovations

- Strained silicon
- Silicon-on-Insulator
- Dual-gated devices
- Very high mobility devices
- MEMS – transistors

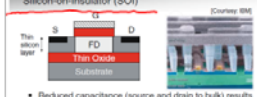


Strained Silicon

Improved ON-Current (10–25%) translates into:
- 84–97% leakage current reduction
- or 15% active power reduction

[Ref: P. Gelsinger, DAC'04]

Silicon-on-Insulator (SOI)



- Reduced capacitance (source and drain to bulk) results in lower dynamic power
- Faster sub-threshold roll-off (close to 60 mV/decade)
- Random threshold fluctuations eliminated in fully-depleted SOI
- Reduced impact of soft-errors
- But
  - More expensive
  - Secondary effects

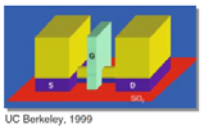
Slide 2.43

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.

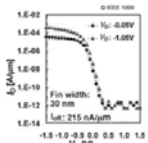
Doing so offers some attractive features. First, as drain and source diffusions extend all the way down to the insulator layer, their junction capacitances are substantially reduced, which translates directly into power savings. Another advantage is the higher sub-threshold slope factor (approaching 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 normal 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!).

FinFETs – An Entirely New Device Architecture



UC Berkeley, 1999



- Suppressed short-channel effects
- Higher on-current for reduced leakage
- Undoped channel – No random dopant fluctuations

[Ref: X. Huang, IEDM'99]

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.

Illustration: Evoluation of Microprocessors

Teraflops Research Chip
Introduced 2006
65nm Technology

80 Processor cores

• 3.16 GHz	62W	1.0 Tflops
• 5.1 GHz	175W	1.6 TFlops
• 5.7 GHz	265W	1.8 TFlops

For comparison: ASCI Red was the first supercomputer to reach Tflops in 1996. That system used nearly 10,000 Pentium® Pro processors running at 200MHz and consumed 500kW of power plus an additional 500kW just to cool the room that housed it.