MOS Transistors, Chapter 3 (PP. 20-28, 44-46)

1. Channel (n-channel) and (p-channel)

Slide 2.8
Simplex curves at a cost function. Comparing the I-V curves produced by the model to those of the actual device (Figure 28), Sketch model, a large discrepancy can be observed for intermediate values of VGS (near VGS). When using the model for the estimation 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 relation of current at the highest value of VGS and VGS are predicted correctly, as these parameters determine the charge and discharge times of the output capacitance. Hence, the propagation delay error is only a couple of percent, which is only a small penalty for a major reduction in model complexity.

For long-channel nMOSFETs,

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

and

\[ V_{TH} = V_{TO} + \gamma \left( \sqrt{1 - \frac{V_{DS}}{V_{TH}}} - \sqrt{1 - \frac{V_{GS}}{V_{TH}}} \right) \]

\[ \gamma = \frac{1}{2} \varepsilon_{Si} \frac{W}{L} \frac{C_{ox}}{K_{ox}} \]

Thresholds and Sub-Threshold Current

CROSS-SECTION OF MOSFET

Models for Sub-100 nm CMOS Transistors
For long channel NMOSFETs, the current density can be expressed as:

$$I_{DS} = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \left[ V_{GS} - V_{TH} \right] \left[ V_{DS} - V_{TH} \right]$$ (3.24)

For saturation region, the current density is given by:

$$V_{GS} = V_{TH} + \frac{V_{DS}}{2}$$

Channel Length Modulation Parameter \( \lambda \):

$$\lambda = \frac{L}{L_d} = \left( 1 - \frac{L}{L_d} \right)$$

\( \lambda \) = empirical parameter

Channel Length Modulation Parameter \( \lambda \):

$$\lambda = \frac{L}{L_d} = \left( 1 - \frac{L}{L_d} \right)$$

\( \lambda \) = empirical parameter

From \( V_{T2} = V_{TH} + \frac{1}{2} \left[ V_{GS} - V_{TH} \right] \):

$$V_{T2} = \frac{1}{2} \left( V_{GS} - V_{TH} \right)$$

Note that due to body effect, \( V_{T2} > V_{T1} \)

With channel length modulation:

$$I_{DS} = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \left[ V_{GS} - V_{TH} \right] \left[ V_{DS} - V_{TH} \right]$$ (3.24)

$$V_{GS} = V_{TH} + \frac{V_{DS}}{2}$$

\( V_{DS} \) in saturation region:

$$V_{DS} = \frac{1}{2} \left( V_{GS} - V_{TH} \right) \left( 1 + \frac{1}{\lambda} \right)$$ (3.29)

In summary for NMOSFETs:

$$I_{DS} = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \left[ V_{GS} - V_{TH} \right] \left[ V_{DS} - V_{TH} \right]$$ (3.24)

$$I_{DS} = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \left[ V_{GS} - V_{TH} \right] \left( 1 + \frac{1}{\lambda} \right)$$ (3.29)
Similarly for PMOS 

\[ I_{D, \text{linear}} = \frac{M_+ \cdot C_{ox} \cdot W}{2} \left[ \frac{(V_{GS}-V_T)^2}{V_{DS}} \right] \]  

(3.58)

which is same as

\[ \frac{M_- \cdot C_{ox} \cdot W}{2} \left[ 2(V_{DS} + V_T) \right] \]  

(3.59)

\[ I_{D, \text{sat}} = \frac{M_+ \cdot C_{ox} \cdot W}{2} \left( V_{GS} - V_T \right) \left( 1 + \frac{V_{DS}}{V_{TH}} \right) \]  

(3.60)

\[ V_{DS} < \frac{(V_{GS} - V_T) \cdot E_L}{E_F} \]  

(3.61)

For NMOS 

\[ I_{D, \text{linear}} = \frac{1}{2} \frac{M_- \cdot C_{ox} \cdot W}{L} (V_{GS} - V_T) \]  

(3.62)

\[ V_{DS} > \frac{(V_{GS} - V_T) \cdot E_L}{E_F} \]  

(3.63)

where \( E_F \) = channel electric field

\[ \frac{V_{DS}}{2} \]  

(3.64)

\[ V_{TH} = V_{PD} + K_N \left( \left( \frac{V_{PD} + V_{DS}}{V_{TH}} \right) - \sqrt{2 \cdot V_{PD}} \right) \]  

(3.65)

\[ - \frac{\Delta V_{TH}}{\Delta V_{DS}} = - \frac{\Delta V_{TH}}{\Delta V_{DS}} \]  

(3.66)

\[ \Delta V_{TH} = \frac{V_{DS}}{2} \]  

(3.67)

\[ I_{D, \text{subthreshold}} = \frac{2 \cdot B \cdot W \cdot n_+ \cdot e^{\frac{V_{DS}}{V_T}}}{L} \left( \frac{V_{TH}}{V_{TH} + V_{DS}} \right) \]  

(3.68)

\[ \text{Sub-threshold Current} \]

- Sub-threshold behavior can be modeled physically

\[ I_{th} = \frac{2 \cdot \pi \cdot C_{ox} \cdot W}{L} \left( \frac{V_{DS}}{V_{TH}} \right) \left( 1 - e^{-\frac{V_{DS}}{V_T}} \right) \]  

(3.69)

where \( N \) is the slope factor (typically around 1.0) and \( I_{th} = 2 \cdot \pi \cdot C_{ox} \cdot W \left( \frac{V_{DS}}{V_{TH}} \right) \left( 1 - e^{-\frac{V_{DS}}{V_T}} \right) \)

- Very often expressed in base 10

\[ I_{th} \]  

(3.70)

where \( S = (\frac{V_{DS}}{V_T}) \), the sub-threshold swing, ranging between 50 mV and 100 mV
Alpha Power Law Model

- Alternate approach, useful for hand analysis of propagation delay

\[ I_{DS} = \frac{V_{DS}}{2L} \cdot C_{ox} (V_{DS} - V_T) \]

- Parameter \( \alpha \) is between 1 and 2.
- In 65-180 nm CMOS technology \( \alpha \approx 1.5-1.8 \)

- This is not a physical model
- Simply empirical:
  - Can fit (in minimum square sense) to a variety of \( \alpha \), \( V_T \)
  - Need to find one with minimum square error — fitted \( V_T \) can be different from physical

Gate-Leakage Mechanisms

- Direct-oxide tunneling dominates for lower \( T_u \)
- Impact ionization tunneling for higher \( T_u \)

in direct-oxide tunneling.

Device and Technology Innovations

- Strained silicon
- Silicon on insulator
- Dual-gate devices
- Very high mobility devices
- MEMS - transistors

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 7 nm. In contrast, direct-oxide tunneling starts to occur at far lower field strengths. The dominant effect under these conditions

Silicon-on-Insulator (SOI)

- Is a technology that has been around for quite a long time and has been managed to be highly successful, though with some technical issues and commercial problems. The SOI process eliminates the 

- Impact ionization tunneling dominates for higher \( T_u \)
- Direct-oxide tunneling dominates for lower \( T_u \)

Direct or deep submicron effects are crucial to understanding and designing successful devices and circuits. These effects include:

- Impact ionization tunneling
- Direct-oxide tunneling

Slide 2.45

The FinFET, also called a tri-gate transistor by IBM, 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 novel, double-gate transistor built on an SOI substrate. The distinguishing characteristics of the FinFET is that the surrounding gate is wrapped around the body of the device. The dimensions of the FinFET 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 achieve in traditional planar devices. In fact, exponential 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 of the gate structure, so that the gate wraps completely around the channel.

Illustration: Evolution of Microprocessors

TeraFLOPS Research Chip

- Introduced 2006
- 65nm Technology

- 85 Processor cores
- 3.16 GHz 62W 1.0 TLOPs
- 5.1 GHz 175W 1.6 TLOPs
- 5.7 GHz 265W 1.8 TLOPs

For comparison: A911Red was the first supercomputer to reach TLOPs in 1994. That system used nearly 10,000 Pentium Pro processors running at 400 MHz and consumed 50 kW of power plus an additional 520 kW just to cool the room that it housed.