TAIWAN TECH National Taiwan University of Science and Technology

Introduction to Analog Integrated Circuit Design

Fall 2023

Short Channel Effects

Yung-Hui Chung

MSIC Lab

DECE, NTUST

Mixed-Signal IC Laboratory NTUST

Outline

- Scaling Theory on CMOS VLSI
- Short Channel Effects

- MOS Device Models
- Process Corners

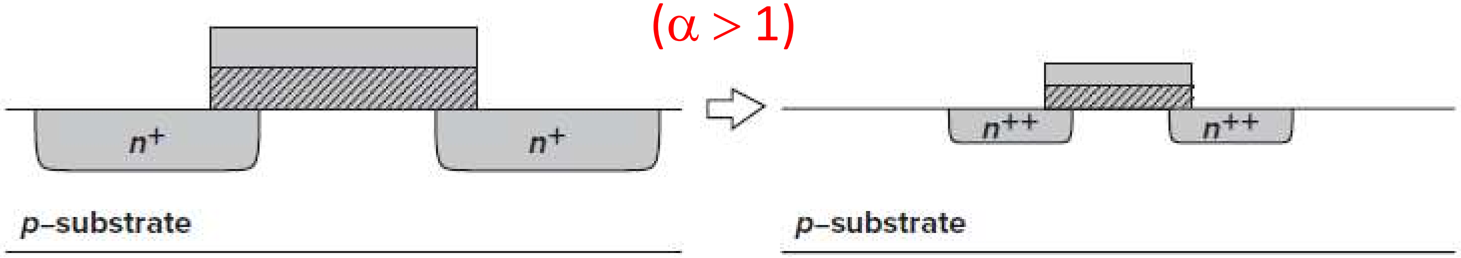
Referred to Textbook, chapter 17

MOS Transistor Scaling (1)

The two principal reasons for the dominance of CMOS technology in today’s semiconductor industry are

(1) the zero static power dissipation of CMOS logic and

(2) the scalability of MOSFETs

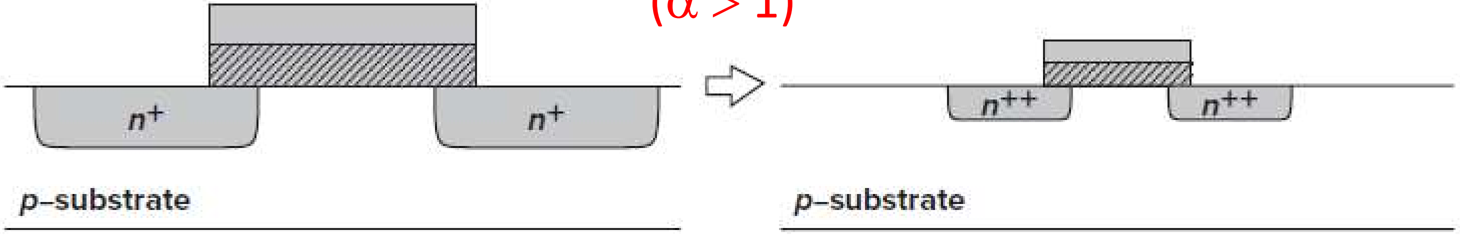


- **Definition:** Scaling factor *α*
- **The ideal scaling theory follows three rules:**
  (1) To reduce all lateral and vertical dimensions by *α* (> 1)
  (2) To reduce the threshold voltage and the supply voltage by *α*
  (3) To increase all of the doping levels by *α* (Fig. 17.1)
- **The above is called “Constant-Field Scaling”**

MOS Transistor Scaling (2)

- **CMOS**: **Zero** static power consumption and **Scalability** of MOSFETs

(α > 1)



$$I_{D,scaled} = \frac{1}{2}\mu_n(\alpha C_{ox})\left(\frac{W/\alpha}{L/\alpha}\right)\left(\frac{V_{GS}}{\alpha} - \frac{V_{TH}}{\alpha}\right)^2$$

$$= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2\frac{1}{\underline{\alpha}}$$

The total channel capacitance,

$$C_{ch,scaled} = \frac{W}{\alpha}\frac{L}{\alpha}(\alpha C_{ox})$$

$$= \frac{1}{\underline{\alpha}}WLC_{ox}$$

$$W_d = \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)(\phi_B + V_R)} \xrightarrow{V_R \gg \phi_B}$$

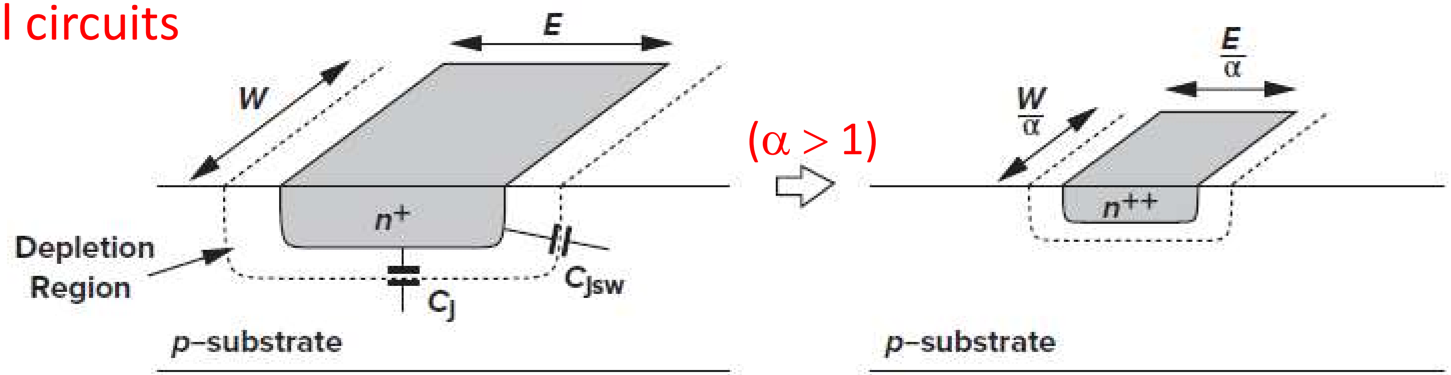
$$W_{d,scaled} \approx \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{\alpha N_A} + \frac{1}{\alpha N_D}\right)\frac{V_R}{\alpha}}$$

$$\approx \frac{1}{\underline{\alpha}}\sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)V_R}$$

W_d: Width of depletion region

MOS Transistor Scaling (3)

Digital circuits



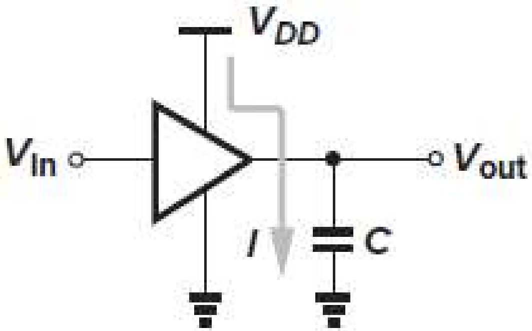
$$C_{S/D,scaled} = \frac{W}{\alpha}\frac{E}{\alpha}(\alpha C_j) + 2\left(\frac{W}{\alpha} + \frac{E}{\alpha}\right)(C_{jsw})$$

$$= [WEC_j + 2(W+E)C_{jsw}]\frac{1}{\underline{\alpha}}$$

$$P = fCV^2$$

$$P_{scaled} = f\left(C/\alpha\right)\left(V/\alpha\right)^2$$

$$= \underline{\frac{1}{\alpha^3}fCV^2}$$

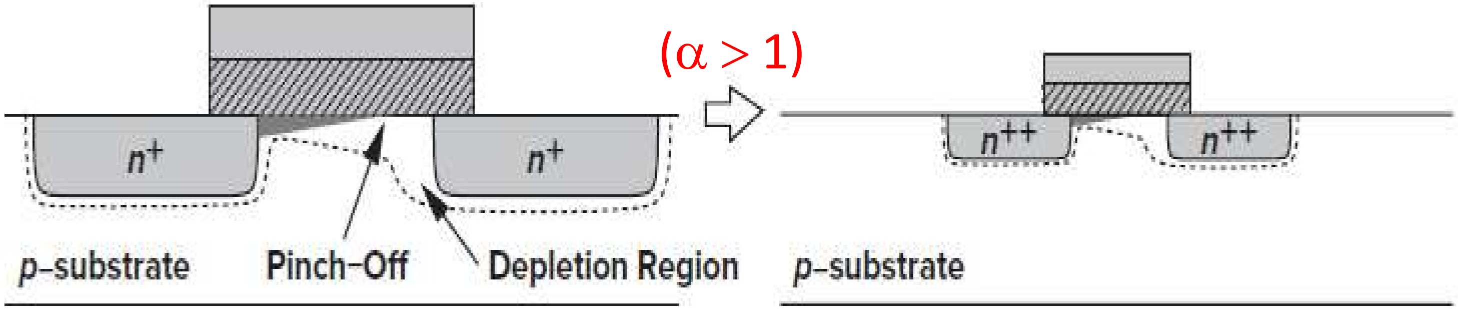


$$T_{d,scaled} = \frac{C/\alpha}{I/\alpha}\frac{V_{DD}}{\alpha}$$

$$= \left(\frac{C}{I}V_{DD}\right)\frac{1}{\underline{\alpha}}$$

MOS Transistor Scaling (4)

Analog circuits



$$g_{m,scaled} = \mu(\alpha C_{ox})\frac{W/\alpha}{L/\alpha}\frac{V_{GS}-V_{TH}}{\alpha}$$

$$= \mu C_{ox}\frac{W}{L}(V_{GS}-V_{TH})$$

$$r_{O,scaled} = \frac{1}{\alpha\lambda\frac{I_D}{\alpha}}$$

$$= \frac{1}{\lambda I_D}$$

$$P = IV$$

$$P_{scaled} = (I/\alpha)(V/\alpha)$$

$$= \frac{1}{\underline{\alpha^2}} \cdot IV$$

Looks good to analog design, but …

MOS Transistor Scaling (5)

- The reluctance of circuit designers to use *a lower supply voltage* and the fundamental limitations in decreasing the MOS threshold voltage have led to another scaling scenario: **Constant-voltage scaling**.
- In this case, the device dimensions shrink by *α, the doping levels increase by α, and the voltages remain* constant, thereby increasing the electric fields by ***α***.
- Such *high electric fields* both raise the possibility of device breakdown and exacerbate short-channel effects.

In reality, technology scaling has followed "a mixture of constant-field and constant-voltage trend," thus demanding innovative device design so as to achieve reliability and performance

Outline

- Scaling Theory on CMOS VLSI
- Short Channel Effects
  - Threshold Voltage Variation (V_{TH})
  - Mobility Degeneration (μ)
  - Velocity Saturation (g_m)
  - Hot Carrier Effects (V_{TH})
  - Output Impedance Variation (r_{ds})
- MOS Device Models
- Process Corners

Short-Channel Effects

- **Constant-Scaling is not easily achieved. Why?**
  (CMOS VLSI can tell you more)

- **Small-geometry effects** arise because of the following five factors
  - The electric fields tend to increase because the **supply voltage has not scaled proportionally**
  - The built-in potential term in Eq. (17.5), ϕ_B, is **neither scalable nor negligible**

$$W_d = \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)(\phi_B + V_R)}$$

  - The depth of S/D junctions cannot be **reduced easily**
  - The **mobility decreases** as the substrate doping increases
  - The subthreshold slope (described below) is **not scalable**

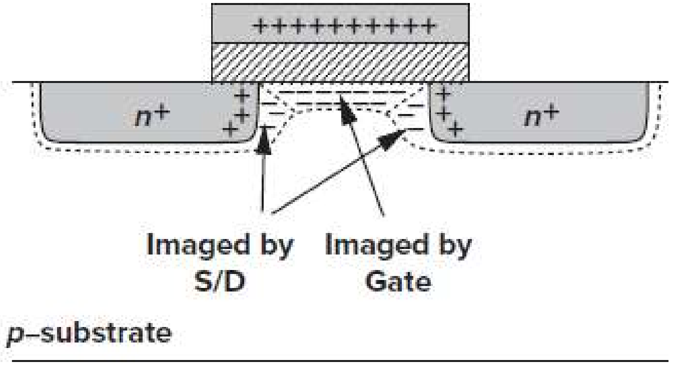
Short-Channel Effects

Threshold Voltage Variation

min V_{TH} is bounded* by

1. Subthreshold behavior
2. Temperature variation
3. Process variation
4. Channel length

* Max $V_{TH}<V_{DD}/4$ is for digital circuit speed



Subthreshold behavior

For long-channel devices, the subthreshold drain current

$$\boxed{I_D = \mu C_d \frac{W}{L} V_T^2 \left(\exp \frac{V_{GS}-V_{TH}}{\zeta V_T}\right)\left(1-\exp\frac{-V_{DS}}{V_T}\right)}$$

where

$$C_d = \sqrt{\epsilon_{si} q N_{sub}/(4\phi_B)}$$
$$V_T = kT/q, \text{ and } \zeta = 1 + C_d/C_{ox}$$

At saturation (V_{DS} > 4*V_T), I_D is independent on V_{DS}

$$\frac{\partial(\log_{10} I_D)}{\partial V_{GS}} = (\log_{10} e)\frac{1}{\zeta V_T}$$

The inverse of this quantity is usually called the "subthreshold slope", ***S***

$(\,)^{-1}$

$$S = 2.3 V_T \left(1 + \frac{C_d}{C_{ox}}\right) \quad \mathrm{V/dec}$$

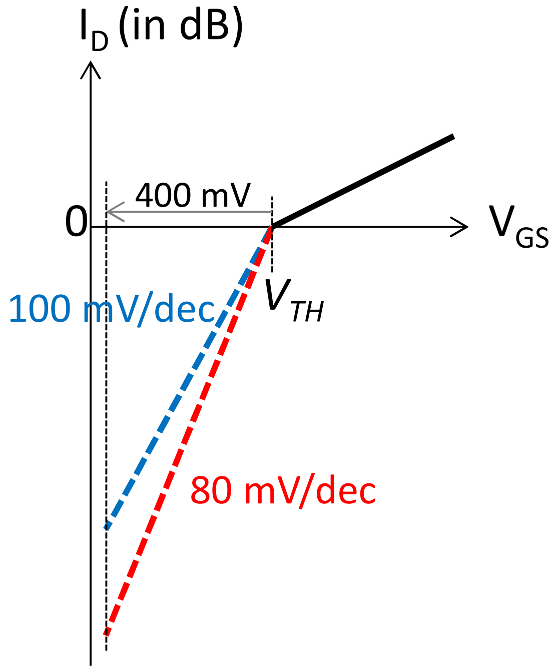
Short-Channel Effects

Threshold Voltage Variation

Subthreshold slope

$$S = 2.3V_T\left(1 + \frac{C_d}{C_{ox}}\right) \quad \text{V/dec}$$

where $C_d = \sqrt{\epsilon_{si} q N_{sub}/(4\phi_B)}$



For example, if $C_d = 0.67C_{ox}$ *, then* ***S*** *= 100 mV/dec,* suggesting that a change of 100 mV in V_{GS} leads to a tenfold reduction in the drain current

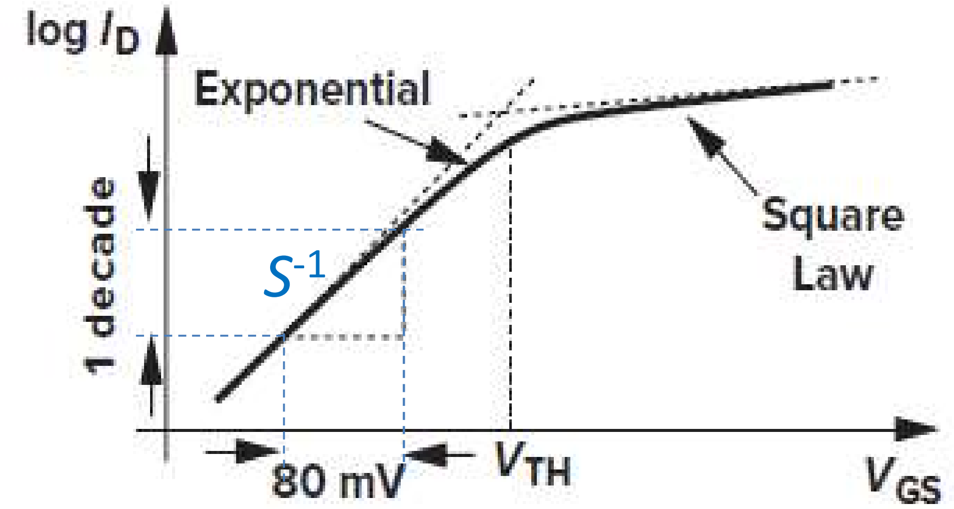
In order to turn off the transistor by lowering V_{GS} below V_{TH} (subthreshold operation), *S* must be as small as possible, i.e., C_d/C_{ox} must be minimized

- *S* = 100 mV/dec => ΔV_{GS} = -100 mV has a current reduction of 1/10
- If *S* is smaller, 80 mV/dec , => ΔV_{GS} = -400 mV can get a current reduction of 10^{-5}. This means the on/off current ratio can have a scaling of 10^5
- That's called the **leakage current**, we care!!

Short-Channel Effects

Threshold Voltage Variation

Leakage Issue



Subthreshold Current

$$\boxed{I_D = \mu C_d \frac{W}{L} V_T^2 \left(\exp \frac{V_{GS}-V_{TH}}{\zeta V_T}\right)\left(1-\exp\frac{-V_{DS}}{V_T}\right)}$$

$$\frac{\partial(\log_{10} I_D)}{\partial V_{GS}} = (\log_{10} e)\frac{1}{\zeta V_T} \quad \Rightarrow \quad S = 2.3V_T\left(1+\frac{C_d}{C_{ox}}\right) \quad \mathrm{V/dec}$$

Comments:

- If V_{TH} is too small => V_{GS}-V_{TH} cannot be more negative => **OFF** current is not low!!
- Temperature variation on V_{TH} causes about 100mV over -40°C ~ +125°C

Short-Channel Effects

Threshold Voltage Variation

Channel Length Issue

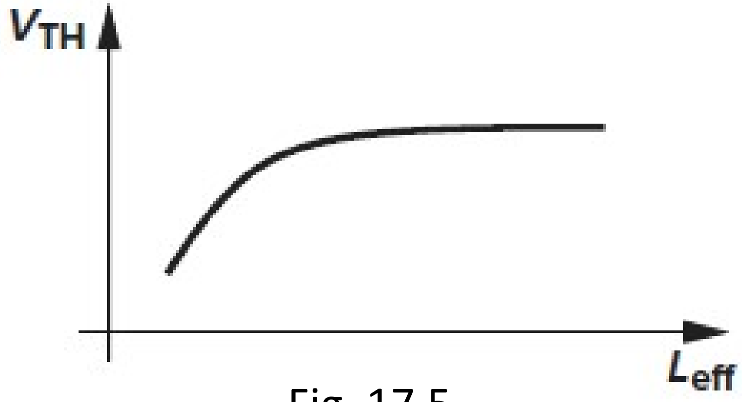


Fig. 17.5

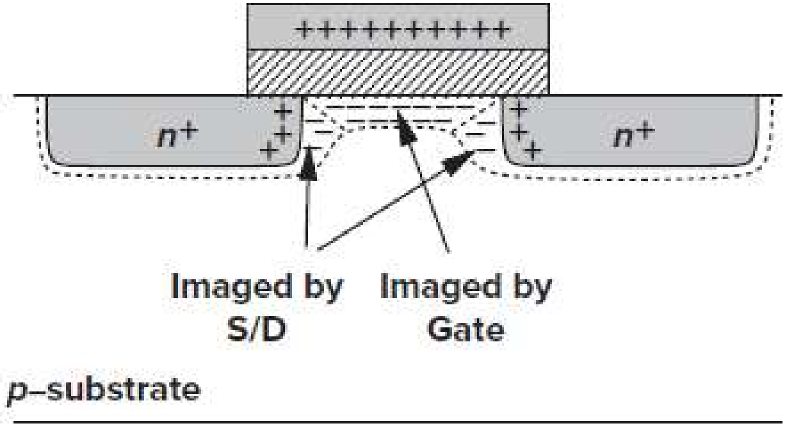


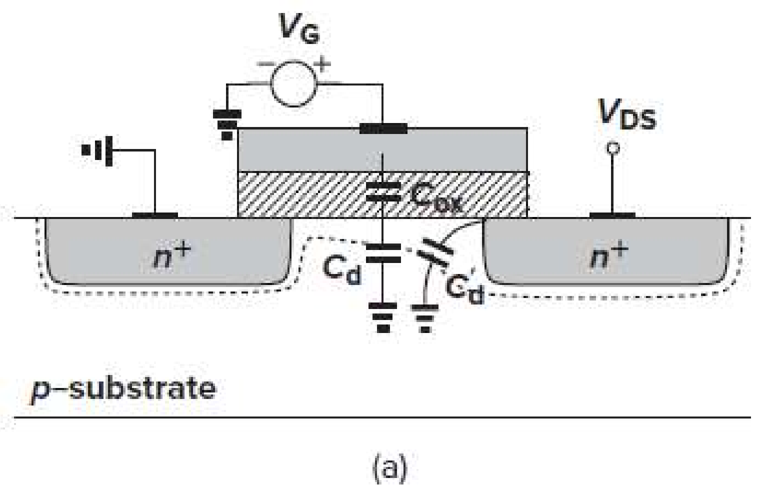
Fig. 17.6

- As shown in Fig. 17.5, transistors fabricated on the same wafer but with different lengths yield lower V_{TH} as L decreases
- This is because the depletion regions associated with the source and drain junctions protrude into the channel area considerably, thereby reducing the immobile charge that must be imaged by the charge on the gate (Fig. 17.6)
- In other words, part of the immobile charge in the substrate is now imaged by the charge inside the source and drain areas rather than by the charge on the gate.
- As a result, the gate voltage required to create an inversion layer **decreases**
- However, things cannot be so simple …

Short-Channel Effects

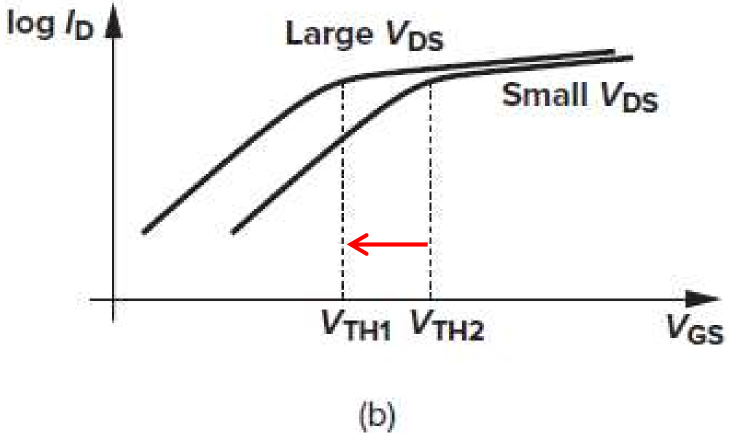
Threshold Voltage Variation

Drain-induced barrier lowering (DIBL)



(a)

V_{DS} effect on V_{TH}

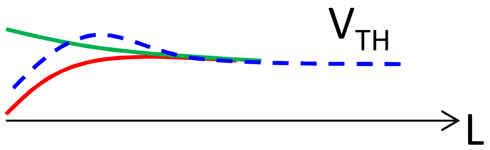


(b)

- In short-channel devices, the drain voltage also makes the surface more positive by creating a two-dimensional field in the depletion region
- The principal impact of ***DIBL*** on circuit design is the degraded output impedance

Short-Channel Effects

Threshold Voltage Variation



Reverse Short-Channel Effect

- In nanometer CMOS technologies, the threshold voltage *decreases as* the channel length increases from its minimum value
- A "***halo***" implant of heavy doping surrounds the source and drain junctions. This implant reduces the penetration of the drain depletion region into the channel area, thereby improving the device characteristics
- Due to the nonuniform substrate doping along the channel in Fig. 17.8, the "local" threshold voltage also varies from the source to the drain
- The channel length increases, the average substrate doping decreases, and so does the threshold voltage (*L* is shorter, avg(N_{sub}) is decreased, V_{TH} is also lower)

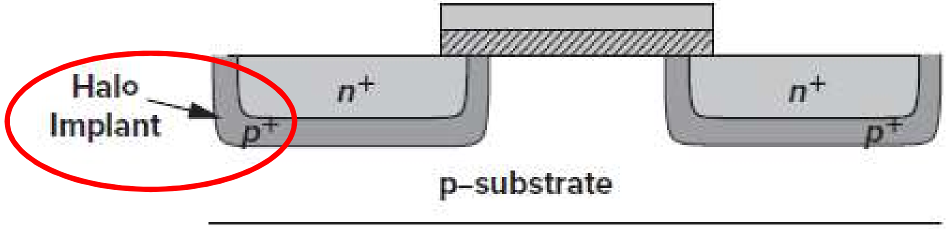


Fig. 17.8

$$V_{TH} = \phi_{MS} + 2\phi_F + \frac{Q_{dep}}{C_{ox}}\uparrow$$

$$\phi_F = (kT/q)\ln(\underline{N_{sub}}/n_i)$$

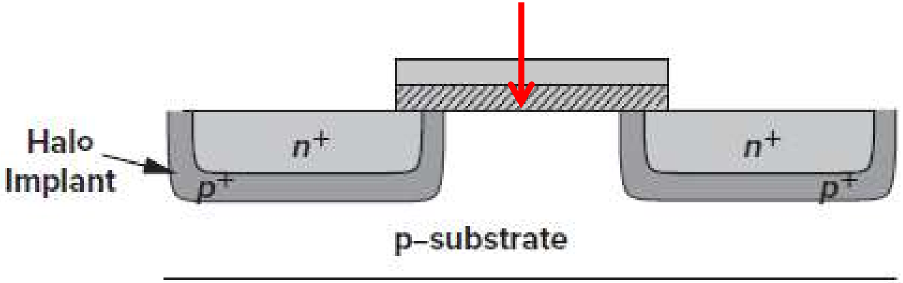
$$Q_{dep} = \sqrt{4q\epsilon_{si}|\phi_F|\underline{N_{sub}}}$$

Short-Channel Effects

Mobility Degeneration

Vertical Field

Large V_{GS} causes crowded carriers



$$\boxed{\mu_{eff} = \frac{\mu_0}{1+\theta(V_{GS}-V_{TH})}}$$

θ is a fitting parameter roughly equal to ($10^{-7}/t_{ox}$) V^{-1} [7]. For example, if t_{ox} = 100 Å, then $\theta \approx 1\ V^{-1}$

↓

The mobility begins to fall as the overdrive exceeds 100 mV

$$I_D = \frac{1}{2}\frac{\mu_0 C_{ox}}{1+\theta(V_{GS}-V_{TH})}\frac{W}{L}(V_{GS}-V_{TH})^2$$

↓ Assume $\theta(V_{GS}-V_{TH}) \ll 1$

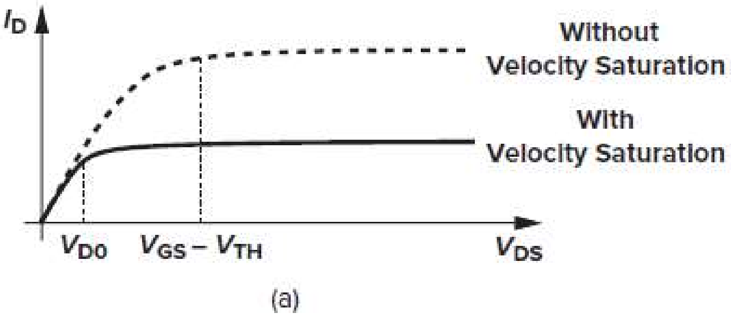
$$I_D \approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[1-\underline{\theta(V_{GS}-V_{TH})}](V_{GS}-V_{TH})^2$$

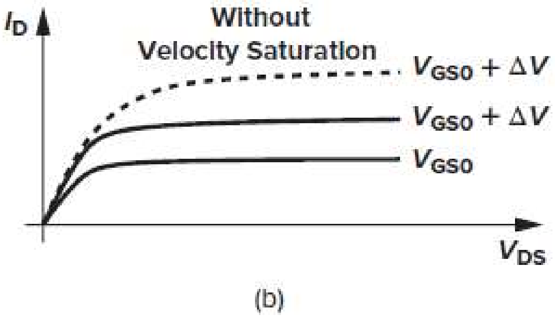
$$\approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[(V_{GS}-V_{TH})^2-\underline{\theta(V_{GS}-V_{TH})^3}]$$

Short-Channel Effects

Velocity Saturation

Lateral Field

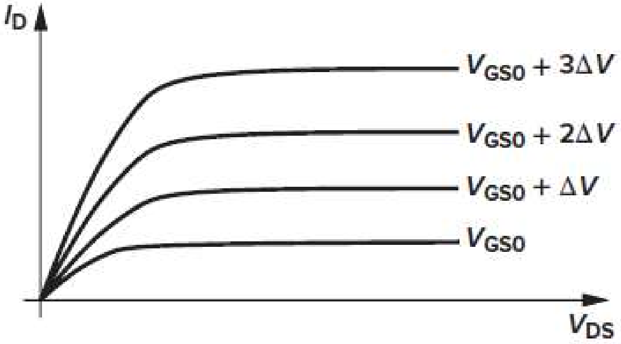




$$v = \mu E$$

v approaches a saturated value, ~10^7 cm/s

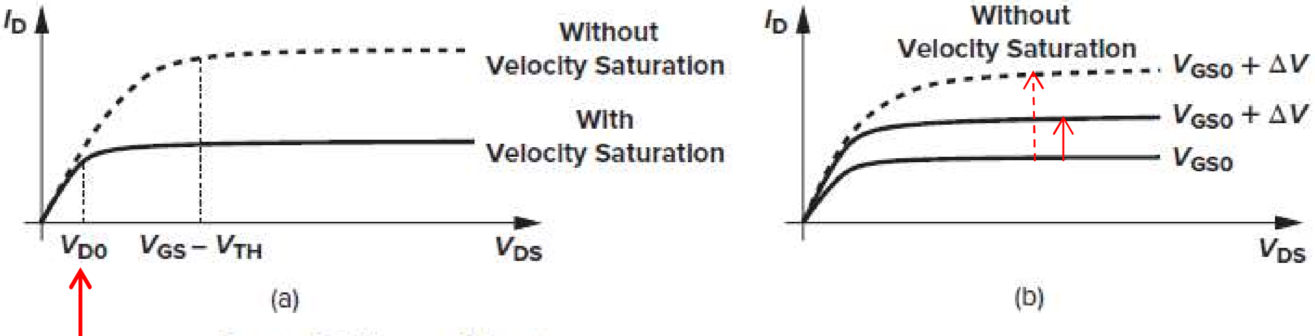
$$I_D = v_{sat} Q_d$$
$$= v_{sat} W C_{ox} (V_{GS} - V_{TH})$$
$$g_m = v_{sat} W C_{ox}$$



Short-Channel Effects

Velocity Saturation

Lateral Field



$$V_{DS,sat} = \frac{2\mu_{eff} L (V_{GS} - V_{TH})}{2\mu_{eff} L + V_{GS} - V_{TH}}$$

Observed from Lab data

$$\boxed{I_D = W C_{ox} v_{sat} \frac{(V_{GS} - V_{TH})^2}{V_{GS} - V_{TH} + 2\frac{v_{sat} L}{\underline{\mu_{eff}}}}}$$

where $\mu_{eff} = \dfrac{\mu_0}{1 + \theta (V_{GS} - V_{TH})}$

If L or v_{sat} is large enough,
$I_D = 1/2 \cdot \mu_0 C_{ox} (W/L) \cdot (V_{GS} – V_{TH})^2$

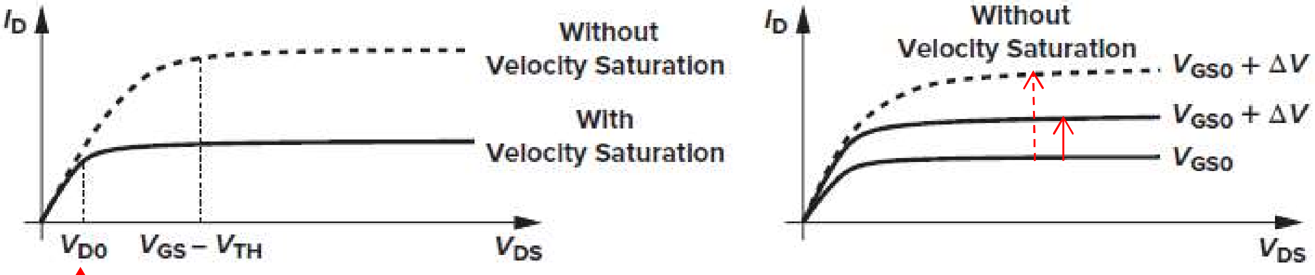
If L or v_{sat} is very small,
$I_D = W C_{ox} v_{sat} (V_{GS} - V_{TH})$

For example, if $v_{sat} \approx 10^7$ cm/s, L = 0.25 μm, and $\mu_0 \approx$ 350 cm^2/V/s, we have $2 v_{sat} L / \mu_0 \approx 1.43$ V, recognizing that for overdrive voltages of a few hundred millivolts, the transistor operation is somewhat close to the square law

Short-Channel Effects

Velocity Saturation

Lateral Field



(a)

(b)

$$\uparrow$$

$$V_{DS,sat} = \frac{2\mu_{eff} L (V_{GS} - V_{TH})}{2\mu_{eff} L + V_{GS} - V_{TH}}$$

$$I_D = W C_{ox} v_{sat} \frac{(V_{GS} - V_{TH})^2}{V_{GS} - V_{TH} + \frac{2 v_{sat} L}{\mu_0}[1 + \theta(V_{GS} - V_{TH})]}$$

Observed from Lab data

$$\boxed{I_D = W C_{ox} v_{sat} \frac{(V_{GS} - V_{TH})^2}{V_{GS} - V_{TH} + 2\frac{v_{sat} L}{\underline{\mu_{eff}}}}}$$

where $\mu_{eff} = \dfrac{\mu_0}{1 + \theta(V_{GS} - V_{TH})}$

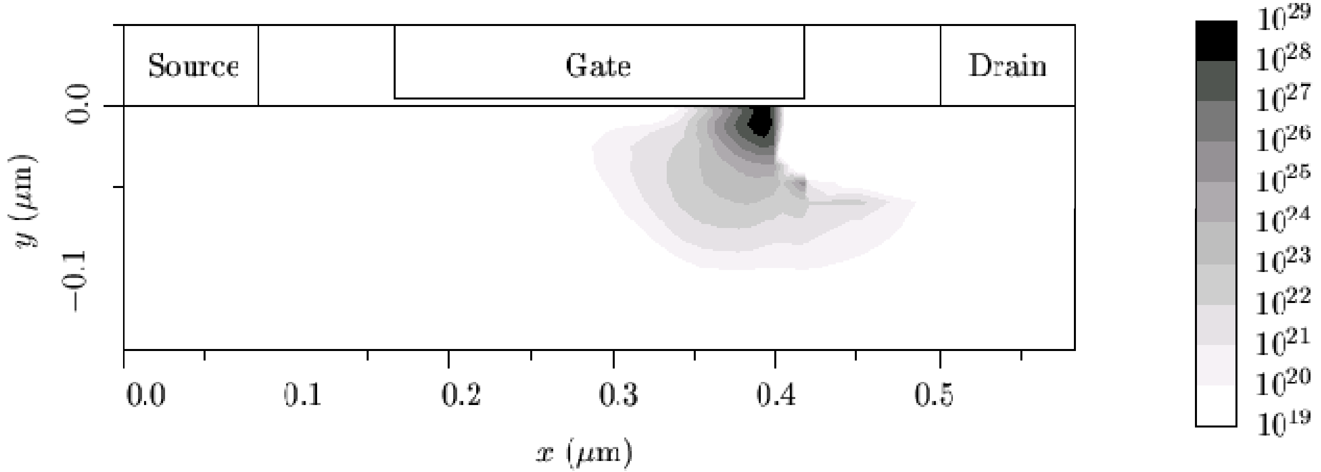
$$= W C_{ox} v_{sat} \frac{(V_{GS} - V_{TH})^2}{\frac{2 v_{sat} L}{\mu_0} + \left(1 + \frac{2 v_{sat} L \theta}{\mu_0}\right)(V_{GS} - V_{TH})}$$

$$= \frac{1}{2}\mu_0 C_{ox} \frac{W}{L} \frac{(V_{GS} - V_{TH})^2}{1 + \left(\frac{\mu_0}{2 v_{sat} L} + \theta\right)(V_{GS} - V_{TH})}$$

For example, the drain current contains “high-order nonlinear terms”

Short-Channel Effects

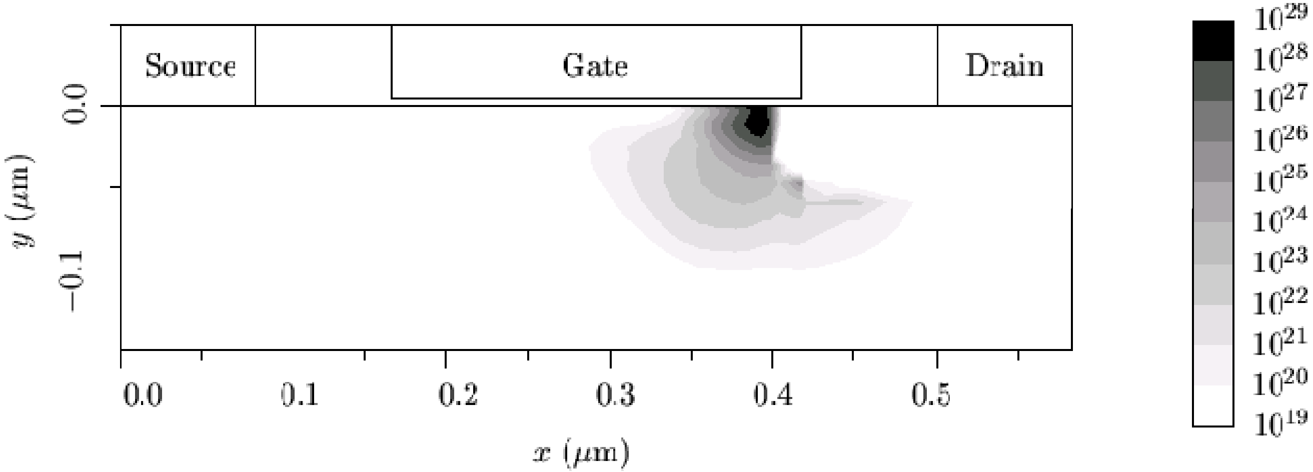
Hot Carrier Effects



- In the vicinity of the drain region, hot carriers may "hit" the silicon atoms at high speeds, thereby creating impact ionization. As a result, new electrons and holes are generated, with the electrons absorbed by the drain and the holes by the substrate.
- Thus, a finite drain-substrate current appears. Also, if the carriers acquire a very high energy, they may be injected into the gate oxide and even flow out the gate terminal, introducing a gate current.
- The substrate and gate currents are often measured to study hot carrier effects.
- In nanometer technologies, hot carrier effects have subsided. This is because the energy required to create an electron-hole pair, $Eg \approx 1.12$ eV, is simply not available if the supply voltage is around 1 V.

Short-Channel Effects

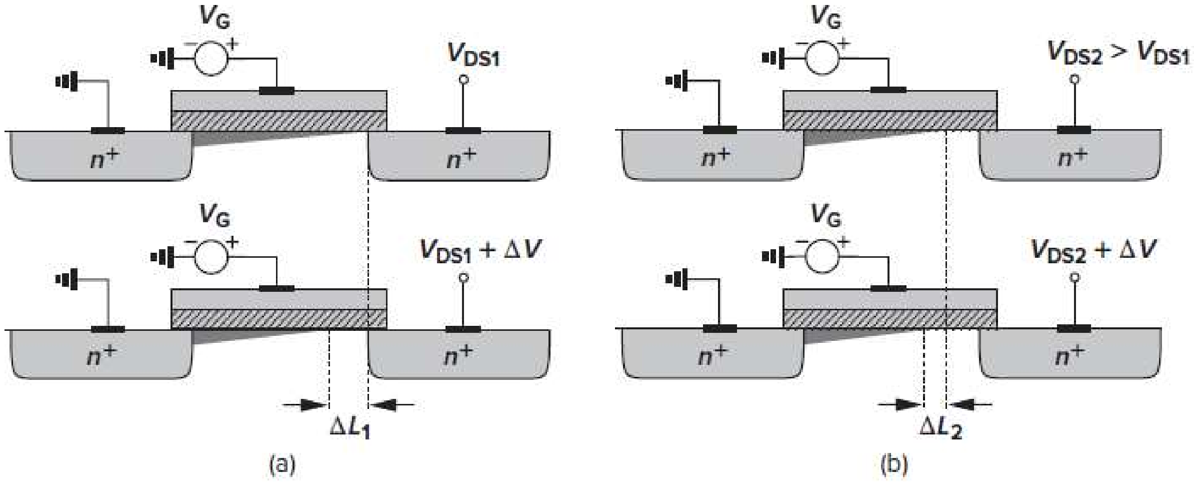
Hot Carrier Effects



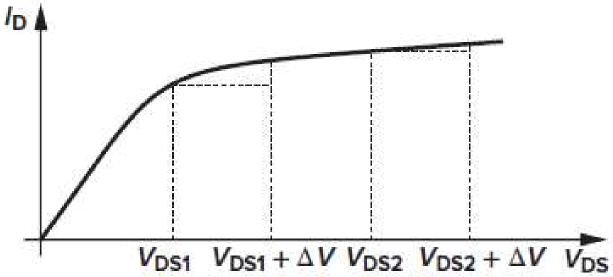
- If a MOS transistor is operated under pinch-off condition, also known as ``saturated case'', hot carriers traveling with saturation velocity can cause parasitic effects at the drain side of the channel known as ``**Hot Carrier Effects**'' (HCE).
  - **Hot Carrier Effect** causes threshold voltage reduction or latch-up
  - **Lightly Doped Drain** (LDD) or **Graded Channel** (GC) help to avoid HCE
- Carrier injection into the gate oxide can lead to hot carrier degradation effects such as threshold voltage changes due to occupied traps in the oxide. Hot carriers can also generate traps at the silicon-oxide interface known as ``fast surface states'' leading to subthreshold swing deterioration and stress-induced drain leakage

Short-Channel Effects

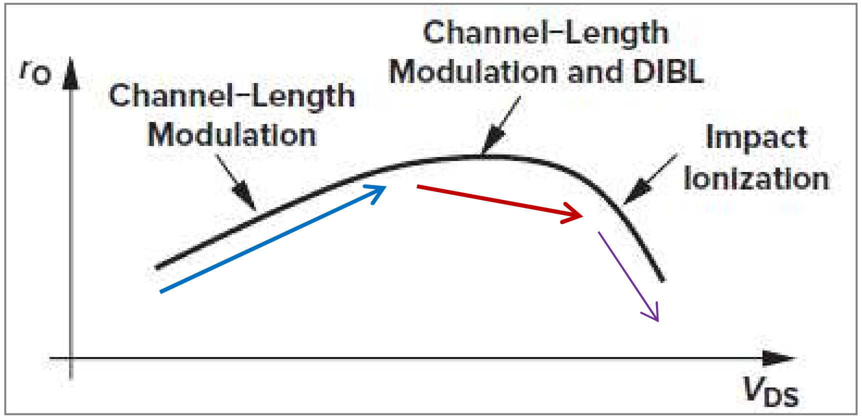
Output Impedance Variation



In reality, however, *r_O varies with V_{DS}. As V_{DS}* increases and the pinch-off point moves toward the source, the rate at which the depletion region around the source becomes wider decreases, resulting in a higher incremental output impedance.



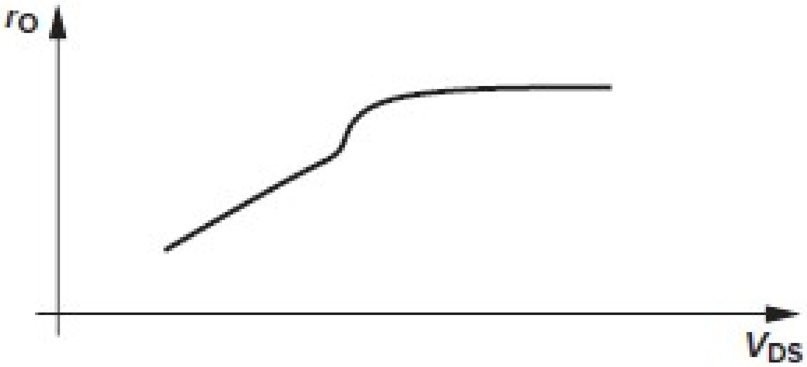
$$r_O = \frac{2L}{1 - \frac{\Delta L}{L}} \frac{1}{I_D} \sqrt{\frac{qN_B}{2\epsilon_{si}}(V_{DS} - V_{DS,sat})}$$

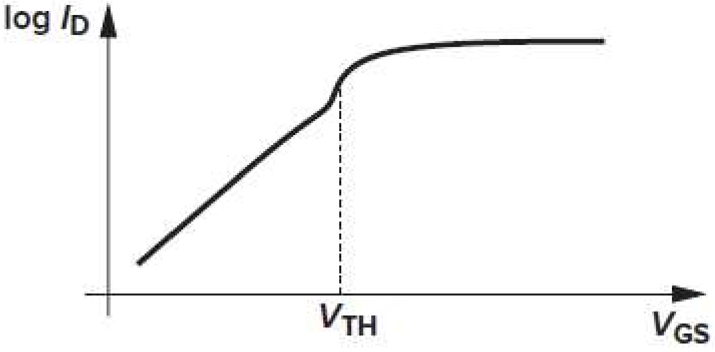


Impact ionization limits the maximum gain that can be obtained from cascode structures because it introduces a small-signal resistance from the drain to the substrate rather than to the source.

Modeling

- Level 1 (You have learned in Microelectronics)
- Level 2 (high-order effects considered)
- Level 3





- BSIM 1, 2, 3, 4
- HSPICE Level 28
- MOS9
- EKV (Subthreshold)

Modeling (Detailed in the textbook)

Level 1 Model

$$I_D = \frac{1}{2}K_P\frac{W}{L-2L_D}[2(V_{GS}-V_{TH})V_{DS}-V_{DS}^2](1+\lambda V_{DS}) \quad \text{Triode Region}$$

$$I_D = \frac{1}{2}K_P\frac{W}{L-2L_D}(V_{GS}-V_{TH})^2(1+\lambda V_{DS}) \quad \text{Saturation Region}$$

$$C_{GS} = \frac{2}{3}WLC_{ox}\left\{1-\frac{(V_{GS}-V_{DS}-V_{TH})^2}{[2(V_{GS}-V_{TH})-V_{DS}]^2}\right\}+WC_{ov}$$

$$C_{GD} = \frac{2}{3}WLC_{ox}\left\{1-\frac{(V_{GS}-V_{TH})^2}{[2(V_{GS}-V_{TH})-V_{DS}]^2}\right\}+WC_{ov}$$

$$C_{GB} = 0.$$

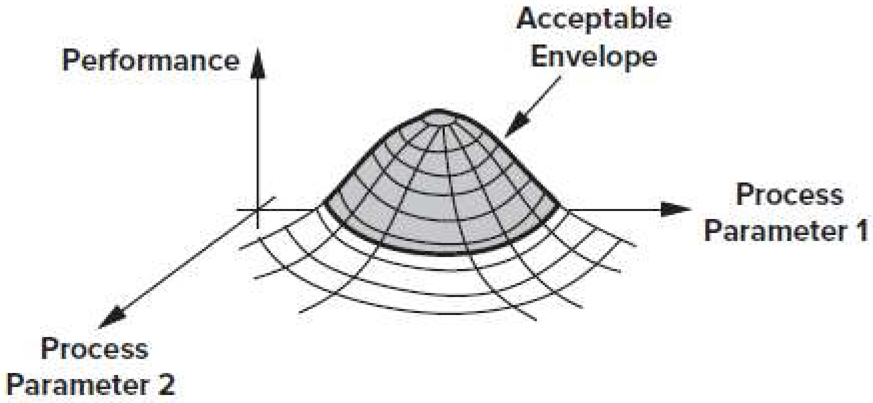
Level 2 Model

$$\begin{aligned} I_D = \mu C_{ox}\frac{W}{L}\{(V_{GS}-V_{TH0})V_{DS}-\frac{V_{DS}^2}{2} \\ -\frac{2}{3}\gamma[(V_{DS}-V_{BS}+2\phi_F)^{3/2}-(-V_{BS}+2\phi_F)^{3/2}]\}\end{aligned}$$

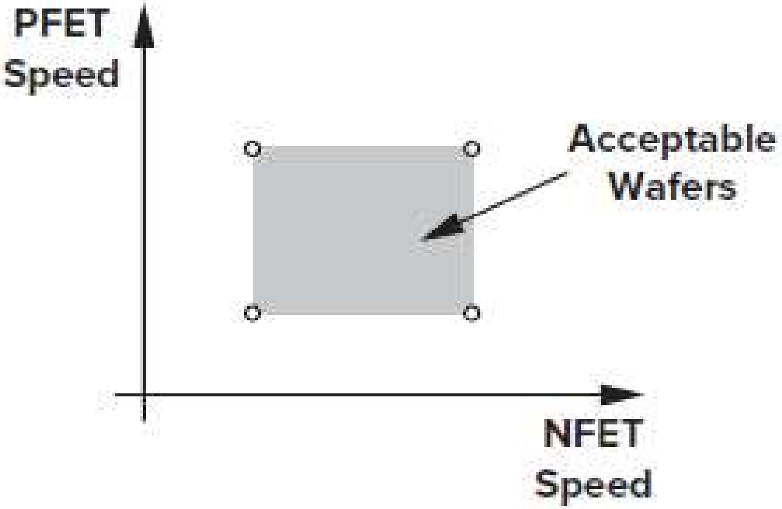
$$V_{D,sat} = V_{GS}-V_{TH0}-\phi_F+\gamma^2\left[1-\sqrt{1+\frac{2}{\gamma^2}(V_{GS}-V_{TH0}+\phi_F)}\right]$$

$$I_{DS} = I_{D,sat}\frac{1}{1-\lambda V_{DS}}$$

Process Corners



In order to facilitate the task of circuit design to some extent, process engineers guarantee a performance envelope for the devices, in essence tightening the anticipated parameter variations by discarding wafers that fall out of the envelope (Fig. 17.17).



Illustrated in Fig. 17.18, the idea is to constrain the speed envelope of the NMOS and PMOS transistors to a rectangle defined by four corners:

- fast NFET and fast PFET
- slow NFET and slow PFET
- fast NFET and slow PFET
- slow NFET and fast PFET