

## Transistors and Scaling

Seth Copen Goldstein  
seth@cs.cmu.edu  
CMU

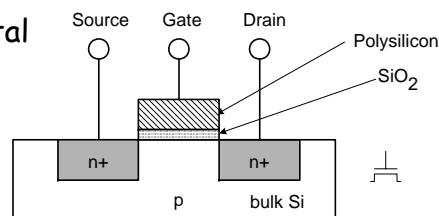
Adapted from "Intro to CMOS VLSI Design, Harris

## Outline

- Transistor theory
- Transistor reality
- Scaling

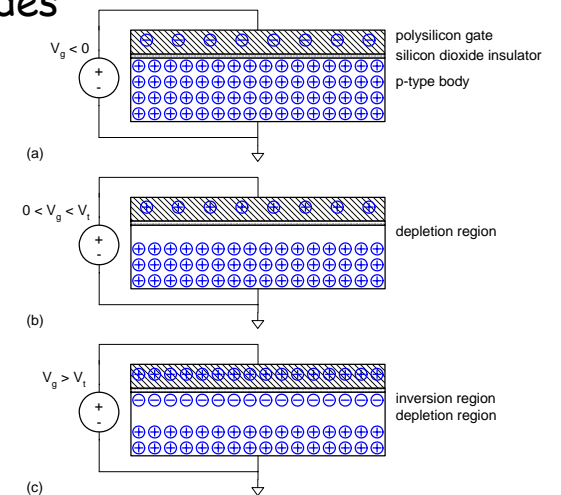
## nMOS Transistor

- Four terminals: gate, source, drain, body
- Gate - oxide - body stack looks like a capacitor
  - Gate and body are conductors
  - $\text{SiO}_2$  (oxide) is a very good insulator
  - Called metal - oxide - semiconductor (MOS) capacitor
  - Even though gate is no longer made of metal



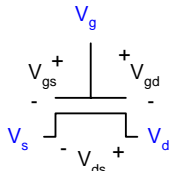
## MOS Capacitor

- Gate and body form MOS capacitor
- Operating modes
  - Accumulation
  - Depletion
  - Inversion



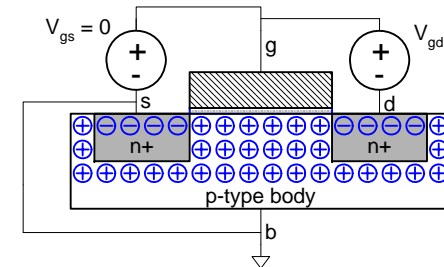
# Terminal Voltages

- Mode of operation depends on  $V_g, V_d, V_s$ 
  - $V_{gs} = V_g - V_s$
  - $V_{gd} = V_g - V_d$
  - $V_{ds} = V_d - V_s = V_{gs} - V_{gd}$
- Source and drain are symmetric diffusion terminals
  - By convention, source is terminal at lower voltage
  - Hence  $V_{ds} \geq 0$
- nMOS body is grounded. First assume source is 0 too.
- Three regions of operation
  - Cutoff
  - Linear
  - Saturation



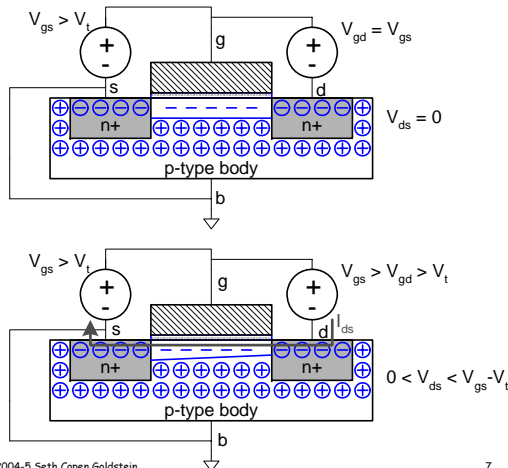
# nMOS Cutoff

- No channel
- $I_{ds} = 0$



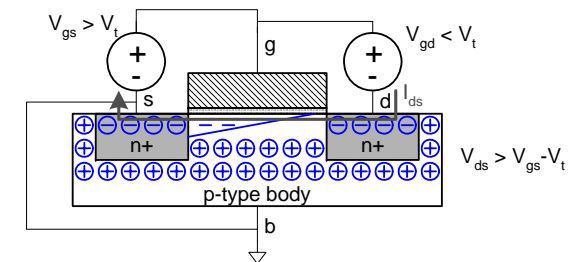
# nMOS Linear

- Channel forms
- Current flows from d to s ( $e^-$  from s to d)
- $I_{ds} \uparrow$  as  $V_{ds} \uparrow$
- Similar to linear resistor



# nMOS Saturation

- Channel pinches off
- $I_{ds}$  independent of  $V_{ds}$
- We say current saturates

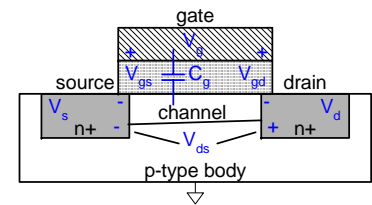
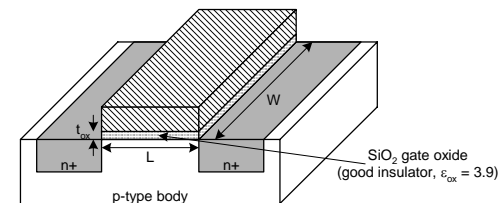


# I-V Characteristics

- In Linear region,  $I_{ds}$  depends on
  - How much charge is in the channel?
  - How fast is the charge moving?

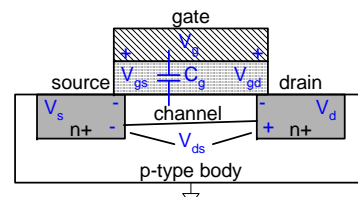
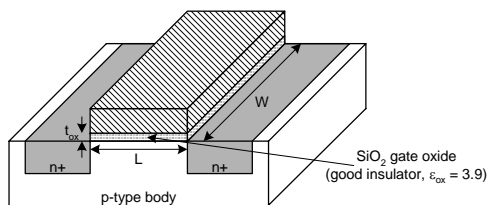
# Channel Charge

- MOS structure looks like parallel plate capacitor while operating in inversion
  - Gate - oxide - channel
- $Q_{channel} =$



# Channel Charge

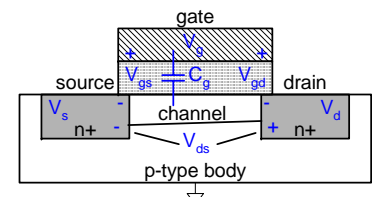
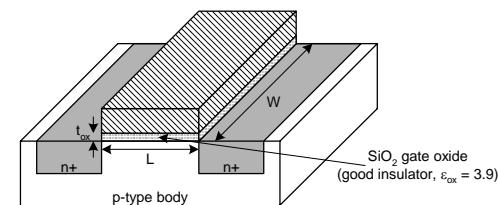
- MOS structure looks like parallel plate capacitor while operating in inversion
  - Gate - oxide - channel
- $Q_{channel} = CV$
- $C =$



# Channel Charge

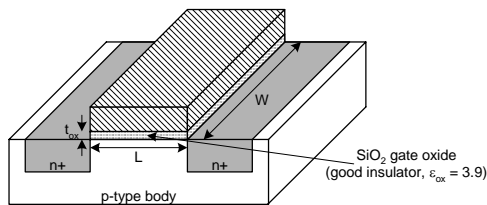
- MOS structure looks like parallel plate capacitor while operating in inversion
  - Gate - oxide - channel
- $Q_{channel} = CV$
- $C = C_g = \epsilon_{ox} WL / t_{ox} = C_{ox} WL$
- $V =$

$$C_{ox} = \epsilon_{ox} / t_{ox}$$



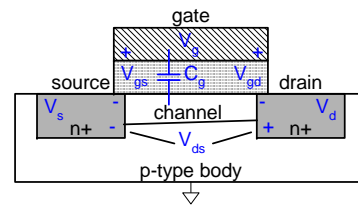
## Channel Charge

- MOS structure looks like parallel plate capacitor while operating in inversion
  - Gate - oxide - channel
- $Q_{\text{channel}} = CV$
- $C = C_g = \epsilon_{\text{ox}} WL / t_{\text{ox}} = C_{\text{ox}} WL$
- $V = V_{g_c} - V_{\text{t}} = (V_{g_s} - V_{d_s}/2) - V_{\text{t}}$   $C_{\text{ox}} = \epsilon_{\text{ox}} / t_{\text{ox}}$



lecture6 15-398

© 2004-5 Seth Copen Goldstein



13

## Carrier velocity

- Charge is carried by e-
- Carrier velocity  $v$  proportional to lateral E-field between source and drain
- $v =$

lecture6 15-398

© 2004-5 Seth Copen Goldstein

14

## Carrier velocity

- Charge is carried by e-
- Carrier velocity  $v$  proportional to lateral E-field between source and drain
- $v = \mu E$   $\mu$  called mobility
- $E =$

lecture6 15-398

© 2004-5 Seth Copen Goldstein

15

## Carrier velocity

- Charge is carried by e-
- Carrier velocity  $v$  proportional to lateral E-field between source and drain
- $v = \mu E$   $\mu$  called mobility
- $E = V_{ds}/L$
- Time for carrier to cross channel:
  - $t =$

lecture6 15-398

© 2004-5 Seth Copen Goldstein

16

## Carrier velocity

- Charge is carried by e-
- Carrier velocity  $v$  proportional to lateral E-field between source and drain
- $v = \mu E$        $\mu$  called mobility
- $E = V_{ds}/L$
- Time for carrier to cross channel:
  - $t = L / v$

## nMOS Linear I-V

- Now we know
  - How much charge  $Q_{\text{channel}}$  is in the channel
  - How much time  $t$  each carrier takes to cross

$$I_{ds} =$$

## nMOS Linear I-V

- Now we know
  - How much charge  $Q_{\text{channel}}$  is in the channel
  - How much time  $t$  each carrier takes to cross

$$I_{ds} = \frac{Q_{\text{channel}}}{t}$$

$$=$$

## nMOS Linear I-V

- Now we know
  - How much charge  $Q_{\text{channel}}$  is in the channel
  - How much time  $t$  each carrier takes to cross

$$I_{ds} = \frac{Q_{\text{channel}}}{t}$$

$$= \mu C_{\text{ox}} \frac{W}{L} \left( V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds}$$

$$= \beta \left( V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds}$$

$\beta$ : gain factor  
Depends on:

- Process
- geometry

$$\beta = \mu C_{\text{ox}} \frac{W}{L}$$

## nMOS Saturation I-V

- If  $V_{gd} < V_t$ , channel pinches off near drain
  - When  $V_{ds} > V_{dsat} = V_{gs} - V_t$
- Now drain voltage no longer increases current

$$I_{ds} =$$

## nMOS Saturation I-V

- If  $V_{gd} < V_t$ , channel pinches off near drain
  - When  $V_{ds} > V_{dsat} = V_{gs} - V_t$
- Now drain voltage no longer increases current

$$I_{ds} = \beta \left( V_{gs} - V_t - \frac{V_{dsat}}{2} \right) V_{dsat}$$

## nMOS Saturation I-V

- If  $V_{gd} < V_t$ , channel pinches off near drain
  - When  $V_{ds} > V_{dsat} = V_{gs} - V_t$
- Now drain voltage no longer increases current

$$I_{ds} = \beta \left( V_{gs} - V_t - \frac{V_{dsat}}{2} \right) V_{dsat}$$

$$= \frac{\beta}{2} (V_{gs} - V_t)^2$$

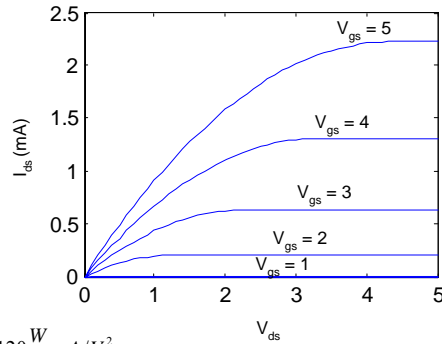
## nMOS I-V Summary

- *Shockley* 1<sup>st</sup> order transistor models

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t & \text{cutoff} \\ \beta \left( V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds} & V_{ds} < V_{dsat} & \text{linear} \\ \frac{\beta}{2} (V_{gs} - V_t)^2 & V_{ds} > V_{dsat} & \text{saturation} \end{cases}$$

## Example

- Consider a 0.6  $\mu\text{m}$  process
  - From AMI Semiconductor
  - $t_{\text{ox}} = 100 \text{ \AA}$
  - $\mu = 350 \text{ cm}^2/\text{V}^*\text{s}$
  - $V_t = 0.7 \text{ V}$
- Plot  $I_{\text{ds}}$  vs.  $V_{\text{ds}}$ 
  - $V_{\text{gs}} = 0, 1, 2, 3, 4, 5$
  - Use  $W/L = 4/2 \lambda$



$$\beta = \mu C_{\text{ox}} \frac{W}{L} = (350) \left( \frac{3.9 \cdot 8.85 \cdot 10^{-14}}{100 \cdot 10^{-8}} \right) \left( \frac{W}{L} \right) = 120 \frac{W}{L} \mu\text{A}/\text{V}^2$$

## pMOS I-V

- All dopings and voltages are inverted for pMOS
- Mobility  $\mu_p$  is determined by holes
  - Typically 2-3x lower than that of electrons  $\mu_n$
  - 120  $\text{cm}^2/\text{V}^*\text{s}$  in AMI 0.6  $\mu\text{m}$  process
- Thus pMOS must be wider to provide same current

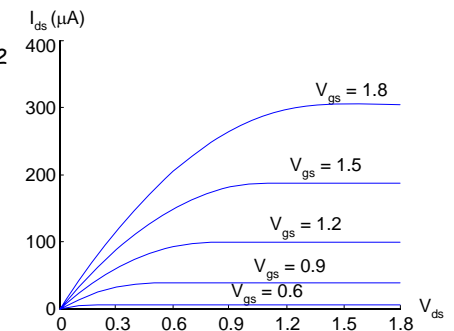
## Ideal Transistor I-V

- Shockley 1<sup>st</sup> order transistor models

$$I_{\text{ds}} = \begin{cases} 0 & V_{\text{gs}} < V_t & \text{cutoff} \\ \beta \left( V_{\text{gs}} - V_t - \frac{V_{\text{ds}}}{2} \right) V_{\text{ds}} & V_{\text{ds}} < V_{\text{dsat}} & \text{linear} \\ \frac{\beta}{2} (V_{\text{gs}} - V_t)^2 & V_{\text{ds}} > V_{\text{dsat}} & \text{saturation} \end{cases}$$

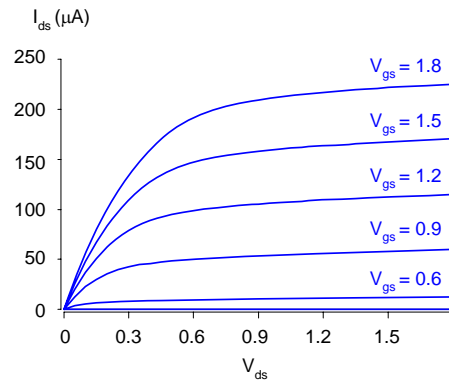
## Ideal nMOS I-V Plot

- 180 nm TSMC process
- Ideal Models
  - $\beta = 155(W/L) \mu\text{A}/\text{V}^2$
  - $V_t = 0.4 \text{ V}$
  - $V_{\text{DD}} = 1.8 \text{ V}$



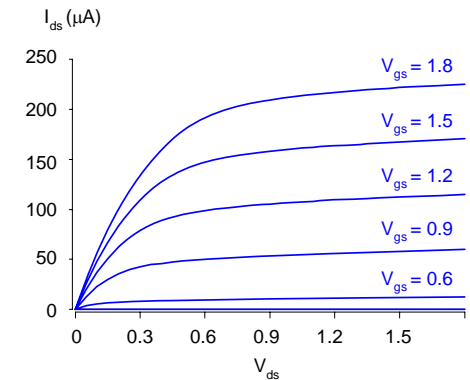
## Simulated nMOS I-V Plot

- 180 nm TSMC process
- BSIM 3v3 SPICE models
- What differs?



## Simulated nMOS I-V Plot

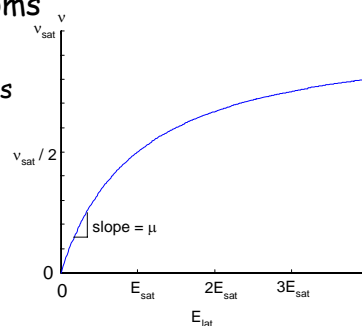
- 180 nm TSMC process
- BSIM 3v3 SPICE models
- What differs?
  - Less ON current
  - No square law
  - Current increases in saturation



## Velocity Saturation

- We assumed carrier velocity is proportional to E-field
  - $v = \mu E_{lat} = \mu V_{ds}/L$
- At high fields, this ceases to be true
  - Carriers scatter off atoms
  - Velocity reaches  $v_{sat}$ 
    - Electrons:  $6-10 \times 10^6$  cm/s
    - Holes:  $4-8 \times 10^6$  cm/s
  - Better model

$$v = \frac{\mu E_{lat}}{1 + \frac{E_{lat}}{E_{sat}}} \Rightarrow v_{sat} = \mu E_{sat}$$



## Vel Sat I-V Effects

- Ideal transistor ON current increases with  $V_{DD}^2$

$$I_{ds} = \mu C_{ox} \frac{W}{L} \frac{(V_{gs} - V_t)^2}{2} = \frac{\beta}{2} (V_{gs} - V_t)^2$$

- Velocity-saturated ON current increases with  $V_{DD}$

$$I_{ds} = C_{ox} W (V_{gs} - V_t) v_{max}$$

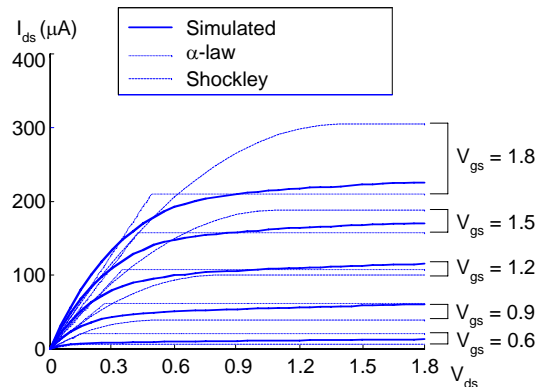
- Real transistors are partially velocity saturated
  - Approximate with  $\alpha$ -power law model
  - $I_{ds} \propto V_{DD}^\alpha$
  - $1 < \alpha < 2$  determined empirically

## $\alpha$ -Power Model

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t & \text{cutoff} \\ I_{dsat} \frac{V_{ds}}{V_{dsat}} & V_{ds} < V_{dsat} & \text{linear} \\ I_{dsat} & V_{ds} > V_{dsat} & \text{saturation} \end{cases}$$

$$I_{dsat} = P_c \frac{\beta}{2} (V_{gs} - V_t)^\alpha$$

$$V_{dsat} = P_v (V_{gs} - V_t)^{\alpha/2}$$



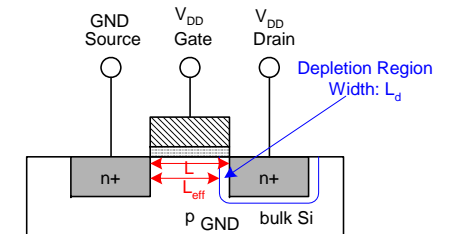
lecture6 15-398

© 2004-5 Seth Copen Goldstein

33

## Channel Length Modulation

- Reverse-biased p-n junctions form a *depletion region*
  - Region between n and p with no carriers
  - Width of depletion  $L_d$  region grows with reverse bias
  - $L_{eff} = L - L_d$
- Shorter  $L_{eff}$  gives more current
  - $I_{ds}$  increases with  $V_{ds}$
  - Even in saturation



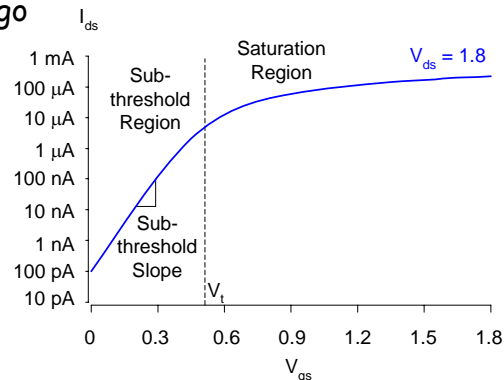
lecture6 15-398

© 2004-5 Seth Copen Goldstein

34

## OFF Transistor Behavior

- What about current in cutoff?
- Simulated results
- What differs?
  - Current doesn't go to 0 in cutoff



lecture6 15-398

© 2004-5 Seth Copen Goldstein

35

## Leakage Sources

- Subthreshold conduction
  - Transistors can't abruptly turn ON or OFF
- Junction leakage
  - Reverse-biased PN junction diode current
- Gate leakage
  - Tunneling through ultrathin gate dielectric
- Subthreshold leakage is the biggest source in modern transistors

lecture6 15-398

© 2004-5 Seth Copen Goldstein

36

## Subthreshold Leakage

- Subthreshold leakage exponential with  $V_{gs}$

$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_t}{nV_T}} \left( 1 - e^{-\frac{V_{ds}}{V_T}} \right) \quad I_{ds0} = \beta V_T^2 e^{1.8}$$

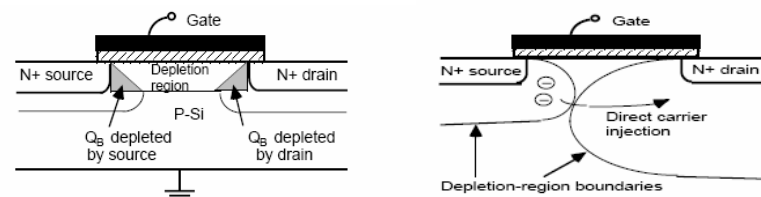
- $n$  is process dependent, typically 1.4-1.5

## DIBL

- Drain-Induced Barrier Lowering
  - Drain voltage also affect  $V_t$

$$V_t' = V_t - \eta V_{ds}$$

- High drain voltage causes subthreshold leakage to \_\_\_\_\_.

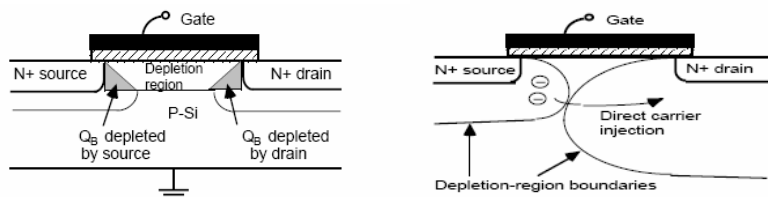


## DIBL

- Drain-Induced Barrier Lowering
  - Drain voltage also affect  $V_t$

$$V_t' = V_t - \eta V_{ds}$$

- High drain voltage causes subthreshold leakage to **increase**.

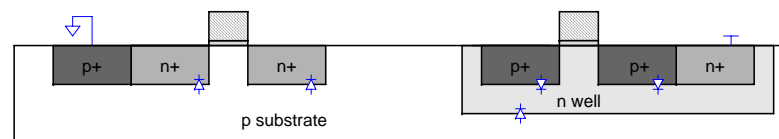


## Junction Leakage

- Reverse-biased p-n junctions have some leakage

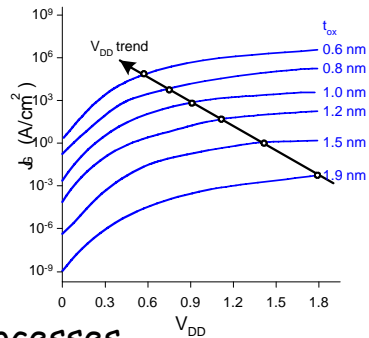
$$I_D = I_S \left( e^{\frac{V_D}{V_T}} - 1 \right)$$

- $I_S$  depends on doping levels
  - And area and perimeter of diffusion regions
  - Typically  $< 1 \text{ fA}/\mu\text{m}^2$



## Gate Leakage

- Carriers may tunnel through very thin gate oxides
- Predicted tunneling current (from [Song01])



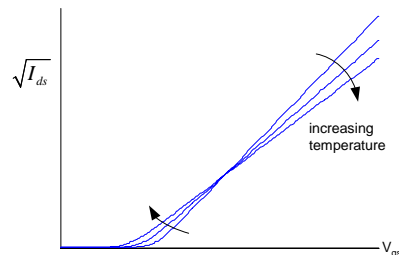
- Negligible for older processes
- May soon be critically important

## Temperature Sensitivity

- Increasing temperature
  - Reduces mobility
  - Reduces  $V_t$
- $I_{ON}$  \_\_\_\_\_ with temperature
- $I_{OFF}$  \_\_\_\_\_ with temperature

## Temperature Sensitivity

- Increasing temperature
  - Reduces mobility
  - Reduces  $V_t$
- $I_{ON}$  **decreases** with temperature
- $I_{OFF}$  **increases** with temperature

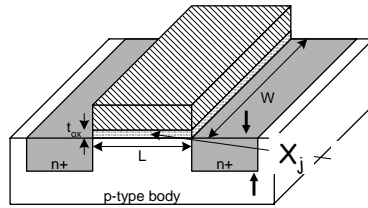


## So What?

- So what if transistors are not ideal?
  - They still behave like switches.
- But these effects matter for...
  - Supply voltage choice
  - Logical effort
  - Quiescent power consumption
  - Pass transistors
  - Temperature of operation

# Scaling Methods

- Dennard: scaling method that maintains constant electric field



<u>Device/Circuit Parameter</u>		<u>Constant Field Scaling Factor</u>	
Dimension :	$x_{ox}, L, W, X_j,$	$1/K$	
Substrate doping :	$N_a$	$K$	
Supply voltage :	$V$	$1/K$	
Supply current :	$I$	$1/K$	
Gate Capacitance :	$W L/x_{ox}$	$1/K$	
Gate delay :	$C V / I$	$1/K$	
Power dissipation :	$C V^2 / \text{delay}$	$1/K^2$	saraswat