Short Channel MOS Transistor

Kaushik Roy

Purdue University
Dept. of ECE

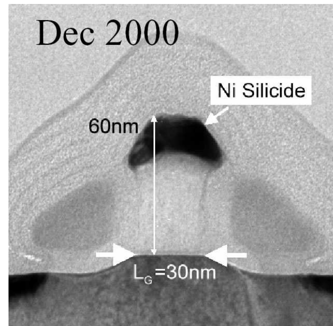
kaushik@purdue.edu

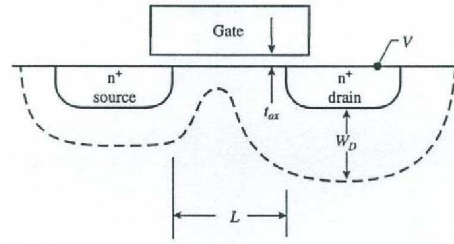
Topics

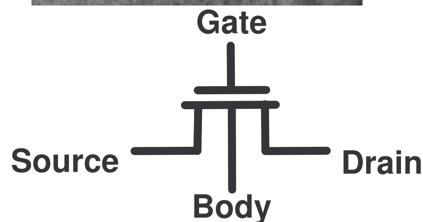
- Long channel MOSFET: review
  - Strong inversion (linear, saturation mode)
- Short channel MOSFET
  - Velocity saturation
  - Vt roll-off
  - Drain induced barrier lowering
  - Series resistance
  - Narrow width effect
  - Weak inversion (V_t, S-swing)

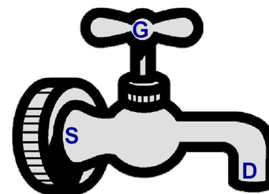
2

Transistor







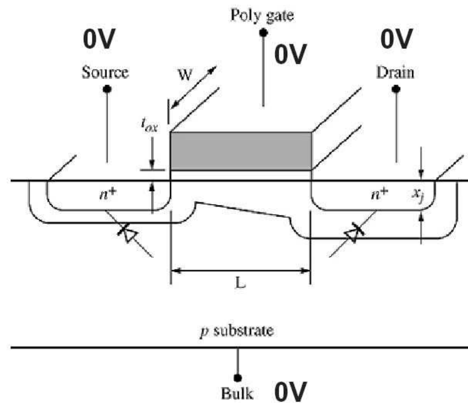


Kuroda, IEDM panel

Compact Modeling

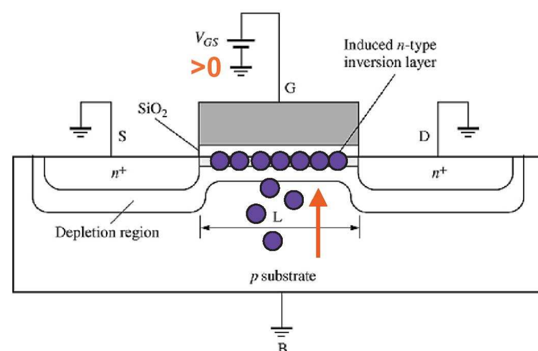
- We already know how to run SPICE. Why do we need to learn about models?
  - SPICE can accurately model the device using many parameters (30-100)
  - SPICE is nothing other than a matrix solver (KCL, KVL, linearized I-V equations)
- We need a reasonable compact model
  - To reason about circuit parameters (functionality, delay, power, robustness, …)
  - For hand/computer (e.g. Matlab) optimization
  - To check the SPICE simulation

Basic Operation (1)



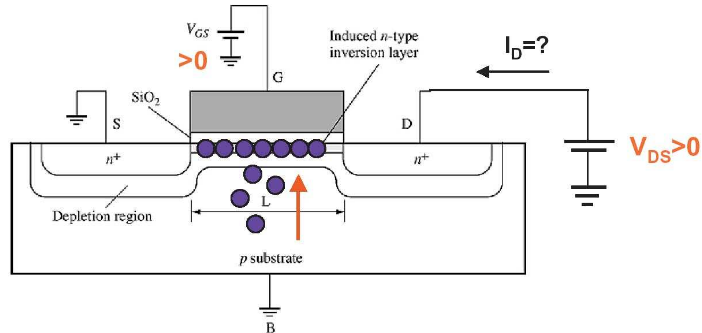
- Device is in cut-off region
- Simply, two back-to-back reverse biased pn diodes.

Basic Operation (2)



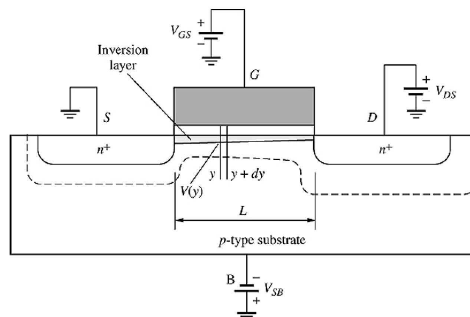
- With a positive gate bias, electrons are pulled toward the positive gate electrode
- Given a large enough bias, the electrons start to "invert" the surface (p→n type), a conductive channel forms
- Threshold voltage V_t

Basic Operation (3)



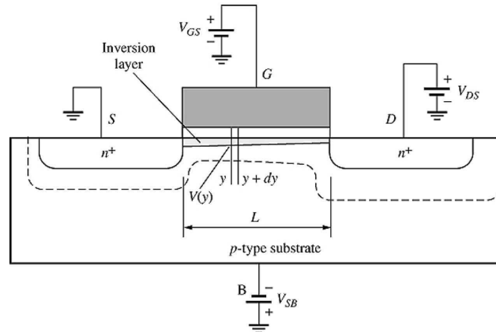
- Current flows from drain to source with a positive drain voltage
- What is current in terms of V_{gs}, V_{ds}, V_{bs}?

Assumptions



- Current is controlled by mobile charge in the channel
- Gradual channel: variation of E-field mainly perpendicular to the channel
- $v=\mu_e E$ (not true in short channel devices)
- Gen. & recomb. current is negligible: same I_{ds} across channel

MOS Current



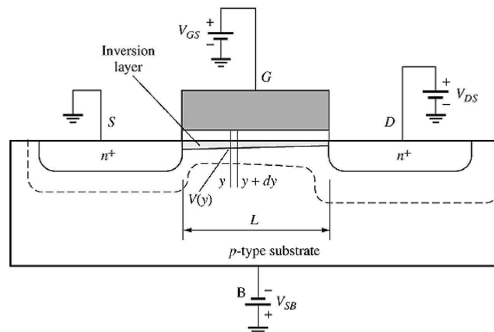
- From EE559, $Q_n = C_{ox}(V_{gs}-V_t-V(y))$
- By Ohm's law, $I_{ds} = Q_n \bullet v \bullet W$

$$= C_{ox}(V_{gs}-V_t-V(y)) \bullet \mu_e E \bullet W$$
$$= C_{ox}(V_{gs}-V_t-V(y)) \bullet \mu_e \bullet (dV(y)/dy) \bullet W$$
$$\rightarrow \quad I_{ds} \bullet dy = C_{ox}(V_{gs}-V_t-V(y)) \bullet \mu_e \bullet W \bullet dV(y)$$

PURDUE UNIVERSITY

IEI NRL

MOS Current



- Integrate this over the channel

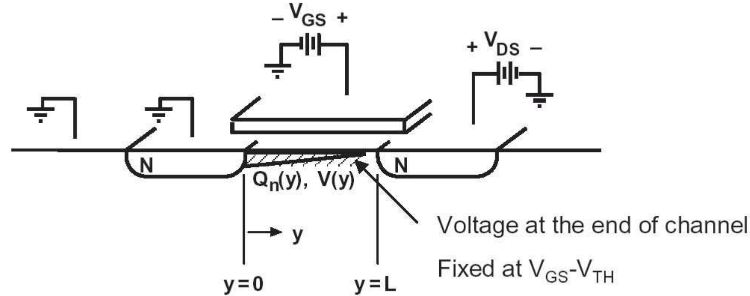
$$I_{ds} \bullet dy = C_{ox}(V_{gs}-V_t-V(y)) \bullet \mu_e \bullet W \bullet dV(y)$$
$$I_{ds} \bullet L = \mu_e C_{ox} W ((V_{gs}-V_t) V_{ds} - 0.5V_{ds}^2)$$

$$I_{ds} = \mu_e C_{ox} W/L ((V_{gs}-V_t) V_{ds} - 0.5V_{ds}^2) \text{ : linear mode}$$

PURDUE UNIVERSITY

IEI NRL

MOS Current



- $Q_n = C_{ox}(V_{gs}-V_t-V(y))$ → what if $V(y) > V_{gs}-V_t$
- Pinch-off: channel near drain disappears
  - Electrons which move along the channel to the pinch-off region are sucked across by the field, and enter the drain
  - Current through the channel is fixed

$I_{ds} = \mu_e C_{ox} W/(2L)\,(V_{gs}-V_t)^2$: saturation mode

Bulk Charge Model

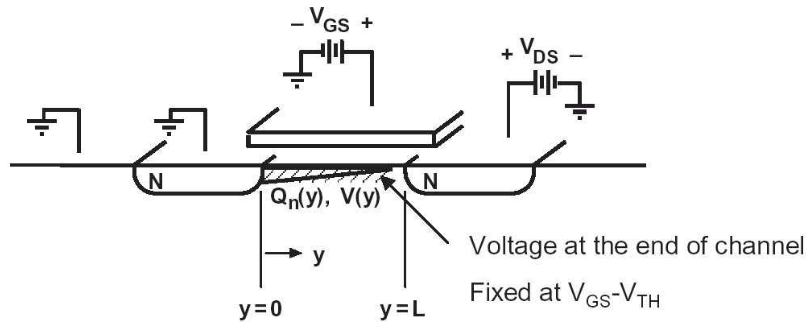
$$I_d = \mu_e C_{ox} \frac{W}{L}\left[(V_{gs}-V_t)V_{ds} - m\frac{1}{2}V_{ds}^2\right]$$

$$I_d = \mu_e C_{ox} \frac{W}{2L}\frac{(V_{gs}-V_t)^2}{m} \quad , \quad V_{dsat} = \frac{V_{gs}-V_t}{m}$$

$$m = 1 + \frac{C_{dep}}{C_{ox}} \quad : \quad body-effect\ coefficient$$

- More accurate than the square law model
- Considers inversion charge and bulk depletion charge
- Due to body effect across the channel

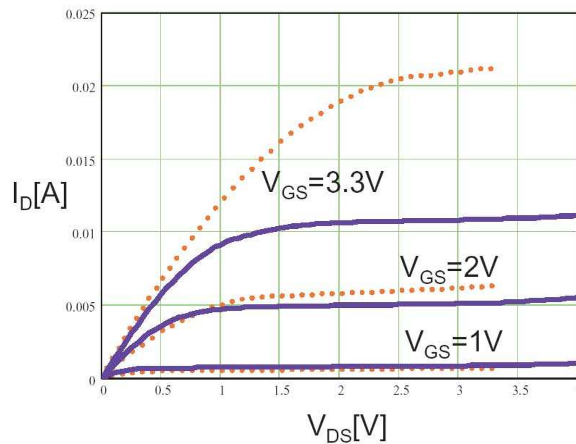
Channel Length Modulation



- **Pinch-off depletion layer width increases as the drain voltage increases**
- **Extreme case of this is punch-through**

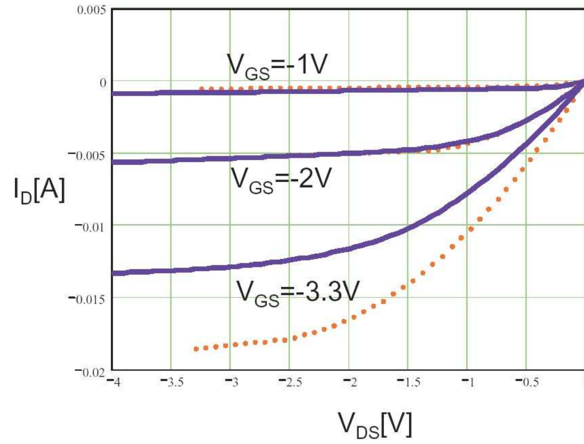
$$L = L_o - \zeta\, V_{ds} \qquad I_{ds} = I_{dsat} \times \frac{L_o}{L_o - \zeta V_{ds}} = I_{dsat} \times (1 + \lambda V_{ds})$$

Simulation versus Model (NMOS)



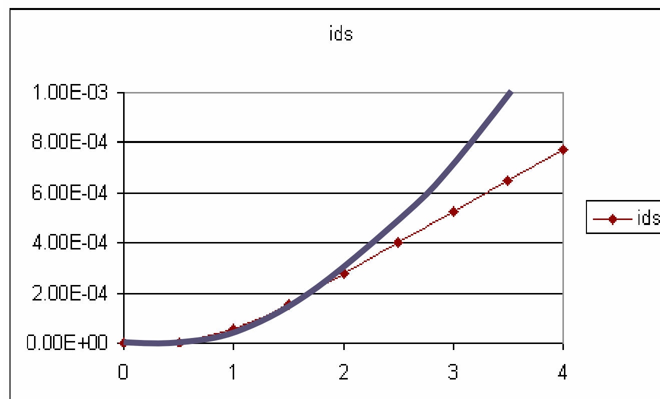
- **The square-law model doesn't match well with simulations**
- **Only fits for low V_{gs}, low V_{ds} (low E-field) conditions**

Simulation versus Model (PMOS)



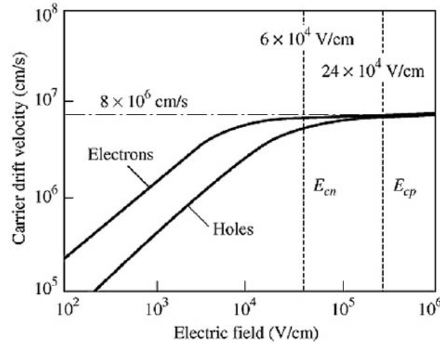
- Not as bad as the NMOS device
- Still large discrepancies at high E-field conditions

Simulation versus Model (I_{ds} vs. V_{gs})



- Saturation current does not increase quadratically
- The simulated curves looks like a straight line
- Main reason for discrepancy: velocity saturation

Velocity Saturation



- E-fields have gone up as dimensions scale
- Unfortunately, carrier velocity in silicon is limited
- Electron velocity saturates at a lower E-field than holes
- Mobility (μ_e=v/E) degrades at higher E-fields
- Simple piecewise linear model can be used

Velocity Saturation

$$v = \frac{\mu_e E}{\left(1+\left(E/E_c\right)^n\right)^{1/n}} \quad for\ E < E_c \qquad E_c = \frac{2v_{sat}}{\mu_e}$$

$$= v_{sat} \quad for\ E > E_c$$

[Toh, Ko, Meyer, JSSC, 8/1988]

μ_e, E_c must be determined for a given V_{gs}

$\mu_e \downarrow, E_c \uparrow$ for larger V_{gs} (almost linear)

- Modeled through a variable mobility
- n=1 for PMOS, n=2 for NMOS
- To get an analytical expression, let's assume n=1

Velocity Saturation

- **Plug it into the original current equation**

$$I_{ds} = C_{ox}(V_{gs} - V_t - mV(y)) \times \frac{\mu_e E}{1 + E/E_c} \times W$$

$$= C_{ox}(V_{gs} - V_t - mV(y)) \times \frac{\mu_e \, dV/dy}{1 + \frac{dV/dy}{E_c}} \times W$$

$$I_{ds} dy = C_{ox}\mu_e W (V_{gs} - V_t - mV(y) - \frac{I_{ds}}{W\mu_e E_c}) dV$$

$$\therefore \quad I_{ds} = \mu_e C_{ox} \frac{W}{L}\left((V_{gs} - V_t)V_{ds} - m\frac{V_{ds}^2}{2}\right) \times \frac{1}{1 + \frac{V_{ds}}{E_c L}}$$

PURDUE UNIVERSITY

NRL

Velocity Saturation Point

$$dI_{ds}/dV_{ds} = 0 \qquad V_{dsat} = \frac{2(V_{gs} - V_t)/m}{1 + \sqrt{1 + 2\mu_e(V_{gs} - V_t)/(mv_{sat}L)}}$$

$$I_{dsat} = C_{ox} W v_{sat}(V_{gs} - V_t)\frac{\sqrt{1 + 2\mu_e(V_{gs} - V_t)/(mv_{sat}L)} - 1}{\sqrt{1 + 2\mu_e(V_{gs} - V_t)/(mv_{sat}L)} + 1}$$

Long channel device

$L \to \infty$

$$V_{dsat} = \frac{V_{gs} - V_t}{m}$$

$$I_{dsat} = \mu_e C_{ox} \frac{W}{L}\frac{(V_{gs} - V_t)^2}{2m}$$

Short channel device

$L \to 0$

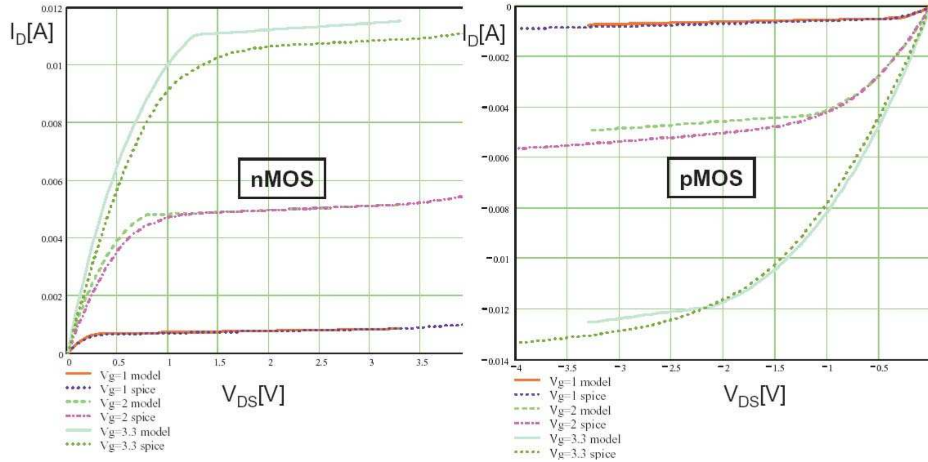
$$V_{dsat} = \sqrt{2v_{sat}L(V_{gs} - V_t)/(m\mu_e)}$$

$$I_{dsat} = C_{ox} W v_{sat}(V_{gs} - V_t)$$

PURDUE UNIVERSITY

NRL

Simulation versus Model



- **Model incorporating velocity saturation matches fairly well with simulation**

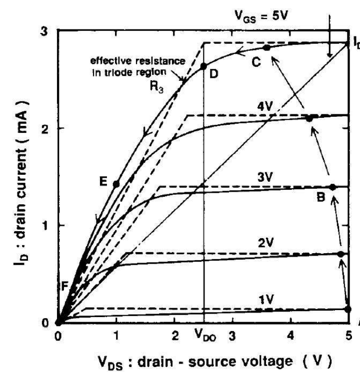
Alpha Power Law

- **Simple empirical model for short channel MOS**

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

[Sakurai and Newton, JSSC 1990]

- **Parameter α is between 1 and 2**
- **α=1-1.2 for short channel devices**
- **Parameters α and V_t are fitted to measured data for minimum square error → fitted V_t can be different from physical V_t**



Improving Short Channel MOS Model

- MOS current model
  = square law device (long channel)
  + body effect across channel (bulk charge model, long channel)
  + channel length modulation (long channel)
  + velocity saturation (short channel)
- V_t model
  = standard expression (long channel)
  + body effect (body bias, long channel)
  + V_t roll-off (barrier lowering, short channel)
  + Drain induced barrier lowering (short channel)
  + ...

PURDUE UNIVERSITY

NRL

Remember the Standard V_t Equation?

$$V_t = V_{fb} + |2\phi_B| + \frac{\sqrt{2qN_a\varepsilon_{si}|2\phi_B|}}{C_{ox}}$$

- Detailed derivation given in Taur's book
- Basically, three terms
  - Flat band voltage
  - $2\Phi_B$: the magic number for on-set of inversion
  - Oxide voltage

PURDUE UNIVERSITY

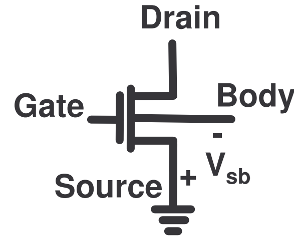
NRL

Body Effect (Back Bias)

$$V_{t0} = V_{fb} + |2\phi_B| + \frac{\sqrt{2qN_a\varepsilon_{si}|2\phi_B|}}{C_{ox}}$$

$$V_t = V_{fb} + |2\phi_B + V_{sb}| + \frac{\sqrt{2qN_a\varepsilon_{si}|2\phi_B + V_{sb}|}}{C_{ox}} - V_{sb}$$

$$V_t = V_{fb} + |2\phi_B| + \frac{\sqrt{2qN_a\varepsilon_{si}|2\phi_B + V_{sb}|}}{C_{ox}}$$



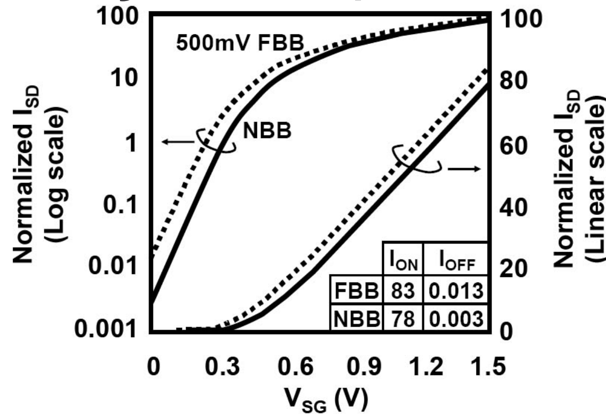
$V_{sb} > 0$: RBB
$V_{sb} < 0$: FBB

- Body effect degrades transistor stack performance
- However, we need a reasonable body effect for post silicon tuning techniques
- Reverse body biasing, forward body biasing

PURDUE UNIVERSITY

NRL

Body Effect (Back Bias)



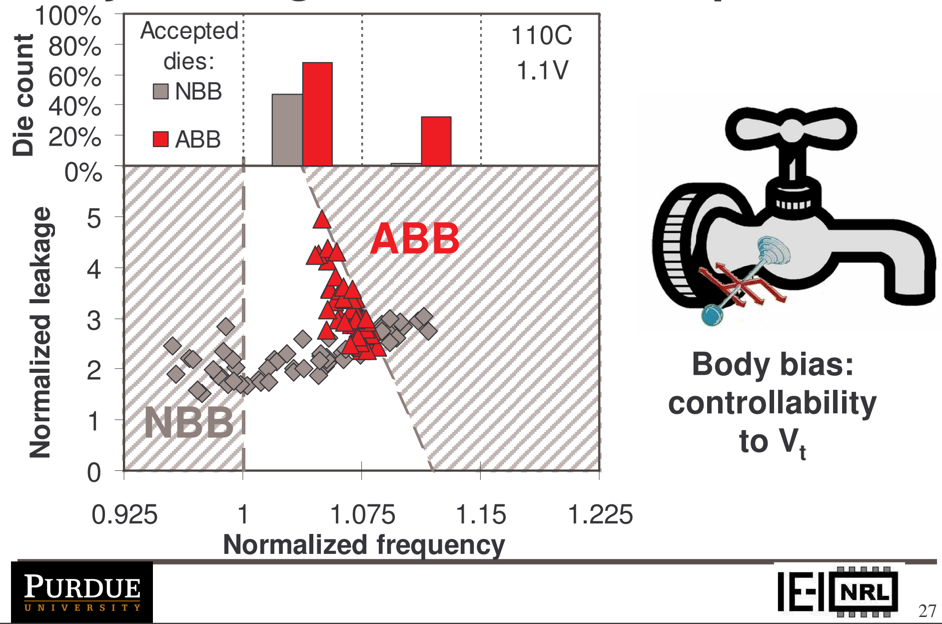
	I_{ON}	I_{OFF}
FBB	83	0.013
NBB	78	0.003

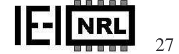
- V_t can be adjusted by applying FBB or RBB
  - Essential for low power and high performance
  - Will talk about body biasing extensively later on

PURDUE UNIVERSITY

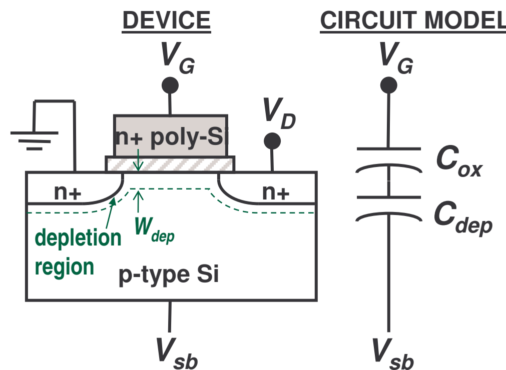
NRL

Body Biasing for Process Compensation





Substrate Sensitivity

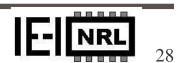


$$\left.\frac{dV_t}{dV_{sb}}\right|_{V_{sb}=0} = \frac{C_{dep}}{C_{ox}} \quad : \ substrate \ \ sensitivit\,y$$

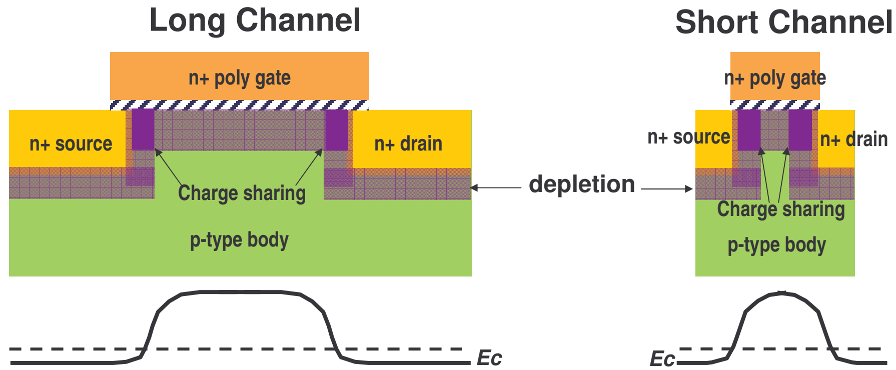
$$C_{dep} = \frac{\varepsilon_{Si}}{W_{dep}} = \sqrt{\frac{qN_a\varepsilon_{Si}}{4\phi_B}}$$

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$$

→ **Some ppl call this the body coeff.**

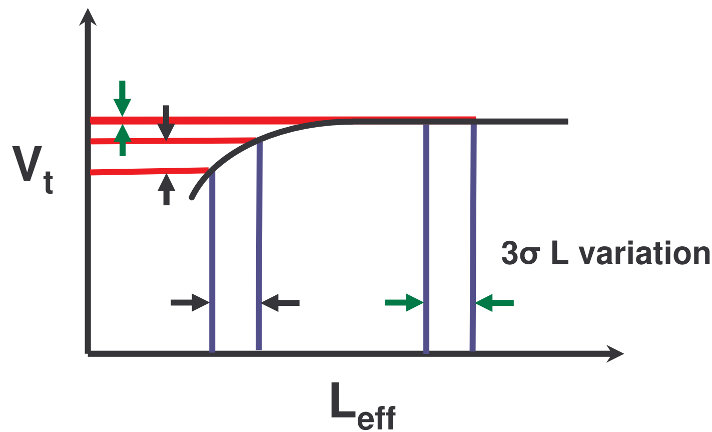


Short Channel Effect: V_t roll-off



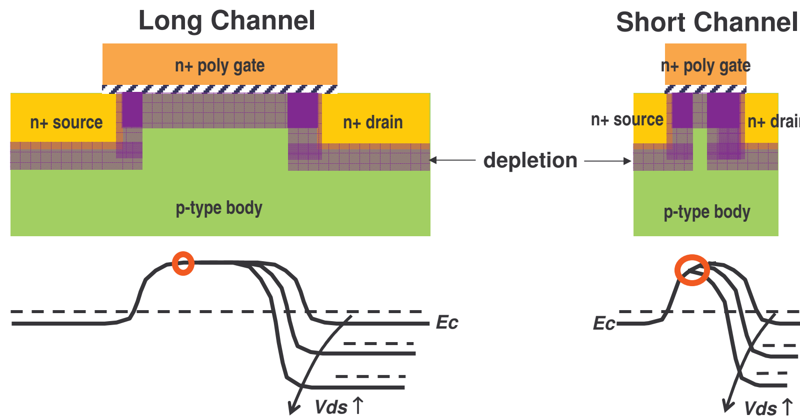
- Ability of gate & body to control channel charge diminishes as L decreases, resulting in Vt-roll-off and body effect reduction

Short Channel Effect: V_t roll-off



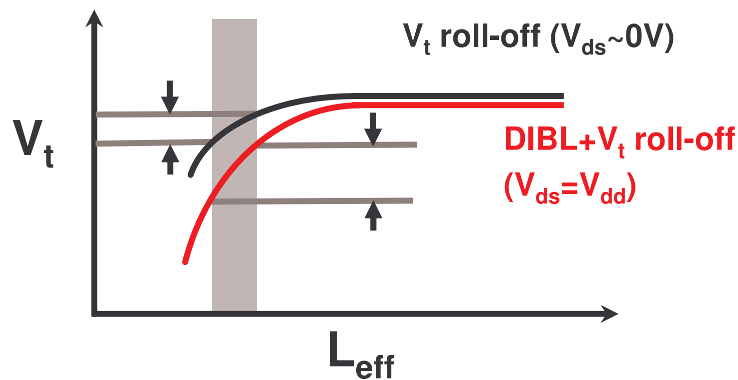
- 3σ V_t variation increases in short channel devices

Short Channel Effect: Drain Induced Barrier Lowering (DIBL)



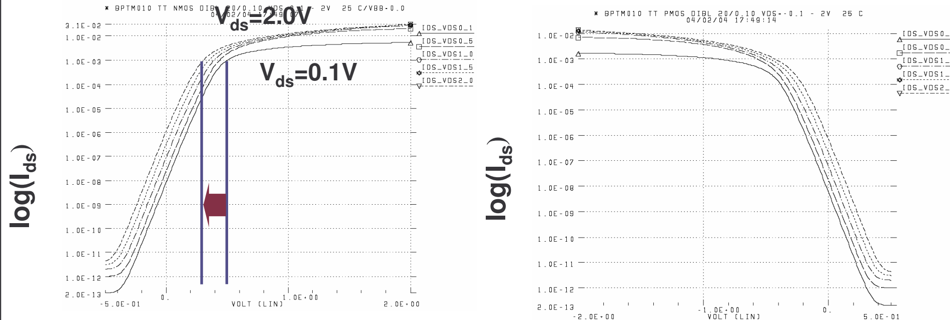
- Increase in V_{DS} reduces V_t and increases V_t-roll-off: DIBL

Short Channel Effect: Drain Induced Barrier Lowering (DIBL)



Short Channel Effect: DIBL

log(I_{ds})



V_{gs} (NMOS) V_{gs} (PMOS)

- DIBL coefficient $\lambda_d = \dfrac{\Delta V_t}{\Delta V_{ds}}$
- DIBL increases leakage current
- Dynamic V_{dd} can reduce leakage because of DIBL

PURDUE UNIVERSITY

NRL

33

Short Channel V_t Equation

$$V_t = V_{fb} + |2\phi_B| + \frac{\lambda_b}{C_{ox}}\sqrt{2qN_a\varepsilon_s(|2\phi_B| + V_{sb})} - \lambda_d V_{ds}$$

$$V_t = V_{fb} + |2\phi_B| + \frac{\sqrt{2qN_a\varepsilon_{si}|2\phi_B|}}{C_{ox}}$$ (Long channel V_t equation)

[Poon, IEDM, 1973]

$$\lambda_b = 1 - \left(\sqrt{1 + \frac{2W}{X_j}} - 1\right)\frac{X_j}{L}$$

[Ng, TED, 1993]

$$\lambda_d = \left[\frac{L}{2.2\mu m^{-2}(T_{ox} + 0.012\mu m)(W_{sd} + 0.15\mu m)(X_j + 2.9\mu m)}\right]^{-2.7}$$

PURDUE UNIVERSITY

NRL

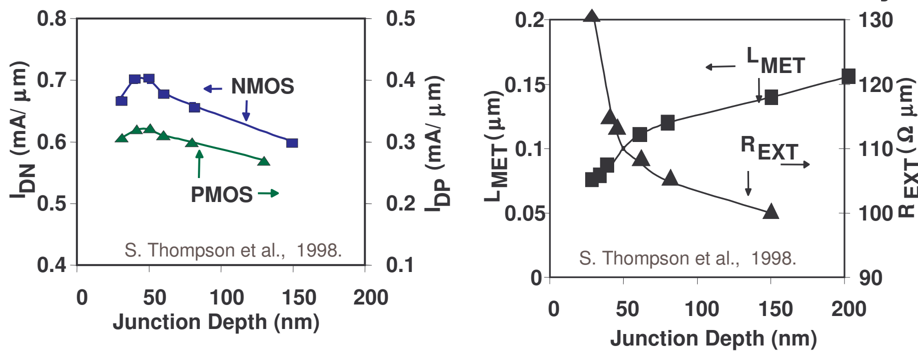
34

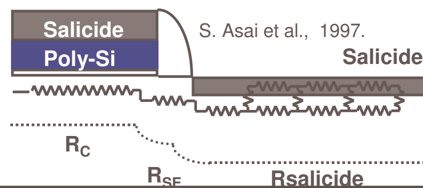
Improving Short Channel MOS Model

- MOS current model
  - = square law device (long channel)
  - + body effect across channel (bulk charge model, long channel)
  - + channel length modulation (long channel)
  - + velocity saturation (short channel)
- V_t model
  - = standard expression (long channel)
  - + body effect (body bias, long channel)
  - + V_t roll-off (barrier lowering, short channel)
  - + Drain induced barrier lowering (short channel)
  - + ...

PURDUE UNIVERSITY

Transistor Scaling Challenges - X_j





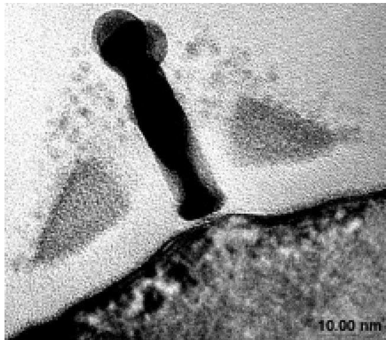
PURDUE UNIVERSITY

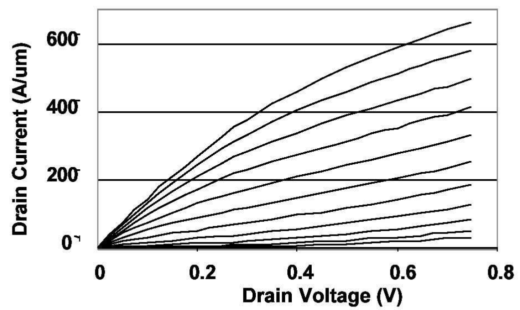
Effect of Series Resistance

PURDUE UNIVERSITY

37

Effect of Series Resistance
(10nm Device)

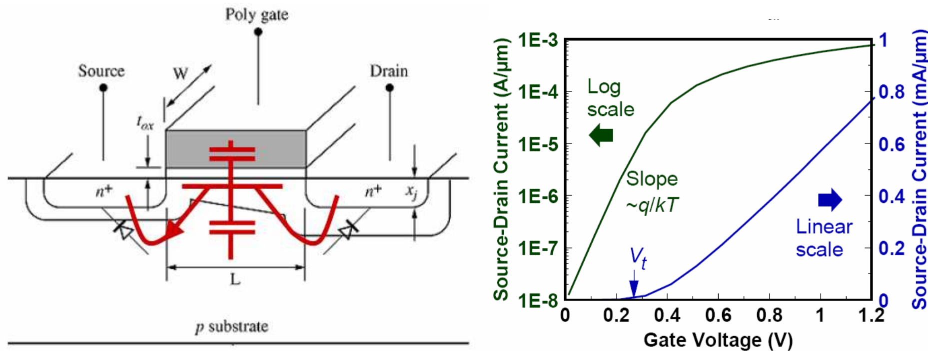




PURDUE UNIVERSITY

38

Sub-Threshold Conduction



- NPN BJT is formed in sub-threshold region
- Only difference with a real BJT is that the base voltage is controlled through a capacitive divider, and not directly by a electrode
- Like in a BJT, current is exponential to V_{be}

PURDUE UNIVERSITY

Sub-Threshold Current

$$I_d = \frac{W}{L}\mu_{eff} C_{ox}\left(\frac{k_B T}{q}\right)^2 (m-1)\, e^{q(V_{gs}-V_t)/mkT}\,(1-e^{-qV_{ds}/kT})$$

PURDUE UNIVERSITY

Sub-Threshold Swing

$$S = m\frac{kT}{q}\ln 10 \;\; (mV/dec) \quad , \quad m = 1 + \frac{C_{dep}}{C_{ox}}$$

- Smaller S-swing is better
- Ideal case: m=1 (C_{ox}>>C_{sub})
  - Fundamental limit = 1 * 26mV * ln10
    = 60 mV/dec @ RT
  - Can only be achieve by device geometry (FD-SOI)
- Typical case: m≈1.3
  - S = 1.3 * 26mV * ln10 ≈ 80 mV/dec @ RT
  - At worst case temperature (T=110C), S ≈ 100 mV/dec

V_{dd} and V_t Scaling

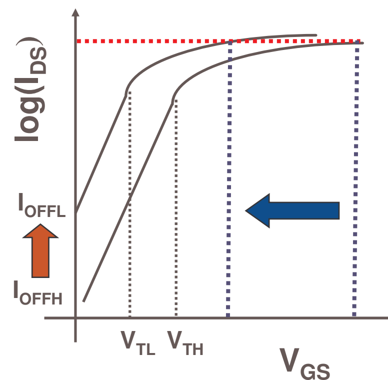
Performance vs Leakage:

V_T ↓ I_{OFF} ↑ $I_D(SAT)$ ↑

$$I_{OFF} \propto \frac{W_{eff}}{L_{eff}} K_1 e^{-V_T / mkT/q}$$

$$I_D(SAT) \propto \frac{W_{eff}}{L_{eff}} K_2 (V_{GS} - V_T)^2$$

$$I_D(SAT) \propto K_3 W_{eff} C_{ox} \upsilon_{SAT} (V_{GS} - V_T)$$



⇨ As V_t decreases, sub-threshold leakage increases

⇨ Leakage is a barrier to voltage scaling

V_{dd} and V_t Scaling

- V_t cannot be scaled indefinitely due to increasing leakage power (constant sub-threshold swing)
- Example

  CMOS device with S=100mV/dec has I_{ds}=10μA/μm @ V_t=500mV

  I_{off}=10μA/μm x 10^{-5} = 0.1 nA/μm

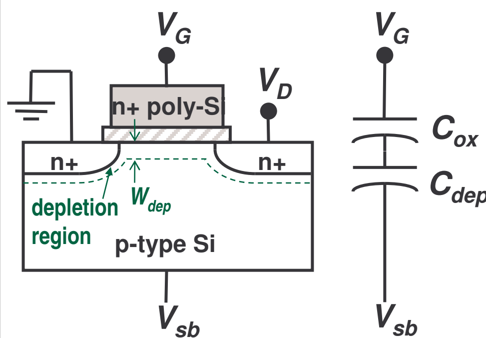
  Now, consider we scale the V_t to 100mV

  I_{off}=10μA/μm x 10^{-1} = 1 μA/μm

  Suppose we have 1B transistors of width 1μm

  I_{sub}=1μA/μm x 1B x 1μm = 100 A !!

S-Swing & Substrate Sensitivity



$$S = m\frac{kT}{q}\ln 10\,, \quad m = 1 + \frac{C_{dep}}{C_{ox}}$$

$$\left.\frac{dV_t}{dV_{sb}}\right|_{V_{sb}} = \frac{C_{dep}}{C_{ox}} = m - 1$$

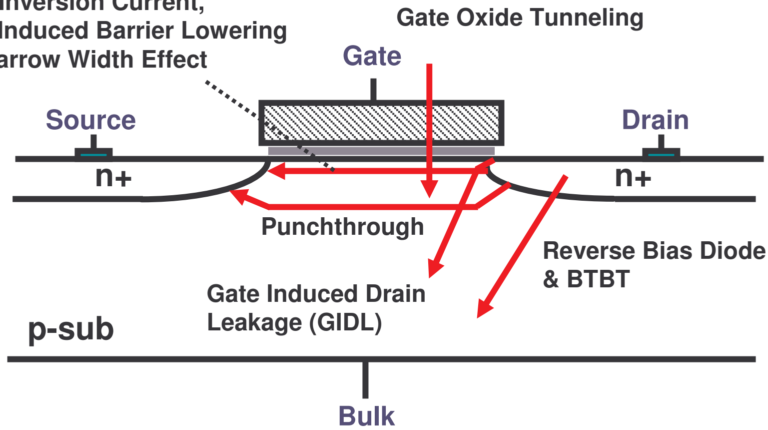
$$C_{dep} = \frac{\varepsilon_{Si}}{W_{dep}} = \sqrt{\frac{qN_a\varepsilon_{Si}}{4\phi_B}}$$

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$$

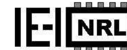
$$m\uparrow (N_a \uparrow or\ t_{ox} \uparrow) \Rightarrow S\uparrow \text{ (bad)} \Rightarrow sub.sensitivty \uparrow \text{ (good)}$$

$$m\downarrow (N_a \downarrow or\ t_{ox} \downarrow) \Rightarrow S\downarrow \text{ (good)} \Rightarrow sub.sensitivty \downarrow \text{ (bad)}$$

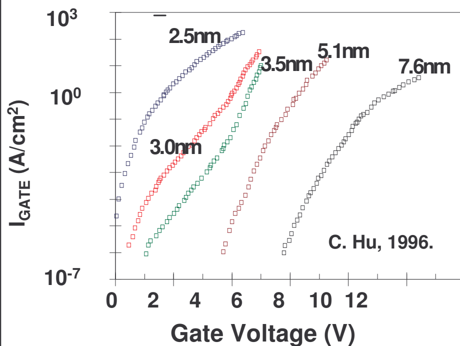
Leakage Components

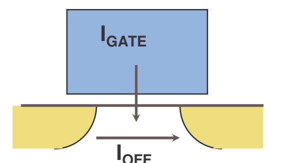


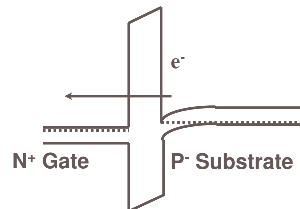
[Keshavarzi, Roy, and Hawkins, ITC 1997]



Gate Oxide Tunneling Leakage







Electrical T_{ox} = Physical T_{ox} + 1nm

50% poly depletion, 50% quantum effect

Problem in analog (diff pairs, current mirror), domino

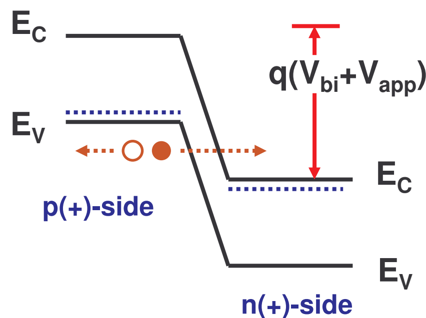
Gate Oxide Tunneling Leakage

- Quantum mechanics tells us that there is a finite probability for electrons to tunnel through oxide
- Probability of tunneling is higher for very thin oxides
- NMOS gate leakage is much larger than PMOS
- Gate leakage has the potential to become one of the main showstoppers in device scaling

$$I_{gate} = AE_{ox}^{2} e^{-B/E_{ox}} \quad , \quad E_{ox} = \frac{V_{dd} - V_t}{t_{ox}}$$

PURDUE UNIVERSITY

Band-to-Band Tunneling Leakage



S/D junction BTBT Leakage

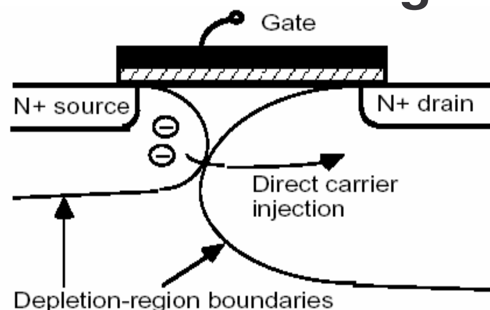
- Reversed biased diode band-to-band tunneling
  - High junction doping: “Halo” profiles
  - Large electric field and small depletion width at the junctions

PURDUE UNIVERSITY

Gate Induced Drain Leakage (GIDL)

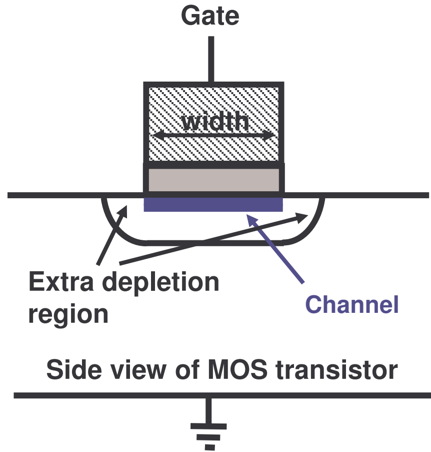
- Appears in high E-field region under gate/drain overlap causing deep depletion
- Occurs at low V_g and high V_d bias
- Generates carriers into substrate from surface traps, band-to-band tunneling
- Localized along channel width between gate and drain
- Thinner oxide, higher V_{dd}, lightly-doped drain enhance GIDL
- High field between gate and drain increases injection of carriers into substrate

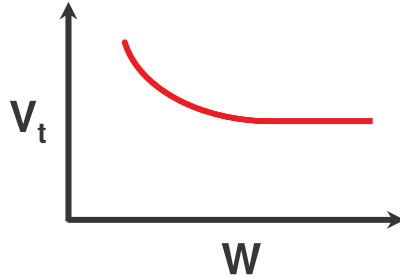
Punch Through



- If the channel length becomes too short, the drain side depletion region can touch the source side
- Reduces the barrier for electron injection from source to drain
- Sub-surface version of weak inversion conduction

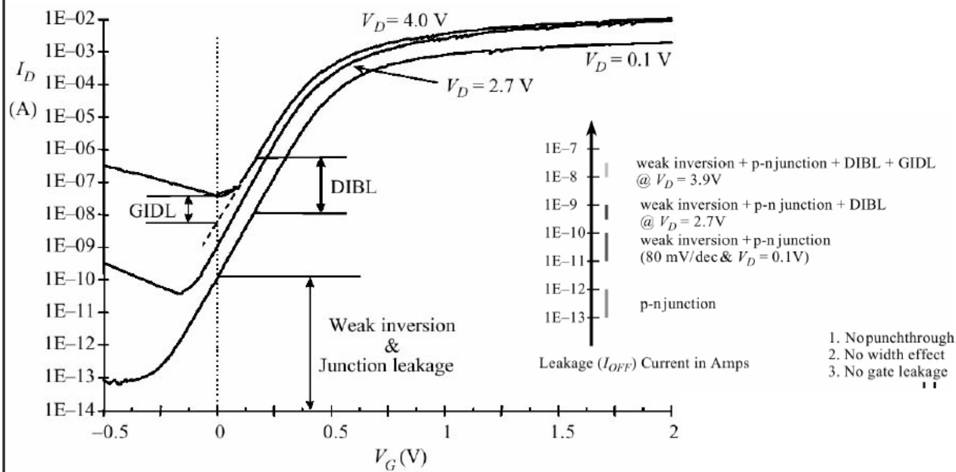
Narrow Width Effect





- Depletion region extends outside of gate controlled region
- Opposite to V_t roll-off
- Depends on isolation technology

Leakage Components



[IEEE press, 2000]

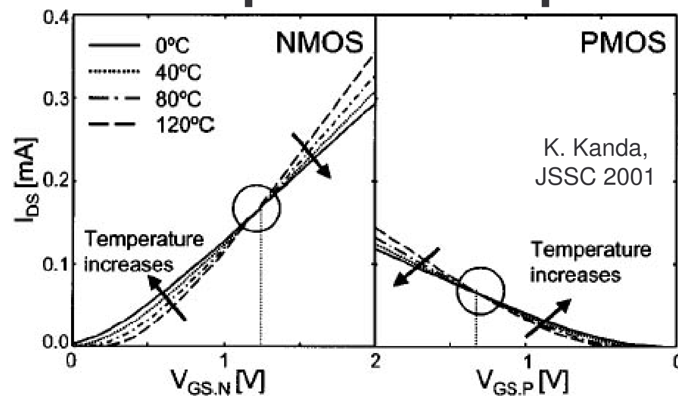
Temperature Dependence

- **Mobility degrades at higher temperatures**

$$\mu = \mu_o \left(\frac{T}{T_o}\right)^{-3/2}$$

  - Scattering increases with vibrating atoms
  - Temp change from 27C to 130C decreases current to 0.65.
  - The circuit will run 1.6 times slower.
- **V_t decreases at higher temperatures**
  - Electrons on the source side gain more energy
  - Sub-threshold leakage will increase
  - The circuit will run faster
- **Question: What happens to circuit performance at high temperatures? Slower or Faster?**

Positive Temperature Dependence



- **Depending on V_{dd} and V_t, positive dependency can occur**
  - Advantageous phenomenon for low V_{dd} design
  - Will change the design validation process for worst case conditions

Summary

- IC designers should be aware of the technology issues
- Short channel behaviors
  - Velocity saturation
  - Vt roll-off
  - Drain induced barrier lowering
  - Series resistance
  - Leakage: sub-threshold, gate, BTBT, punchthrough
  - Positive temperature dependence
  - Gate depletion, quantum confinement
- The issues can be dealt with at different levels of abstraction (technology, circuits, CAD, architecture, software, etc)

PURDUE UNIVERSITY

NRL