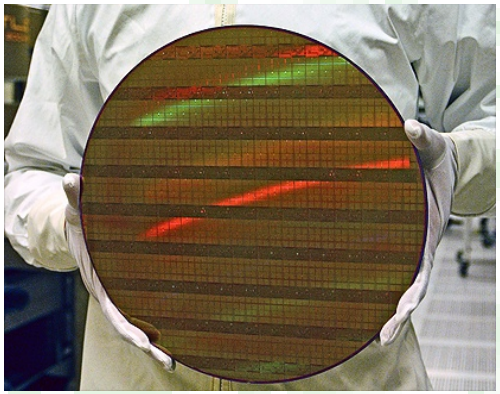
# Compact Modeling of Semiconductor Devices: MOSFET

Yogesh S. Chauhan
Assistant Professor and Ramanujan Fellow
*Nanolab*, Department of Electrical Engineering
IIT Kanpur
Email: chauhan@iitk.ac.in
Homepage – http://home.iitk.ac.in/~chauhan/

# Outline

- Compact Modeling
- MOSFET
- Drain Current in MOSFET
- Smoothing Functions
- Terminal Charges
- Scaling
- FinFET

# Compact Modeling or SPICE Modeling



**Medium of information exchange**

- Good model should be
  - **Accurate:** Trustworthy simulations.
  - **Simple:** Parameter extraction is easy.
- Balance between accuracy and simplicity depends on end application

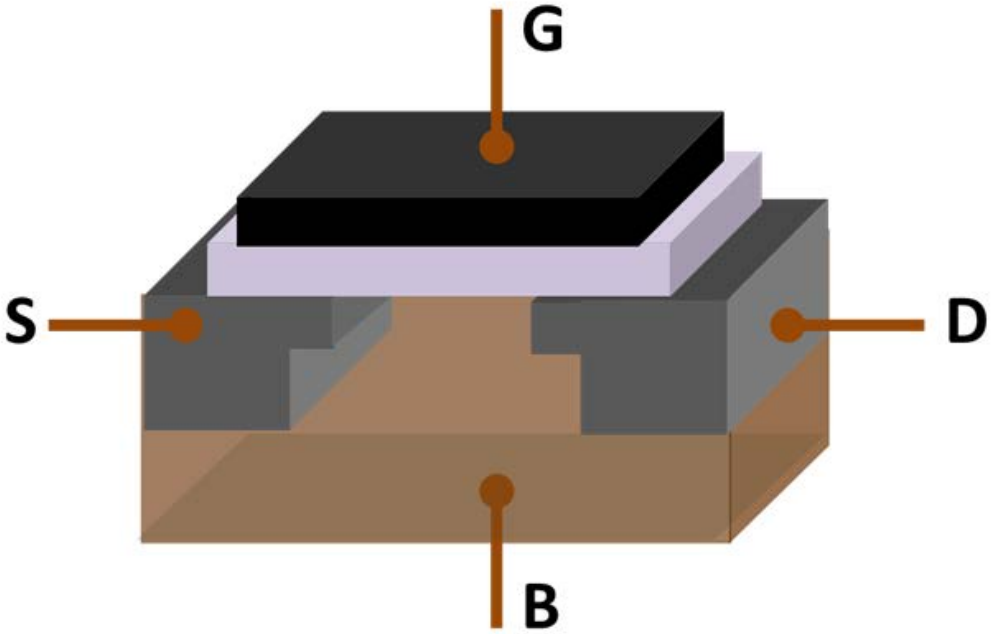
- Excellent Convergence
- Simulation Time – ~μs*ec*
- Accuracy requirements
  - ~ 1% RMS error after fitting
- Example: BSIM6, BSIM-CMG

# Industry Standard Compact Models

- Standardization Body – Compact Model Coalition
- CMC Members – EDA Vendors, Foundries, IDMs, Fabless, Research Institutions/Consortia
- CMC is by the industry and for the industry

# What is MOSFET?

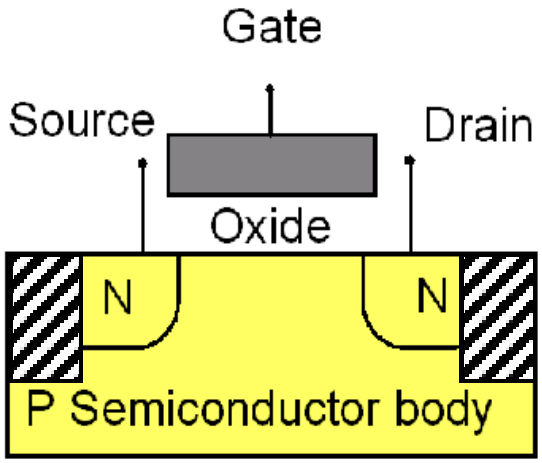
- MOSFET is a transistor used for amplifying or switching electronic signals.

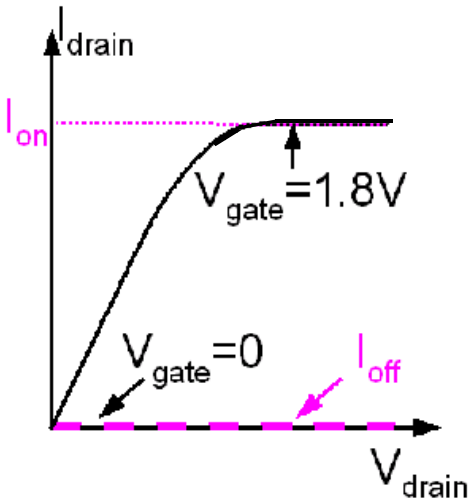



# Introduction to MOSFET

- Building block of Gb memory chips, GHz microprocessors, analog, and RF circuits.

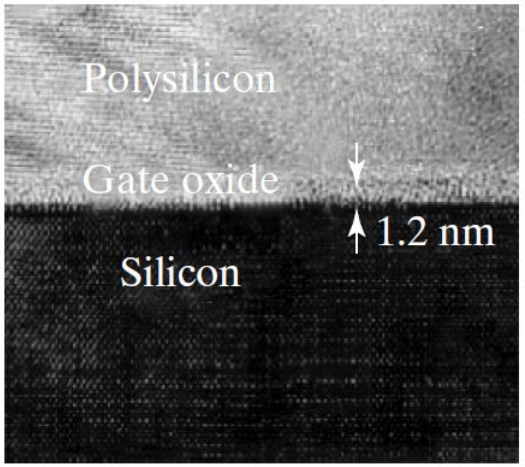
Basic MOSFET structure and IV characteristics




What is desirable: large $I_{on}$, small $I_{off}$

*Polysilicon gate & $SiO_2$*



# How can we simulate MOSFET based circuits?

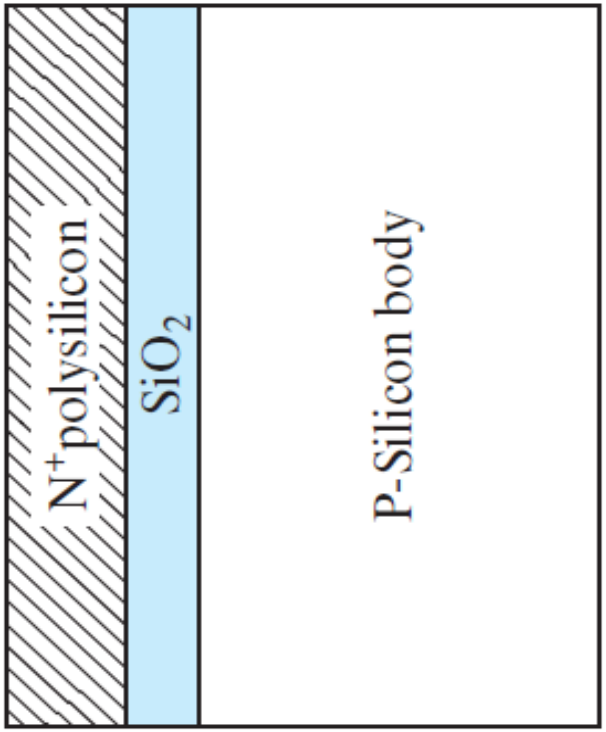
- We need
  - Currents

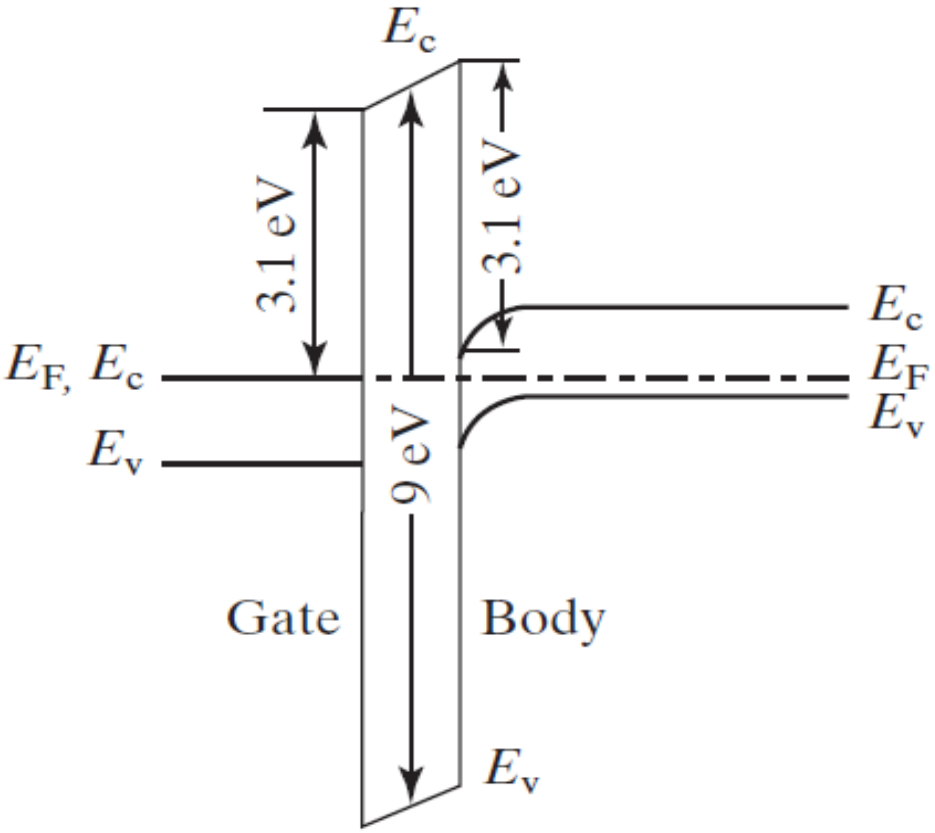
$$I_{ds} = W.Q_{inv}.v = W.Q_{inv}.\mu_{ns}E$$

$$I_{ds} = W.C_{ox}(V_{gs} - V_t).\mu_{ns}\frac{dV_{ch}}{dx}$$

  - Charges (for capacitance)

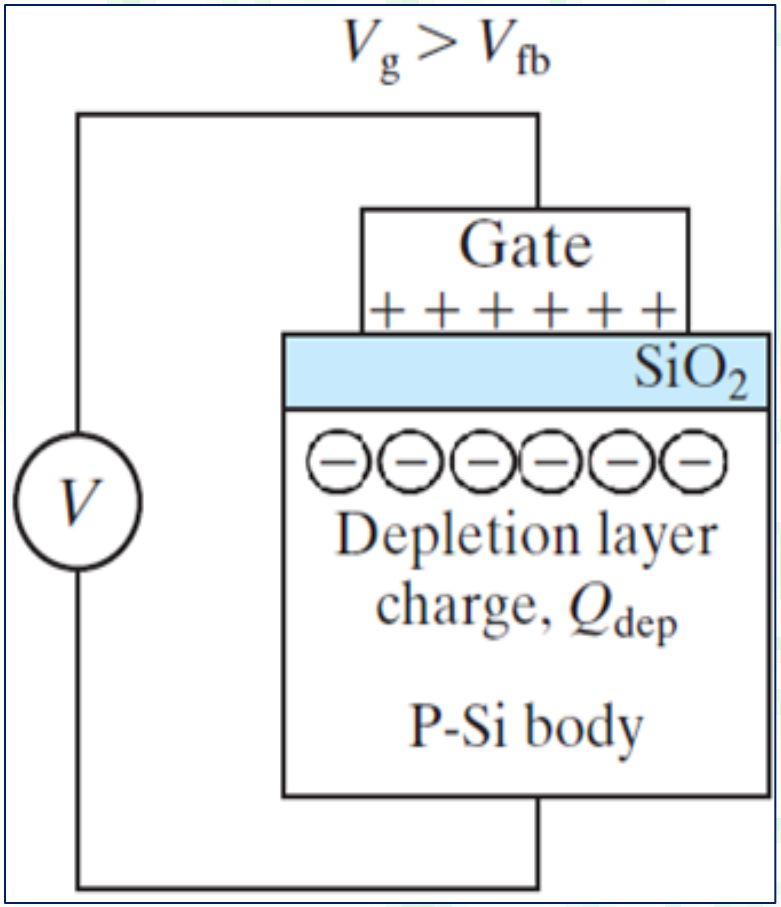
# *Energy Band Diagram in Equilibrium*



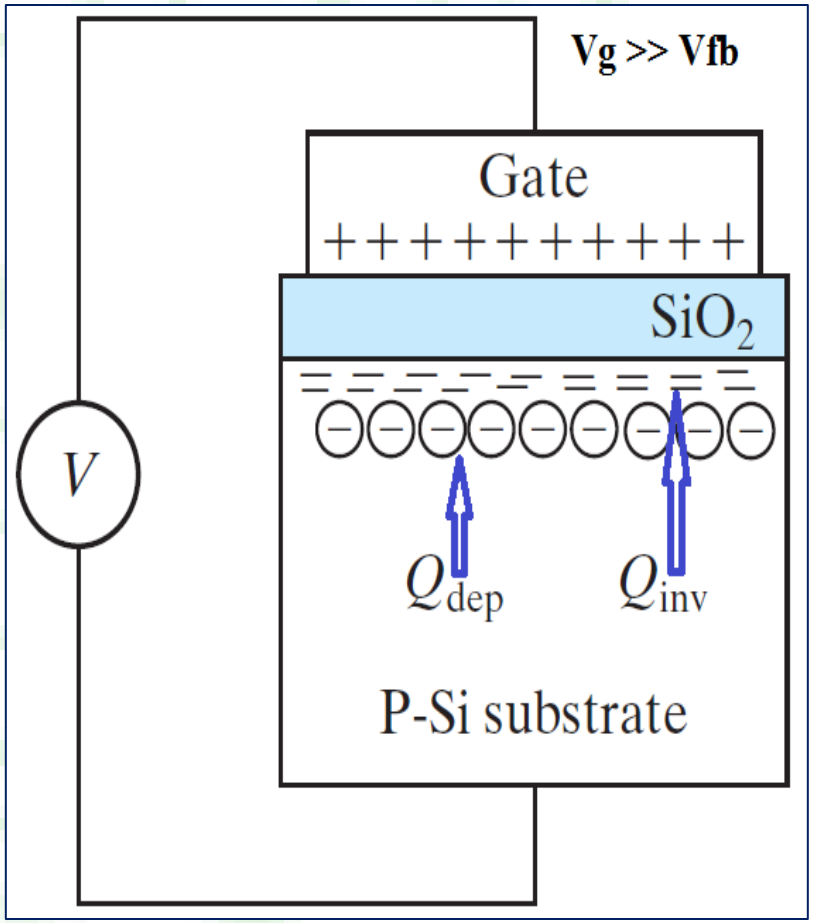



# Depletion and Inversion

Surface is depleted of holes




Surface is inverted



# Threshold Condition and Threshold Voltage

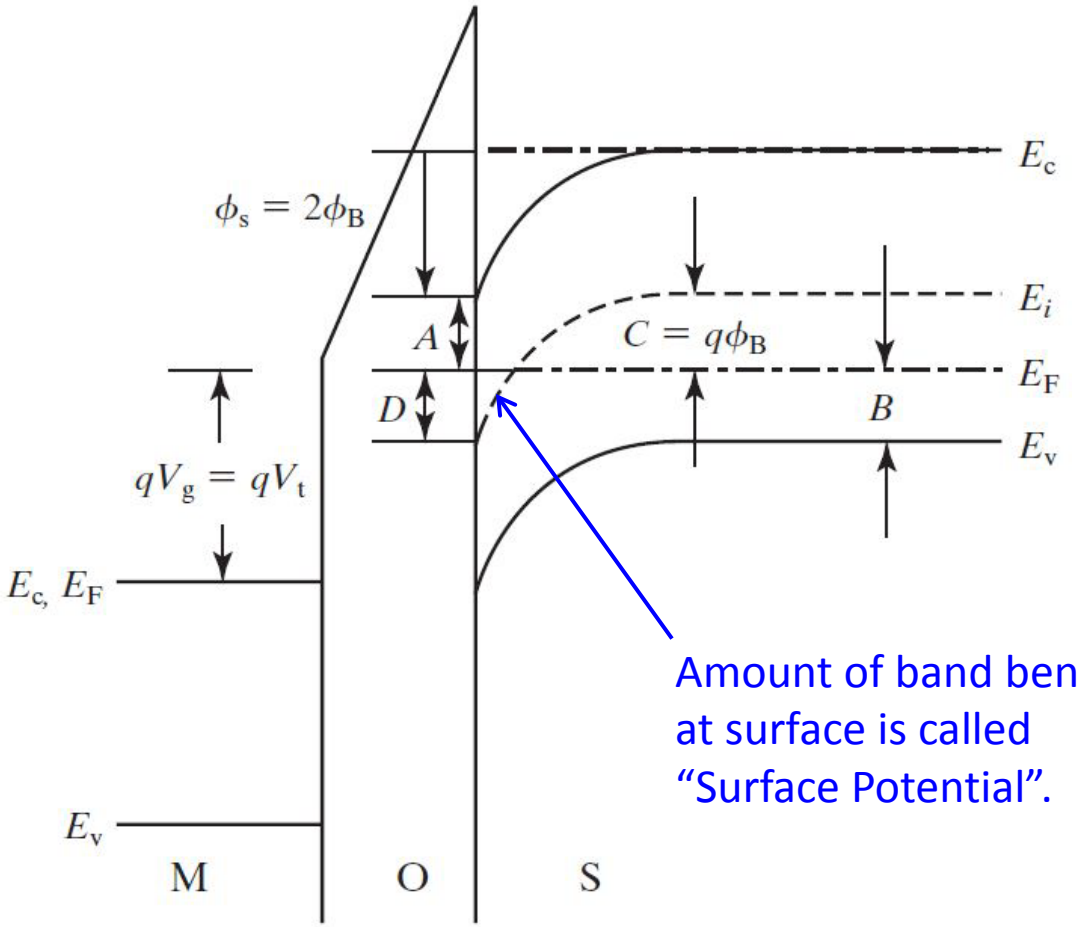
**Threshold (of inversion):**

$n_s = N_a$, or

$(E_c - E_f)_{surface} = (E_f - E_v)_{bulk}$, or

✧ $A=B$, and $C = D$

$$\psi_{st} = 2\phi_b = 2\frac{kT}{q} ln\left(\frac{N_a}{n_i}\right)$$




$$V_{ox} = \frac{\sqrt{qN_a 2\varepsilon_s 2\phi_B}}{C_{ox}}$$

$$V_g = V_{fb} + \psi_s + V_{ox}$$

$$V_t = V_g \text{ at threshold} = V_{fb} + 2\phi_B + \frac{\sqrt{qN_a 2\varepsilon_s 2\phi_B}}{C_{ox}}$$

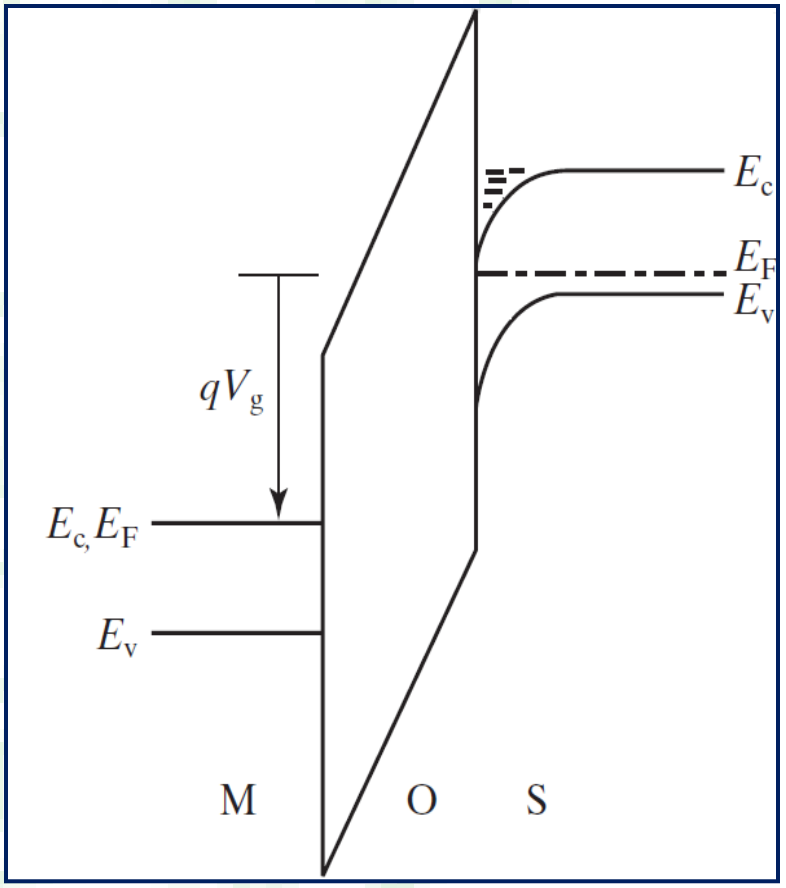
# Inversion Layer Charge

- Applied Gate Voltage

$$V_g = V_{fb} + 2\phi_B - \frac{Q_{dep}}{C_{ox}} - \frac{Q_{inv}}{C_{ox}}$$

$$= V_{fb} + 2\phi_B + \frac{\sqrt{qN_a 2\varepsilon_s 2\phi_B}}{C_{ox}} - \frac{Q_{inv}}{C_{ox}}$$

$$= V_t - \frac{Q_{inv}}{C_{ox}}$$



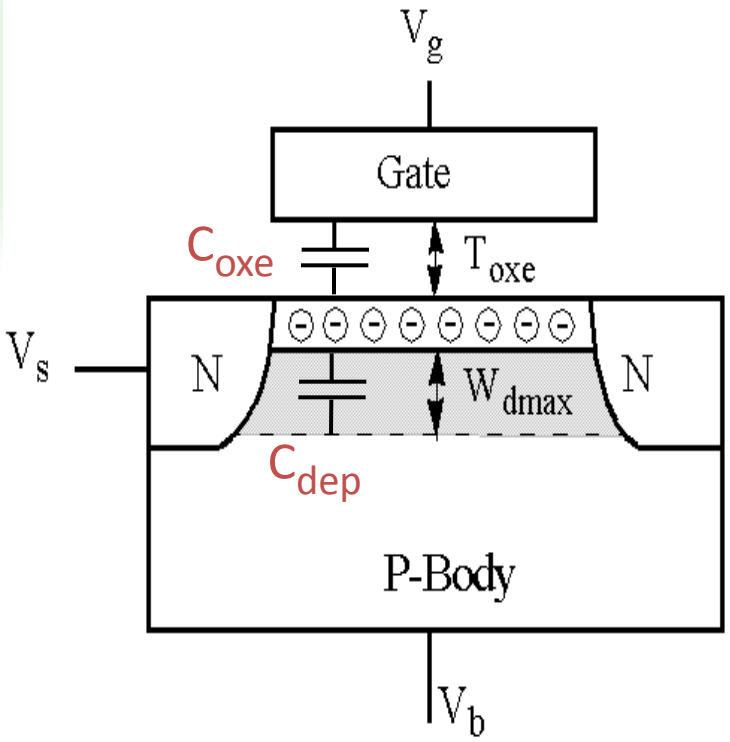

***Inversion Layer Charge*** ➡

$$Q_{inv} = -C_{ox}(V_g - V_t)$$

# MOSFET $V_t$ and the Body Effect

- **Two capacitors => two charge components**

$$C_{dep} = \frac{\varepsilon_s}{W_{d\max}}$$




$$Q_{inv} = -C_{oxe}(V_{gs} - V_t) + C_{dep}V_{sb}$$

$$= -C_{oxe}(V_{gs} - (V_t + \frac{C_{dep}}{C_{oxe}}V_{sb}))$$

- Redefine $V_t$ as

$$V_t(V_{sb}) = V_{t0} + \frac{C_{dep}}{C_{oxe}}V_{sb} = V_{t0} + \alpha V_{sb}$$

# Uniform Body Doping

- In earlier generations of MOSFETs, the body doping density is more or less uniform and $W_{dmax}$ varies with $V_{sb}$.
- In that case, the theory for the body effect is more complicated.

$$V_t = V_{t0} + \frac{\sqrt{qN_a 2\varepsilon_s}}{C_{oxe}}\left(\sqrt{2\phi_B + V_{sb}} - \sqrt{2\phi_B}\right)$$

$$\equiv V_{t0} + \gamma\left(\sqrt{2\phi_B + V_{sb}} - \sqrt{2\phi_B}\right)$$

$\gamma$ is the ***body-effect parameter***.

# Threshold Voltage Modeling

- Long/Wide Channel Model With Uniform Doping

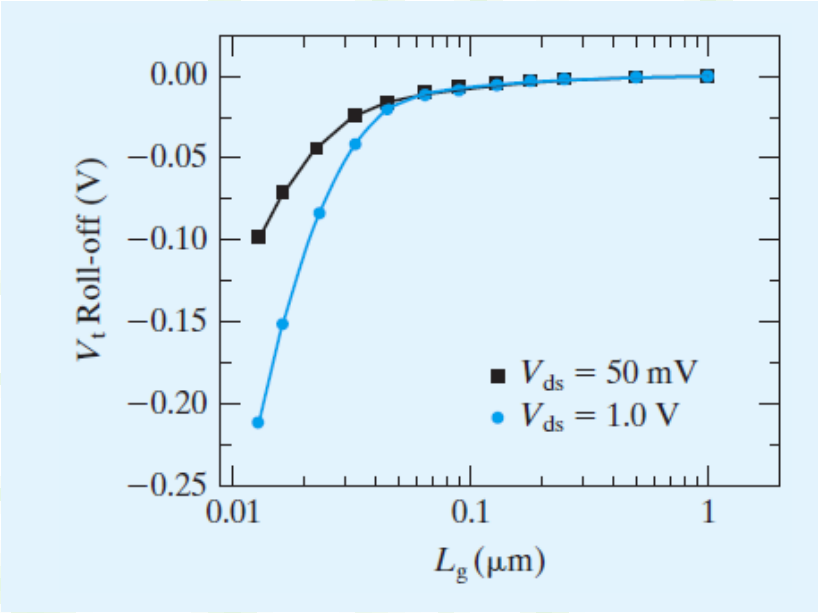
$$V_{th} = VFB + \Phi_s + \gamma\sqrt{\Phi_s - V_{bs}} = VTH0 + \gamma\left(\sqrt{\Phi_s - V_{bs}} - \sqrt{\Phi_s}\right)$$

  - $V_{FB}$=flat band voltage
  - $V_{TH0}$=threshold voltage of device at zero substrate bias
  - $\gamma$ is the body bias coefficient given by

$$\gamma = \frac{\sqrt{2\,q\,\varepsilon_{si}\,N_{substrate}}}{C_{oxe}}$$

# Threshold Voltage Modeling

- As channel length gets shorter $V_{th}$ shows a greater dependence on
  - Short-channel effect ← *Higher Off Current Leakage*
  - DIBL ← *Higher Off Current Leakage at high* $V_{ds}$




$$\Delta V_{th}\left(SCE, DIBL\right) = -\theta_{th}\left(L_{eff}\right)\cdot\left[2\left(V_{bi}-\Phi_s\right)+V_{ds}\right]$$

$$V_{bi} = \frac{k_B T}{q}\ln\left(\frac{NDEP\cdot NSD}{n_i^2}\right)$$

# Threshold Voltage Modeling

- The complete $V_{th}$ model implemented in SPICE as

$$V_{th} = VTH0 + \left(K_{1ox}\cdot\sqrt{\Phi_s - V_{bseff}} - K1\cdot\sqrt{\Phi_s}\right)\sqrt{1+\frac{LPEB}{L_{eff}}} - K_{2ox}V_{bseff}$$

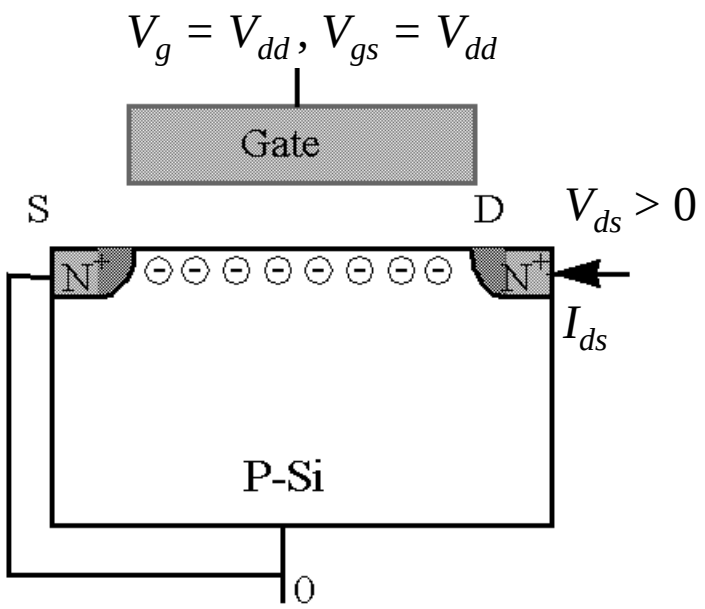
$$+ K_{1ox}\left(\sqrt{1+\frac{LPE0}{L_{eff}}} - 1\right)\sqrt{\Phi_s} + \left(K3 + K3B\cdot V_{bseff}\right)\frac{TOXE}{W_{eff}' + W0}\Phi_s$$

$$- 0.5\cdot\left[\frac{DVT0W}{\cosh\left(DVT1W\frac{L_{eff}W_{eff}'}{l_{tw}}\right)-1} + \frac{DVT0}{\cosh\left(DVT1\frac{L_{eff}}{l_t}\right)-1}\right](V_{bi} - \Phi_s)$$

$$- \frac{0.5}{\cosh\left(DSUB\frac{L_{eff}}{l_{t0}}\right)-1}\left(ETA0 + ETAB\cdot V_{bseff}\right)\cdot V_{ds} - nv_t.\ln\left(\frac{L_{eff}}{L_{eff} + DVTP0.\left(1+e^{-DVTP1V_{DS}}\right)}\right)$$

$$- \left(DVTP5 + \frac{DVTP2}{L_{eff}^{DVTP3}}\right)\cdot\tanh\left(DVTP4\cdot V_{ds}\right)$$

# Surface Mobility




$$I_{ds} = W \times Q_{inv} \times v = W Q_{inv} \mu_{ns} \mathbf{E}$$

$$= W Q_{inv} \mu_{ns} V_{ds} / L$$

$$= W C_{oxe} (V_{gs} - V_t) \mu_{ns} V_{ds} / L$$

- Scattering mechanisms
  - Phonon scattering
  - Coulomb scattering
  - Interface roughness scattering

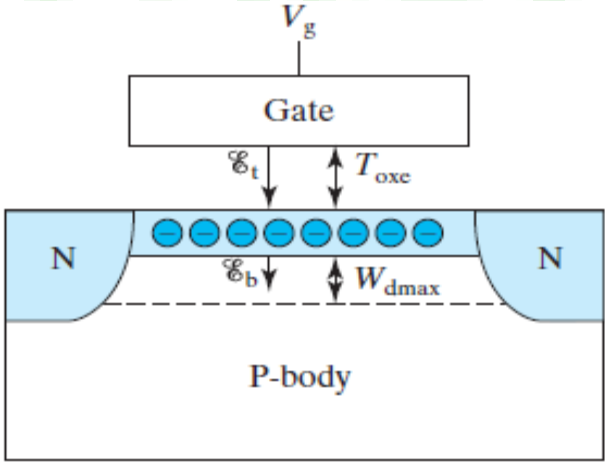
# Surface Mobility

- Mobility is a function of the average of the fields at the bottom and the top of the inversion charge layer, $E_b$ and $E_t$.

From Gauss's Law, $\mathcal{E}_b = -Q_{dep}/\varepsilon_s$

$$V_g = V_{fb} + \psi_s + V_{ox} \quad \Rightarrow \quad V_t = V_{fb} + \phi_{st} - Q_{dep}/C_{oxe}$$

Therefore, $\mathcal{E}_b = \dfrac{C_{oxe}}{\varepsilon_s}(V_t - V_{fb} - \phi_{st})$




$$\mathcal{E}_t = -(Q_{dep} + Q_{inv})/\varepsilon_s$$

$$= \mathcal{E}_b - Q_{inv}/\varepsilon_s = \mathcal{E}_b + \frac{C_{oxe}}{\varepsilon_s}(V_{gs} - V_t)$$

$$= \frac{C_{oxe}}{\varepsilon_s}(V_{gs} - V_{fb} - \phi_{st})$$

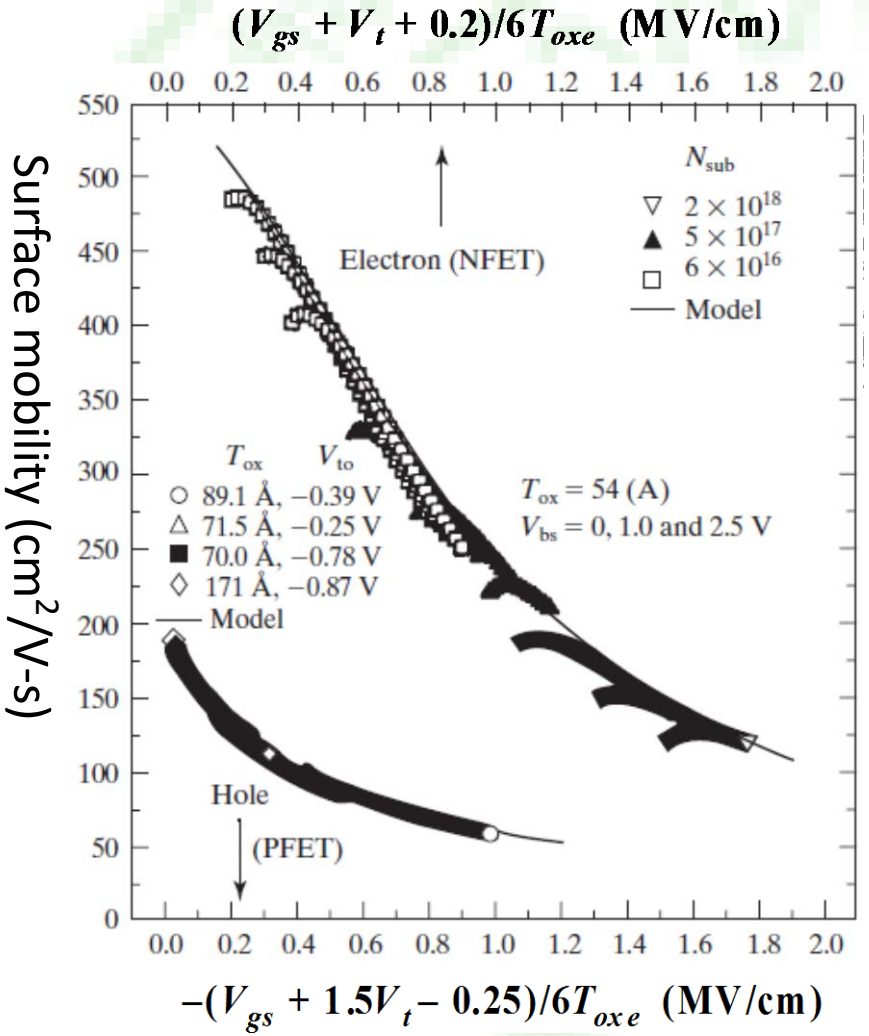
$$\therefore \frac{1}{2}(\mathcal{E}_b + \mathcal{E}_t) = \frac{C_{oxe}}{2\varepsilon_s}(V_{gs} + V_t - 2V_{fb} - 2\phi_{st})$$

$$\approx \frac{C_{oxe}}{2\varepsilon_s}(V_{gs} + V_t + 0.2\ \mathrm{V})$$

$$= \frac{V_{gs} + V_t + 0.2\ \mathrm{V}}{6T_{oxe}}$$

NMOS with n+ poly-Si gate
$V_{fb}$≈-0.5 and $\psi_{st}$≈0.4

# Universal Surface Mobilities




- Surface roughness scattering is stronger (mobility is lower) at higher $V_g$, higher $V_t$, and thinner $T_{oxe}$.

Empirical fitting

$$\mu_{ns} = \frac{540\ \mathrm{cm}^2/\mathrm{Vs}}{1 + \left(\frac{V_{gs} + V_t + 0.2\ \mathrm{V}}{5.4 T_{oxe}}\right)^{1.85}}$$

$$\mu_{ps} = \frac{185\ \mathrm{cm}^2/\mathrm{Vs}}{1 - \left(\frac{V_{gs} + 1.5V_t - 0.25\ \mathrm{V}}{3.38 T_{oxe}}\right)}$$

# Mobility Modeling in BSIM4

***mobMod*** = 0

$$\mu_{eff} = \frac{U0 \cdot f(L_{eff})}{1+\left(UA+UCV_{bseff}\right)\left(\frac{V_{gsteff}+2V_{th}}{TOXE}\right)+UB\left(\frac{V_{gsteff}+2V_{th}}{TOXE}\right)^2+UD\left(\frac{V_{th}\cdot TOXE}{V_{gsteff}+2\sqrt{V_{th}^{\ 2}+0.0001}}\right)^2} \tag{5.6}$$

***mobMod*** = 1

$$\mu_{eff} = \frac{U0 \cdot f(L_{eff})}{1+\left[UA\left(\frac{V_{gsteff}+2V_{th}}{TOXE}\right)+UB\left(\frac{V_{gsteff}+2V_{th}}{TOXE}\right)^2\right]\left(1+UC\cdot V_{bseff}\right)+UD\left(\frac{V_{th}\cdot TOXE}{V_{gsteff}+2\sqrt{V_{th}^{\ 2}+0.0001}}\right)^2} \tag{5.7}$$

$$E_{eff} = \frac{V_{gs}+V_t}{6T_{oxe}}$$

$$E_{eff} = \frac{V_{gs}-Vt+2V_t}{6T_{oxe}}$$

***mobMod*** = 2

$$\mu_{eff} = \frac{U0 \cdot f\left(L_{eff}\right)}{1+\left(UA+UC\cdot V_{bseff}\right)\left[\frac{V_{gsteff}+C_0\cdot\left(VTHO-VFB-\Phi_s\right)}{TOXE}\right]^{EU}+UD\left(\frac{V_{th}\cdot TOXE}{V_{gsteff}+2\sqrt{V_{th}^{\ 2}+0.0001}}\right)} \tag{5.8}$$

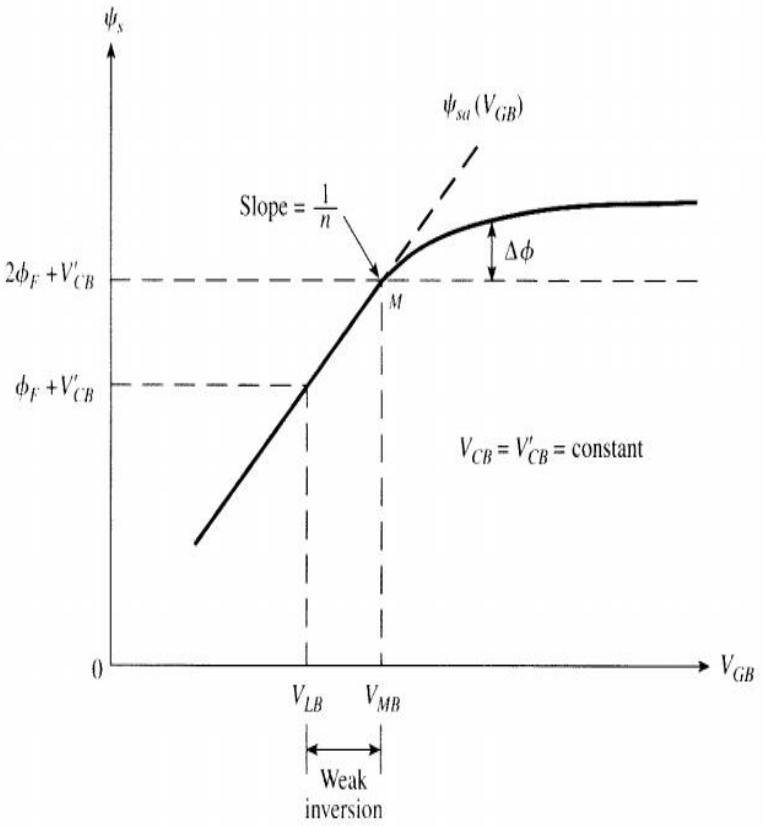
$$E_{eff} = \frac{V_{gsteff}+2V_t}{6T_{oxe}}$$

where the constant $C0 = 2$ for NMOS and 2.5 for PMOS.

$$f(L_{eff}) = 1-UP\cdot\exp\left(-\frac{L_{eff}}{LP}\right) \tag{5.9}$$

# Current in subthreshold region

- Subthreshold conduction
  - Transistor is in depletion
- Surface potential is determined by depletion under the gate, which is constant everywhere ($\psi_S \approx \psi_{sa}$).




$$\psi_{sa} \approx \psi_{sa} = \left( -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2$$

$$Q_I' = -\frac{\sqrt{2q\varepsilon_s N_A}}{2\sqrt{2\phi_f + V_{CB}'}} \left( \phi_t e^{\frac{V_{GC} - V_M}{n\phi_t}} \right)$$

$$I_{ds}(x) = \mu_{eff} W Q_i \frac{dV_{CB}}{dx}$$

# Current in subthreshold region

- Integrating from source to drain,

$$\int_0^L I_{ds}(x)dx = \int_0^L \mu_{eff} \cdot W \cdot Q_i \cdot dV_{CB}$$

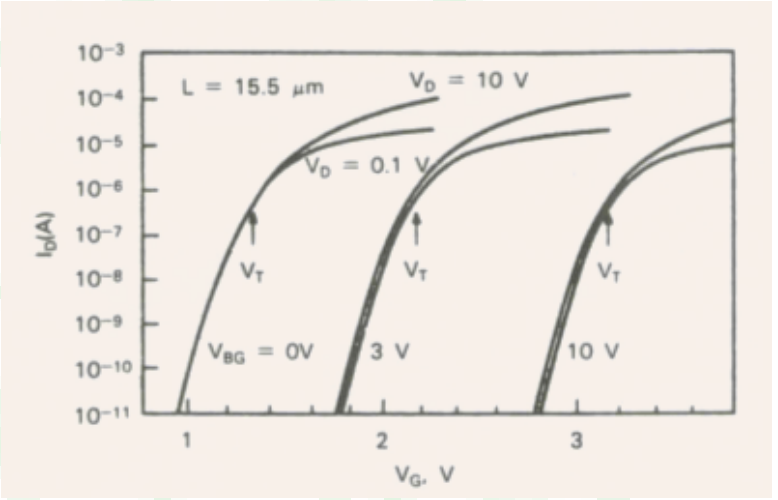
$$I_{ds} = \mu_{eff} \frac{W}{L} \int_{V_{SB}}^{V_{DB}} Q_i \cdot dV_{CB}$$

$$I_{ds} = \mu_{eff} \frac{W}{L} \int_{V_{SB}}^{V_{DB}} \left( \frac{\sqrt{2q\varepsilon_s N_A}}{2\sqrt{2\phi_f + V'_{CB}}} \phi_t \left( e^{\frac{V_{GB} - V_{CB} - V_{TH}}{n\phi_t}} \right) \right) \cdot dV_{CB}$$

$$I_{ds} = I_0 \left( e^{\frac{V_{GS} - V_{TH}}{n\varphi_t}} - e^{\frac{V_{GD} - V_{TH}}{n\varphi_t}} \right) \Rightarrow I_{ds} = I_0 e^{\frac{V_{GS} - V_{TH}}{n\varphi_t}} \left( 1 - e^{-\frac{V_{DS}}{n\varphi_t}} \right)$$

# Current in subthreshold region

$$I_{ds} = I_0 e^{\frac{V_{GS}-V_{TH}}{n\varphi_t}}\left(1-e^{-\frac{V_{DS}}{n\varphi_t}}\right)$$



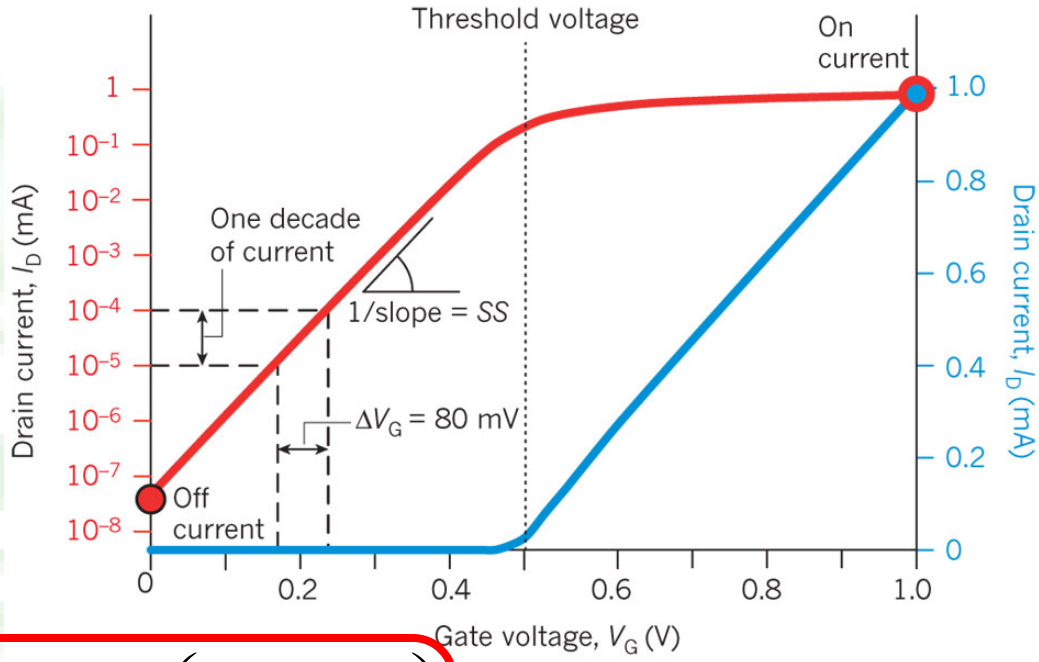

- Note – $V_{TH}$ is a function of body bias.
  - If $V_B$ increases in negative direction, $V_{TH}$ increases.

$$V_{TH} = V_{T0} + \gamma\left(\sqrt{\phi_0 + V_{SB}} - \sqrt{\phi_0}\right)$$

# Subthreshold slope

- It is defined as the amount of gate voltage required to change the gate current by 1-decade.

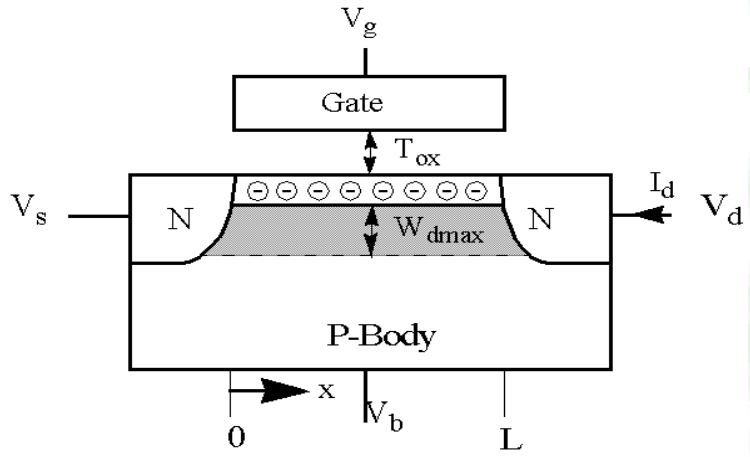
$$S = \frac{dV_{GS}}{d(\log I_{ds})}$$




**At room temperature**

$$S = (25.85)(2.30)\left(1+\frac{C_{dep}}{C_{ox}}\right) \approx 60mV\left(1+\frac{C_{dep}}{C_{ox}}\right)$$

# Drain Current and $Q_{inv}$ in MOSFET




- Channel voltage
  $V_c = V_s$ at $x = 0$ and
  $V_c = V_d$ at $x = L$.

$$Q_I' = -C_{ox}'\left(V_{GC} - V_{TH}\right) = -C_{ox}'\left(V_{GC} - V_{t0} - \alpha V_{cb}\right)$$

- $Q_{inv} = -C_{ox}(V_{gs} - V_{cs} - V_{t0} - \alpha\,(V_{sb} + V_{cs})$
- $= -C_{ox}(V_{gs} - V_{cs} - (V_{t0} + \alpha\, V_{sb}) - \alpha\, V_{cs})$
- $\boldsymbol{Q_{inv} = -C_{ox}(V_{gs} - mV_{cs} - V_t)}$
- 
- $m \equiv 1 + \alpha = 1 + 3T_{oxe}/W_{dmax} \approx 1.2$
- m is called the body-effect factor or bulk-charge factor.

# Drain Current Calculation

Now,

$$I_{ds} = WC_{oxe}(V_{gs} - mV_{cs} - V_t)\mu_{ns} dV_{cs}/dx$$

Integrating the above equation over the channel length L, gives the current voltage relation as follows:

$$\int_0^L I_{ds}\,dx = WC_{oxe}\mu_{ns}\int_0^{V_{ds}} (V_{gs} - mV_{cs} - V_t)\,dV_{cs}$$

$$I_{ds}L = WC_{oxe}\mu_{ns}(V_{gs} - V_t - mV_{ds}/2)V_{ds}$$

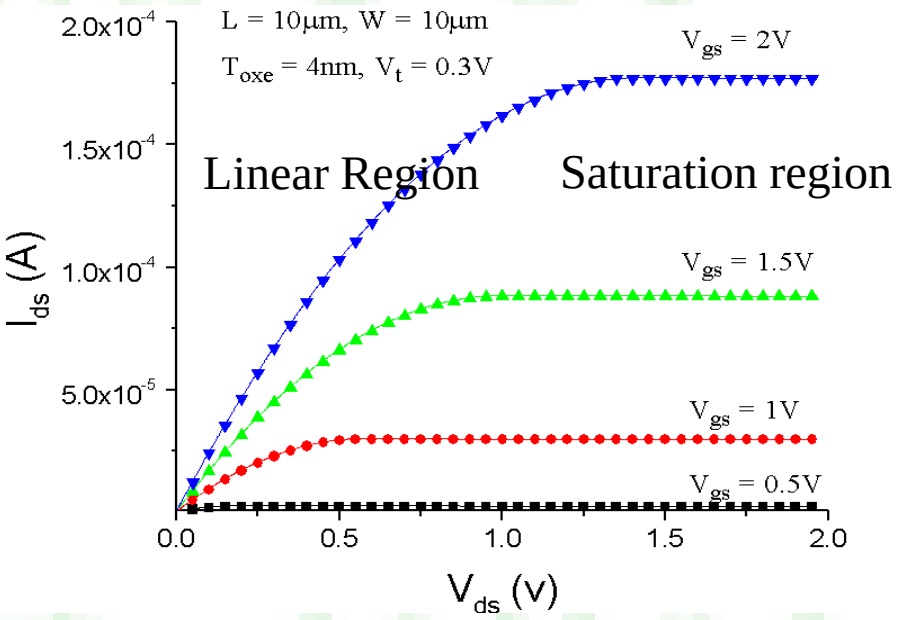
$$\boxed{I_{ds} = \mu_{ns} C_{oxe} \frac{W}{L}\left(V_{gs} - V_t - \frac{m}{2}V_{ds}\right)V_{ds}}$$

# I-V characteristics

$$\frac{dI_{ds}}{dV_{ds}} = 0$$

$$\frac{W}{L} C_{oxe} \mu_{ns} (V_{gs} - V_t - mV_{dsat}) = 0$$

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




$V_{ds} < V_{dsat}$ ➡ Linear Region

$V_{ds} \geq V_{dsat}$ ➡ Saturation region

Drain current in saturation regio

$$I_{dsat} = \frac{W}{2mL} C_{oxe} \mu_{ns} (V_{gs} - V_t)^2$$

- transconductance: $g_m = dI_{ds}/dV_{gs}$ $g_{msat} = \frac{W}{mL} C_{oxe} \mu_{ns} (V_{gs} - V_t)$

# I-V characteristics

What happens at $V_{ds}=V_{dsat}$ & why $I_{ds}$ remains constant beyond $V_{dsat}$


- At $V_{ds}=V_{dsat}$ , $Q_{inv}$ near the drain end of the channel becomes zero ! i.e. Pinch off.

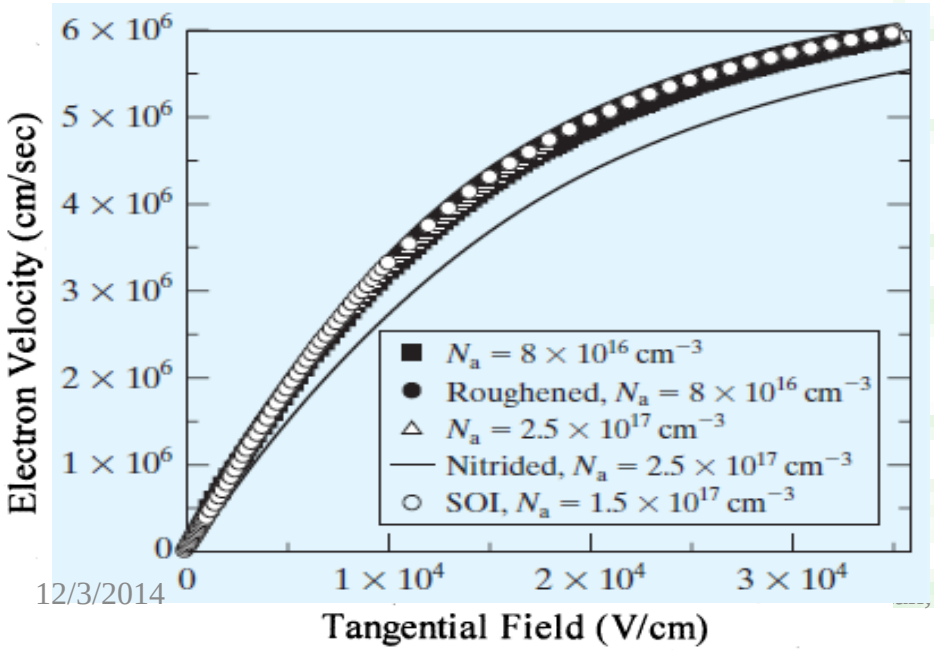
$$I_{ds} = WQ_{inv}\mu_{ns}E \text{ (Large } E \text{ and and negligible } Q_{inv})$$

- At $V_{ds} > V_{dsat}$ , A very short region near the drain end where the $Q_{inv} = 0$, a very high electric field exist due to the drop of the additional $V_{ds} - V_{dsat}$ .

# Velocity Saturation

- At low $E$ $\Rightarrow$ $v = \mu_{ns} E$
- The inversion-layer electron velocity saturates at high field

$$v = \frac{\mu_{ns}\xi}{1 + \xi/\xi_{sat}}$$




$$v = \mu_{ns}\xi \quad , \quad \xi << \xi_{sat}$$

$$v = v_{sat} = \mu_{ns}\xi_{sat} \quad , \quad \xi \geq \xi_{sat}$$

$\xi_{sat}$ is the field at which velocity saturation becomes dominant.

# Velocity Saturation and I-V Model

$$I_{ds} = WQ_{inv}v \qquad v = \frac{\mu_{ns}\xi}{1+\xi/\xi_{sat}}$$

$$I_{ds} = WC_{oxe}\left(V_{gs} - mV_{cs} - V_t\right)\frac{\mu_{ns}\,dV_{cs}/dx}{1+\dfrac{dV_{cs}/dx}{\xi_{sat}}}$$

$$\int_0^L I_{ds}\,dx = \int_0^{V_{ds}} \left[WC_{oxe}\mu_{ns}(V_{gs} - mV_{cs} - V_t) - I_{ds}/\xi_{sat}\right]dV_{cs}$$

**Drain current when $v < v_{sat}$**

$$I_{ds} = \frac{\dfrac{W}{L}C_{oxe}\mu_{ns}\left(V_{gs} - \dfrac{m}{2}V_{ds} - V_t\right)V_{ds}}{1+\dfrac{V_{ds}}{L\xi_{sat}}}$$

# Velocity Saturation and I-V Model

- If L is large then $\frac{V_{ds}}{L\xi_{sat}}$ will be negligible, then:

$$I_{ds} = \frac{W}{L} C_{oxe} \mu_s \left(V_{gs} - V_t - \frac{m}{2} V_{ds}\right) V_{ds}$$

It is called the long channel I-V model

- Effect of *velocity saturation* on $I_{ds}$:

$$I_{ds} = (\text{Long channel } I_{ds}) / \left(1 + \frac{V_{ds}}{L\xi_{sat}}\right)$$

- In short channel devices  $1 + \frac{V_{ds}}{L\xi_{sat}} > 1$

# Velocity Saturation and I-V Model

- Drain current for $V_{ds} \geq V_{dsat}$

$$I_{dsat} = \frac{W}{2mL} C_{oxe} \mu_{ns} \frac{(V_{gs} - V_t)^2}{1 + \frac{V_{gs} - V_t}{m\xi_{sat} L}} = \text{Long channel } I_{dsat} / \left(1 + \frac{V_{gs} - V_t}{m\xi_{sat} L}\right)$$

***Very short channel case:***

$$E_{sat} L << V_{gs} - V_t$$

$$I_{dsat} = W v_{sat} C_{oxe} \left(V_{gs} - V_t - m E_{sat} L\right)$$

$$I_{dsat} \approx W v_{sat} C_{oxe} \left(V_{gs} - V_t\right)$$

- $I_{dsat}$ is proportional to $V_{gs} - V_t$ rather than $(V_{gs} - V_t)^2$,
- Not as sensitive to $L$ as than long channel case ($\propto 1/L$).

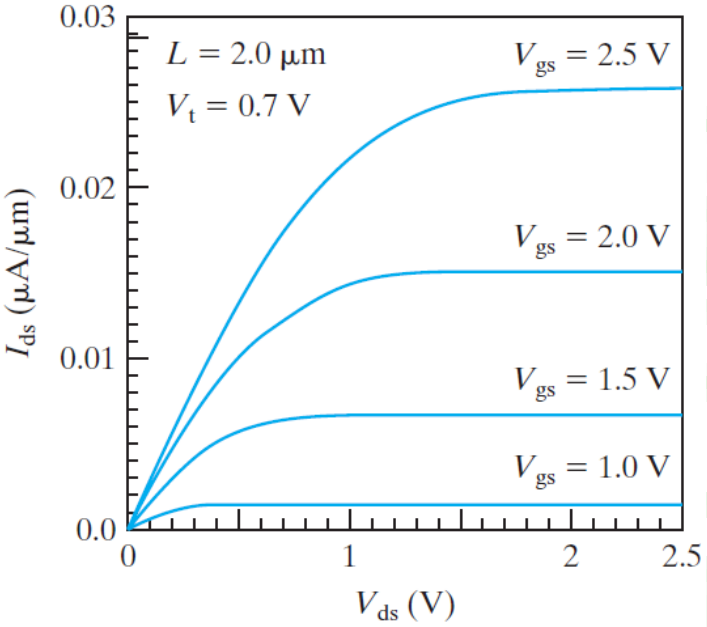
***Long channel case:***

$$E_{sat} L >> V_{gs} - V_t$$

$$I_{dsat} \approx \frac{W}{2mL} C_{oxe} \mu_{ns} \left(V_{gs} - V_t\right)^2$$

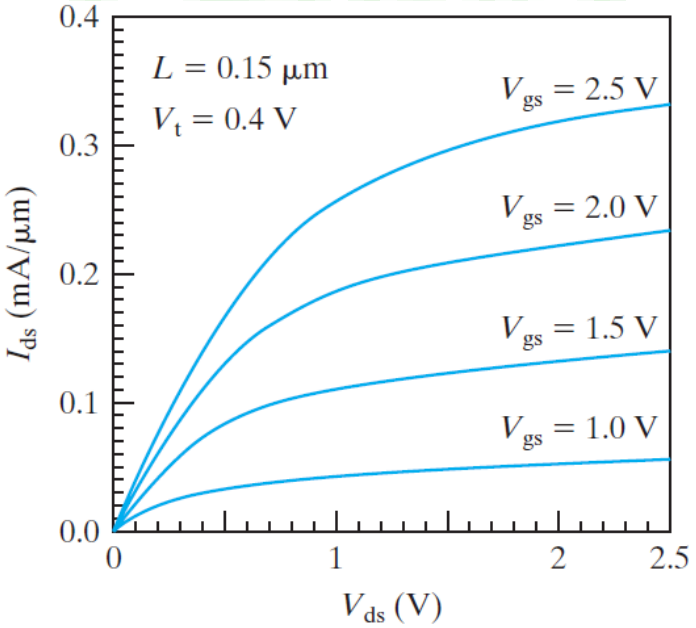
# I-V Characteristics

- Long Channel




- $I_{ds} \propto (V_{gs}-V_t)^2$

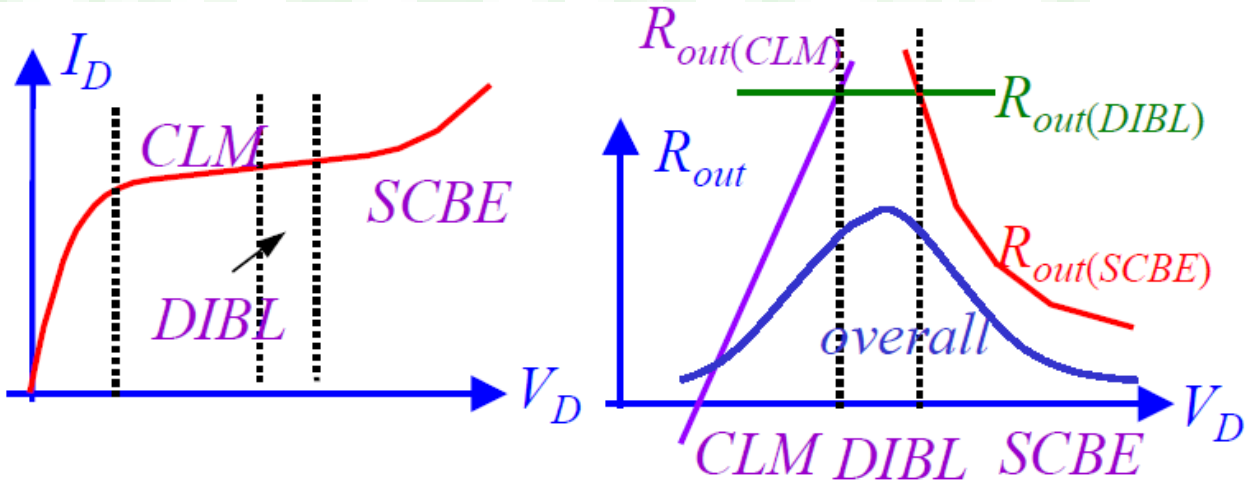
Short Channel



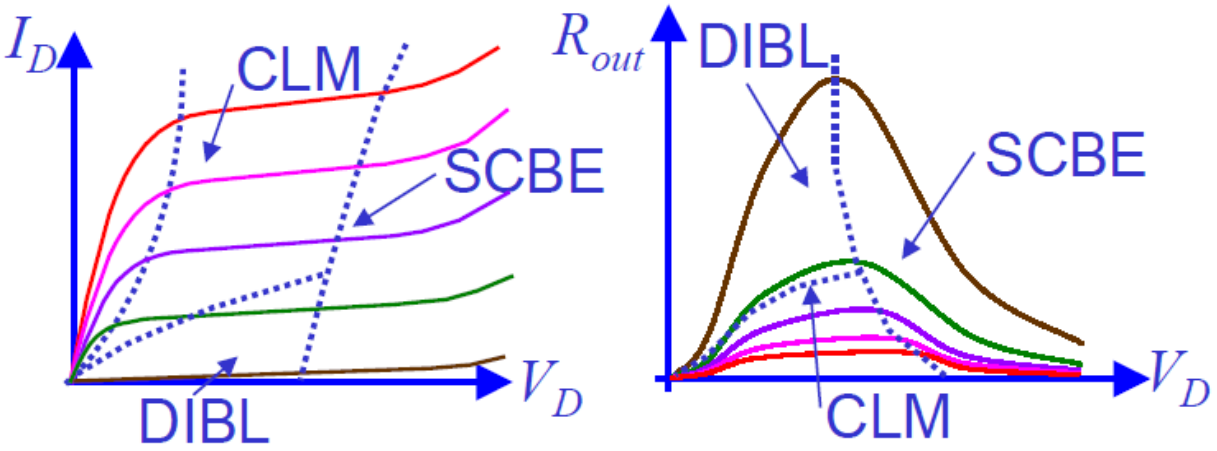

$I_{ds} \propto V_{gs}-V_t$ DITS +CLM+ DIBL

# I-V Characteristics

Curves for a particular gate voltage




Curves for a different gate voltages



# Compact Modeling is Art based on Science

# Smoothing function and I-V Model

- Smoothing function is required for a smooth transition between two functions.
  - This stems from the need to have a single equation valid in all regions of operation.
- BSIM3 introduced use of smoothing functions to get single equation valid in all regions of biases.
  - This gave continuous and smooth I-V and C-V making it popular model for analog design.

# Linear to Saturation transition

- First generation SPICE models used this kind of equation,

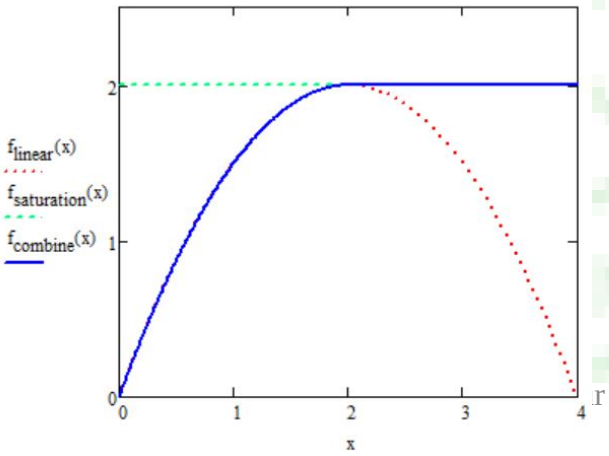
$$I_D = \begin{cases} \frac{W}{L}\mu C'_{ox}\left[(V_{GS}-V_T)V_{DS}-\frac{m}{2}V_{DS}^2\right], V_{DS} < V_{DS,sat} \\ \frac{W}{L}\mu C'_{ox}\left[\frac{(V_{GS}-V_T)^2}{2m}\right], V_{DS} \geq V_{DS,sat} \end{cases}$$

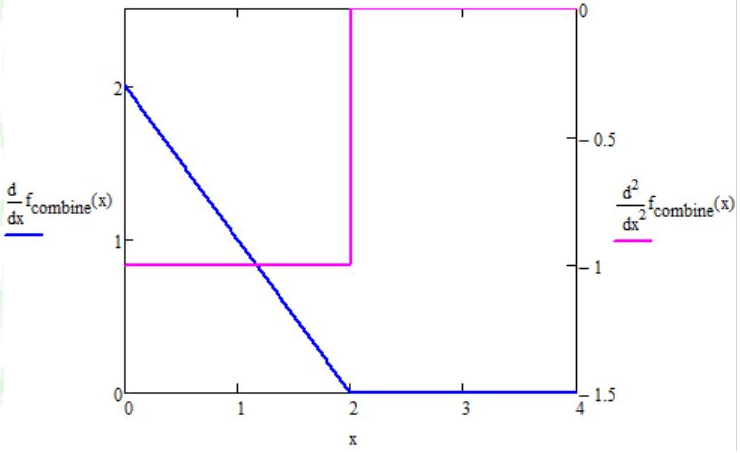
$I_D$ and $\frac{dI_D}{dV_{DS}}$ are continuous at $V_{DS} = V_{DS,sat}$ but $\frac{d^2 I_D}{dV_{DS}^2}$ is not.

$f_{linear}(x) := 2\cdot x - \frac{1}{2}\cdot x^2$

$f_{saturation}(x) := 2$

$f_{combine}(x) := \begin{cases} 2\cdot x - \frac{1}{2}\cdot x^2 & \text{if } x < 2 \\ 2 & \text{otherwise} \end{cases}$



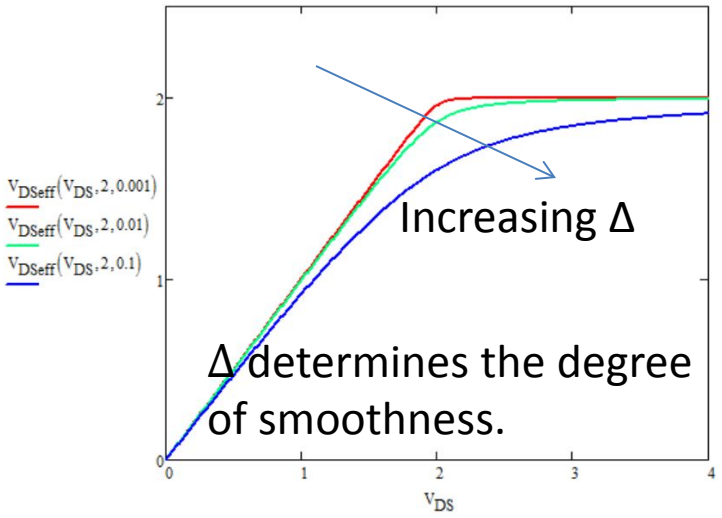
# Linear to Saturation transition

- For numerical robustness, the derivatives of arbitrary order must be continuous at all voltage values of interest. This property is sometimes referred to as ∞-differentiability.
- **Single equation approach used in BSIM3**. Define an effective drain-source bias $V_{DSeff}$,

$$V_{DS,eff} = V_{DS,sat} - \frac{1}{2}\left(V_{DS,sat} - V_{DS} - \Delta + \sqrt{(V_{DS,sat} - V_{DS} - \Delta)^2 + 4\Delta V_{DS,sat}}\right)$$

- $V_{DS} \ll V_{DS,sat}$, $V_{DS,eff} \approx V_{DS}$
- For $V_{DS} \gg V_{DS,sat}$, $V_{DS,eff} \approx V_{DS,sat}$

- Drain current equation becomes ($V_{GS} > V_T$),
- $I_D = \frac{W}{L}\mu C'_{ox}\left[(V_{GS} - V_T)V_{DS,eff} - \frac{m}{2}V_{DS,eff}^2\right]$
- Derivatives are continuous.



# Sub-threshold to strong inversion transition

- For $V_{GS} \ll V_T$,

$$I_D = I_0 e^{\frac{(V_{GS}-V_T-V_{off})}{nkT/q}} \left[1 - e^{-\frac{V_{DS}}{kT/q}}\right]$$

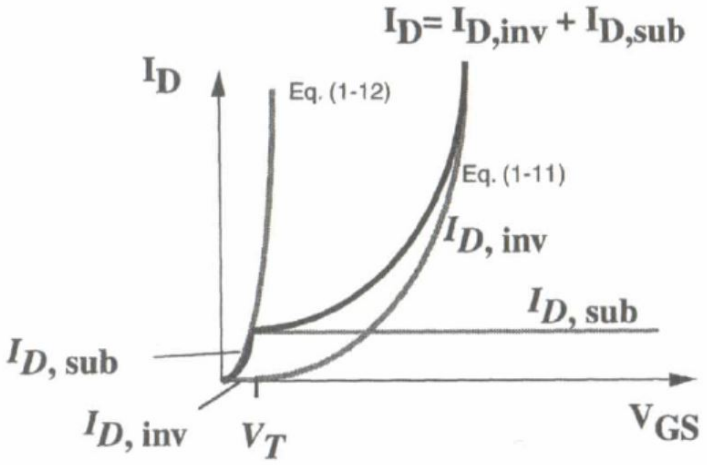
  - This is not valid in strong inversion. It leads to excessively high current for $V_{GS} \gg V_T$
- For $V_{GS} \gg V_T$

$$I_D = \frac{W}{L}\mu C'_{ox}\left[(V_{GS} - V_T)V_{DS,eff} - \frac{m}{2}V_{DS,eff}^2\right]$$

  - This is not valid in sub-threshold and leads to negative current for $V_{GS} < V_T$
- First method – Single equation:

$$I_D = I_{D,sub} + I_{D,inv}$$

- Good enough for Digital applications, but the derivatives are discontinuous making it unsuitable for Analog cases.

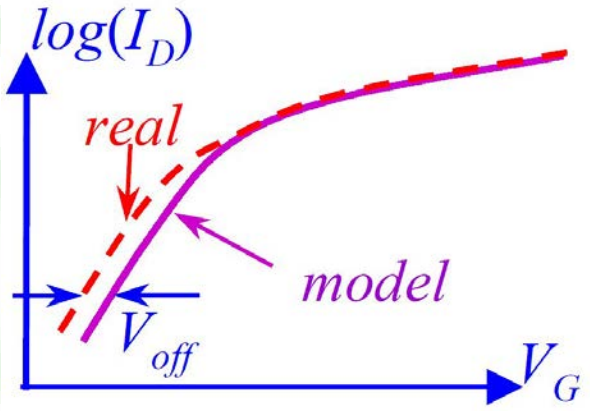


Current not drawn to scale

# Sub-threshold to strong inversion transition

- Second method – Single equation: Use effective $V_{GS} - V_T$ as

$$V_{GST,eff} = \frac{\frac{2nkT}{q} ln\left[1 + \exp\left(\frac{V_{GS} - V_T}{2nkT/q}\right)\right]}{1 + 2n\left[exp\left(-\frac{V_{GS} - V_T - 2V_{off}}{2nkT/q}\right)\right]}$$




- $n$ is the ideality factor and lies between 1 and 2.
- $V_{off}$ is a parameter for fringing from width side. Assume $V_{off}$=0 for further analysis.
- For $V_{GS} \gg V_T$, the exponential term inside ln() is larger than 1 making $V_{GST,eff} = V_{GS} - V_T$
- For $V_{GS} \ll V_T$,

$$V_{GST,eff} \approx \frac{kT}{q} exp\left(\frac{V_{GS} - V_T}{nkT/q}\right)$$

# Single equation for drain current

- Use $V_{DS,sat} = \frac{V_{GST,eff}+2kT/q}{m}$, where $2kT/q$ is added for numerical stability when $V_{GST,eff} \ll 2kT/q$.
- We have written ID as follows valid from linear to saturation-

$$I_D = \frac{W}{L}\mu C'_{ox}\left[(V_{GS}-V_T)V_{DS,eff} - \frac{m}{2}V_{DS,eff}^2\right]$$

$$I_D = \frac{W}{L}\mu C'_{ox}\left[V_{GST,eff} - \frac{m}{2}V_{DS,eff}\right]V_{DS,eff}$$

- Now drain current becomes,

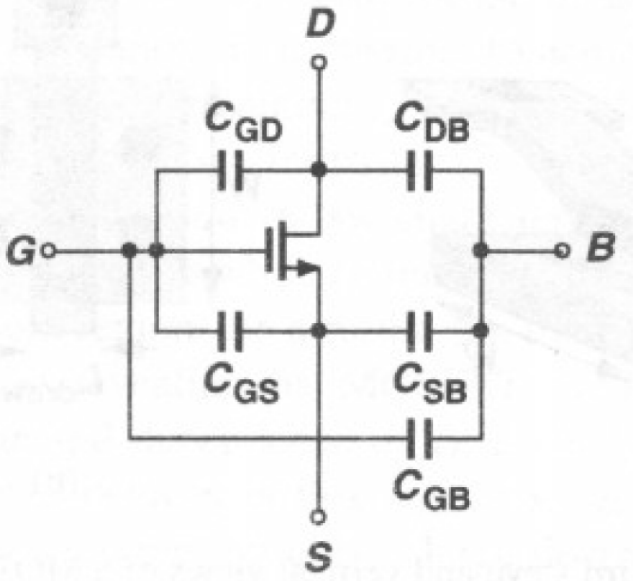
$$I_D = \frac{W}{L}\mu C'_{ox}\left[V_{GST,eff} - \frac{m}{2}V_{DS,eff}\frac{V_{GST,eff}}{V_{GST,eff}+2kT/q}\right]V_{DS,eff}$$

$$I_D = \frac{W}{L}\mu C'_{ox}V_{GST,eff}\left[1 - \frac{m}{2}\frac{V_{DS,eff}}{V_{GST,eff}+2kT/q}\right]V_{DS,eff}$$

- This is valid for all $V_{gs}$ and $V_{ds}$.

# Terminal Charges and Charge Partition

- AC and Transient simulation need capacitances.
- Quasi-static approximation
  - The Channel charge is assumed to respond instantaneously to any change in the bias voltage.
- From $Q_i'$, we need to find $Q_G$, $Q_B$, $Q_S$ and $Q_D$.



# Total inversion charge

- Before delving into $Q_G$, $Q_S$ and $Q_D$. Let us find the total inversion charge in the channel, $Q_i$.
- The charge per unit area, $Q_I'$, is given as

$$Q_I' = -C_{ox}'(V_{GS} - V_T - mV_{CS}(x))$$

- We need to know $V_{CS}(x) = ?$

- Define $\alpha = 1 - \frac{V_{DS,eff}}{V_{DS,sat}}$ which gives, $\frac{V_{DS,eff}}{V_{DS,sat}} = 1 - \alpha$, Thus

$$\boldsymbol{V_{CS} = \frac{V_{GST,eff}}{m}\left[1 - \sqrt{1 - \frac{x}{L}(1-\alpha)^2}\right]}$$

# Total inversion charge

- The charge per unit area, $Q_I'$, is given as

$$Q_I' = -C_{ox}'(V_{GS} - V_T - mV_{CS}(x))$$

$$Q_I' = -C_{ox}' V_{GST,eff} \sqrt{1 - \frac{x}{L}(1-\alpha)^2}$$

- Thus total inversion charge

$$Q_I = -WC_{ox}' \int_0^L V_{GST,eff} \sqrt{1 - \frac{x}{L}(1-\alpha^2)}\, dx$$

- Total inversion charge

$$Q_I = -\frac{2}{3} WLC_{ox}' V_{GST,eff} \frac{1+\alpha+\alpha^2}{1+\alpha}$$

# Source-Drain charge partitioning

- We know, $Q_I = Q_S + Q_D$ but not the exact share of each.
- The assignment of the channel charge to the source and drain charges is called *charge partition*.
- **Charge partitioning**
  - *50/50 partition:* Arbitrarily assign 50% of $Q_I$ to $Q_S$ and 50% to $Q_D$.This is valid only when $V_{DS}$ is small. For $V_{DS}\sim 0$, MOSFET is symmetrical and $Q_S \sim Q_D \sim Q_I/2$.
  - *0/100 partition:* This is based on the logic that in saturation, the pinch off region implies that $Q_D = 0$, which is actually not correct.
  - *40/60 partition:* This is a more physical distribution of charges. One should note that the charge distribution under this scheme is 40/60 only in saturation. However this partition scheme is valid in all regions.

# Source-Drain charge

- Ward-Dutton partitioning scheme

$$Q_D = W\int_0^L \frac{x}{L} Q_I' dx,\ Q_S = W\int_0^L \left(1-\frac{x}{L}\right) Q_I' dx$$

- **Drain charge**

$$Q_D = W\int_0^L \frac{x}{L} Q_I' dx = -WC_{ox}' V_{GST,eff}\int_0^L \frac{x}{L}\sqrt{1-\frac{x}{L}(1-\alpha^2)}\,dx$$

$$Q_D = -\frac{2}{15} WLC_{ox}' V_{GST,eff}\frac{3\alpha^3+6\alpha^2+4\alpha+2}{(1+\alpha)^2}$$

- Similarly,

$$Q_S = W\int_0^L \left(1-\frac{x}{L}\right) Q_I' dx = -\frac{2}{15} WLC_{ox}' V_{GST,eff}\frac{2\alpha^3+4\alpha^2+6\alpha+3}{(1+\alpha)^2}$$

Ref.: S.-Y. OH, D. E. WARD and, A. W. DUTTON, "Transient Analysis of MOS Transistors," IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. SC-15, NO. 4, AUGUST 1980.

# Gate and Bulk charge

- Total gate charge

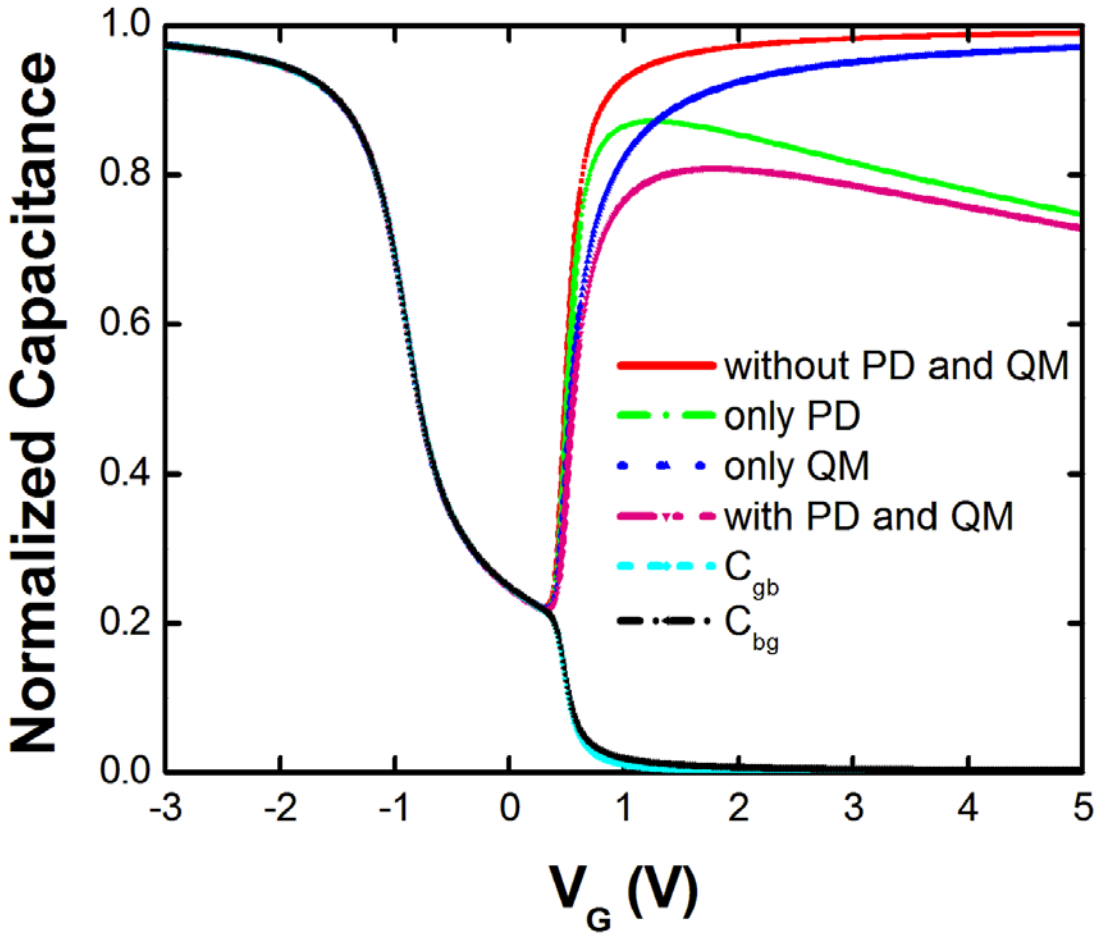
$$Q_G = WLC'_{ox}\left[\frac{V_{GST,eff}}{m}\left\{(m-1)+\frac{2}{3}\cdot\frac{1+\alpha+\alpha^2}{1+\alpha}\right\}\right.$$
$$\left.+2(m-1)(2\phi_F - V_{BS})\right]$$

- Finally Bulk charge,

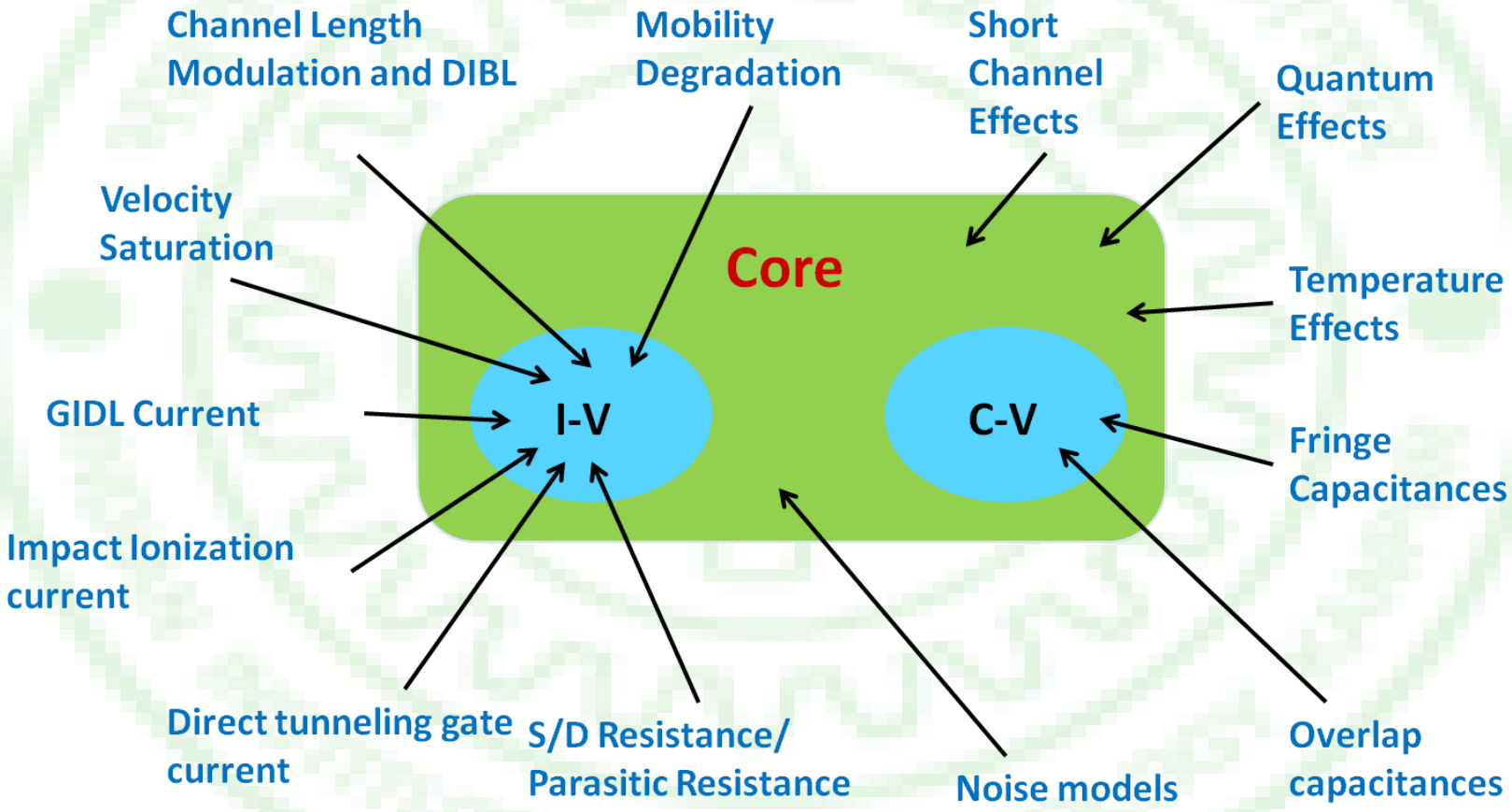
$$Q_B = -(Q_G - Q_I)$$

- Capacitance $C_{mn} = \frac{\partial Q_m}{\partial V_n}$

# Charge and Capacitance plots

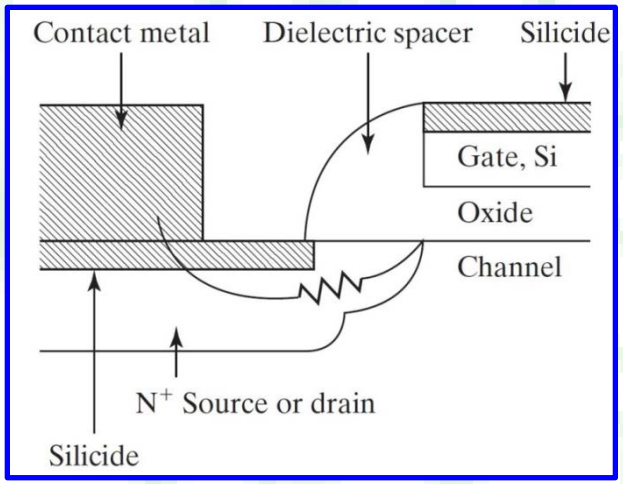


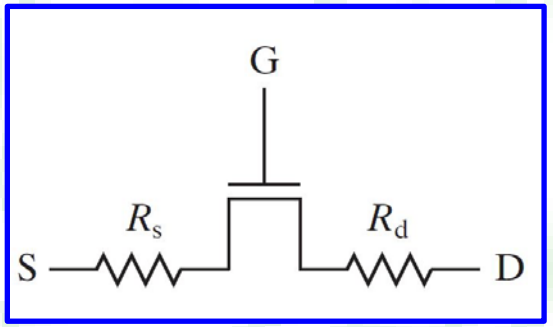
# Real Device Effects



# Parasitic Source-Drain Resistance

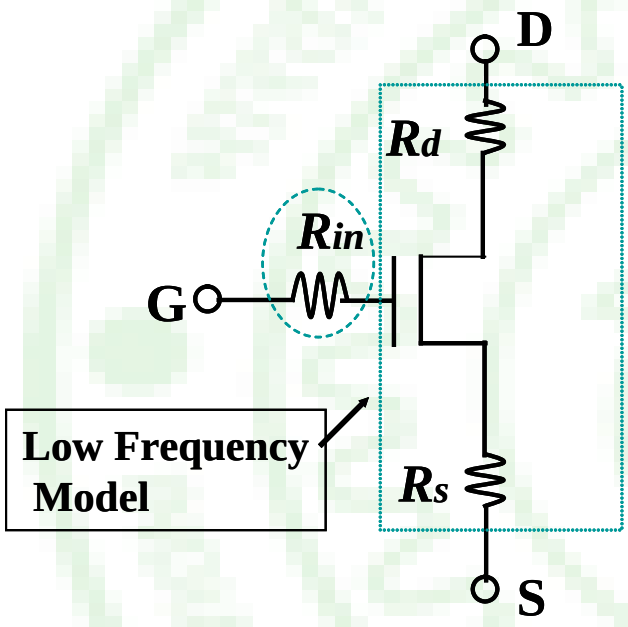
- The main effect of the parasitic resistance is that $V_{gs}$ in the $I_{ds}$ equations is reduced by $R_s \cdot I_{ds}$




$$I_{dsat} \approx \frac{I_{dsat0}}{1 + \frac{I_{dsat0} R_s}{(V_{gs} - V_t)}}$$

# High-Frequency performance



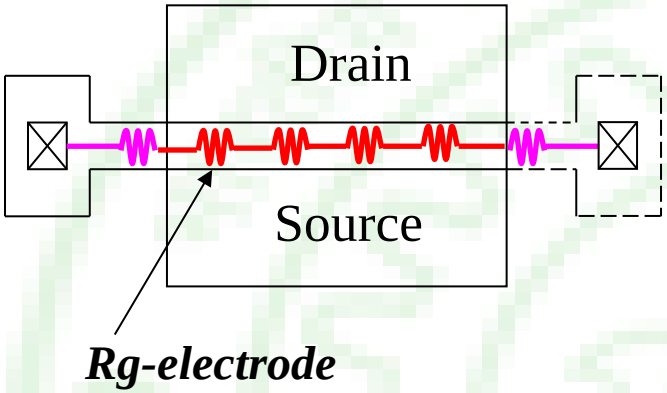

High-frequency performance is limited by input R and/or C.

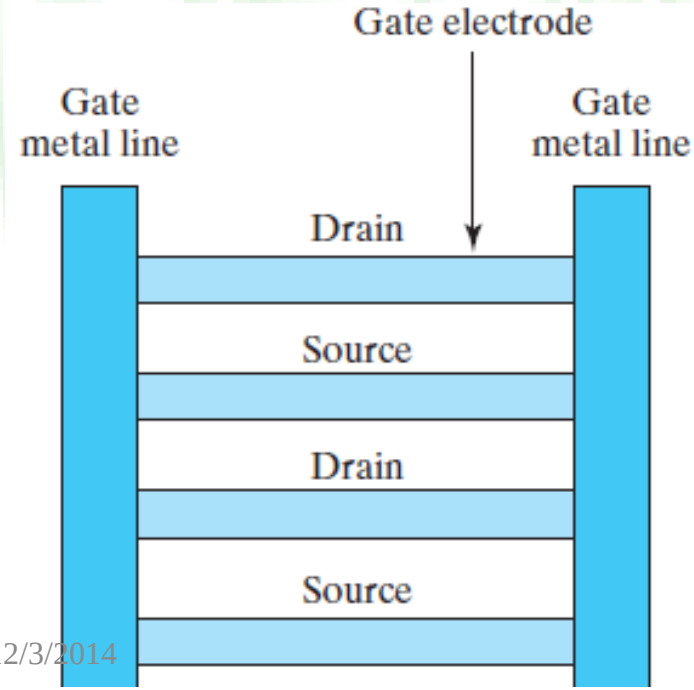
$$R_{in} = R_{g-electrode} + R_{ii}$$

**Gate-electrode resistance**

**Intrinsic input resistance**

# Gate-Electrode Resistance




**Multi-finger layout** greatly reduces the gate electrode resistance

$$R_{g-electrode} = \rho W / 12 T_g L_g N_f^2$$

$\rho$ : resistivity of gate material,
$W_f$ : width of each gate finger,
$T_g$ : gate thickness,
$L_g$ : gate length,
$N_f$ : number of fingers.

# Bulk MOSFET

- Drain current in MOSFET (ON operation)

$$I_{ON} = \mu \frac{W}{L} C_{ox} (V_{DD} - V_{TH})^2$$

- Drain current in MOSFET (OFF operation)

$$I_{OFF} \propto 10^{\left(\frac{V_{GS} - V_{TH}}{S}\right)}$$

$C_{ox}=\varepsilon_{ox}/t_{ox}$=oxide cap.
S – Subthreshold slope

- **Desired**
- **High $I_{ON}$ (↓L, ↑Cox, ↑$V_{DD}$-$V_{TH}$)**
- **Low $I_{OFF}$ (↑$V_{TH}$, ↑S)**

# Technology Scaling

- Each time the minimum line width is reduced, we say that a new ***technology node*** is introduced.
- Example: 90 nm, 65 nm, 45 nm
  - Numbers refer to the minimum metal line width.
  - Poly-Si gate length may be even smaller.

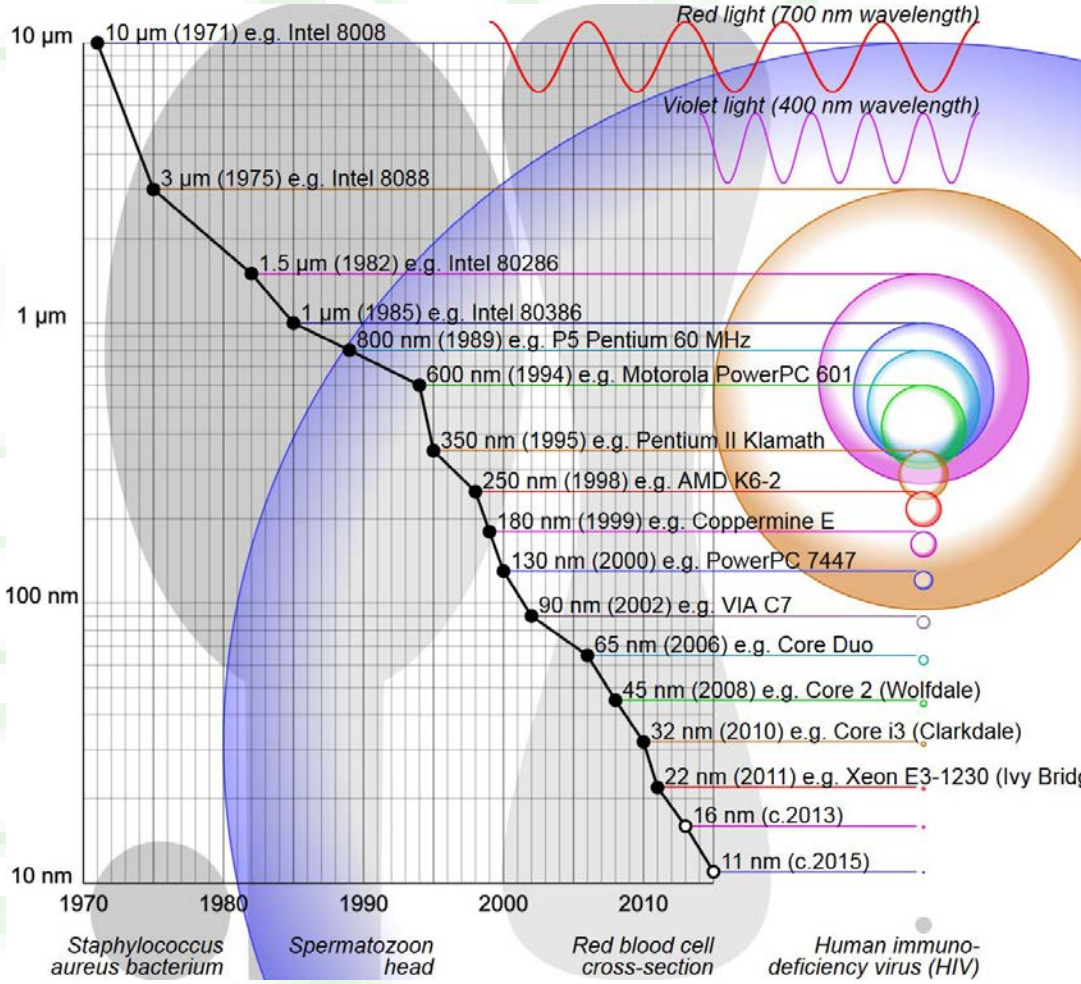



Figure source - Wikipedia

# Technology Scaling

- Scaling – At each new node, all geometrical features are reduced in size to 70% of the previous node.
- Reward – Reduction of *circuit size by half*. (~50% reduction in area, i.e., $0.7 \times 0.7 = 0.49$.)
  - Twice number of circuits on each wafer
  - Cost per circuit is reduced significantly.

- Ultimately – **Scaling drives down the cost of ICs.**

# Scaling and Moore's Law

- Number of components per IC function will double every two years – April 19, 1965 (Electronics Magazine)
- Shorthand for rapid technological change!

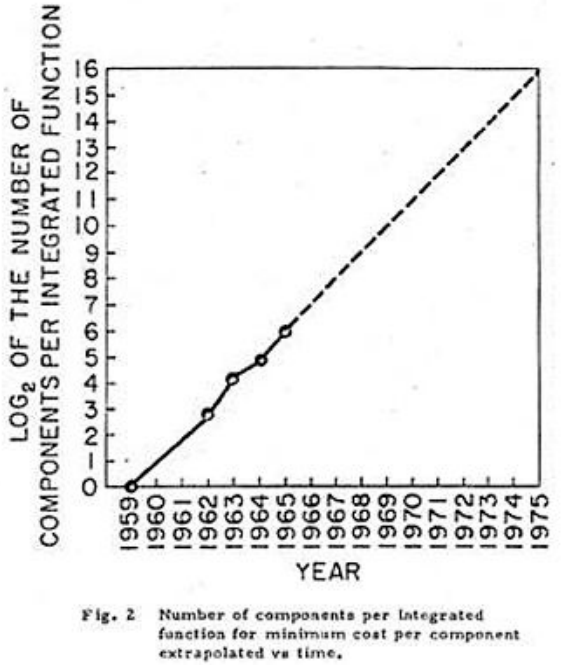
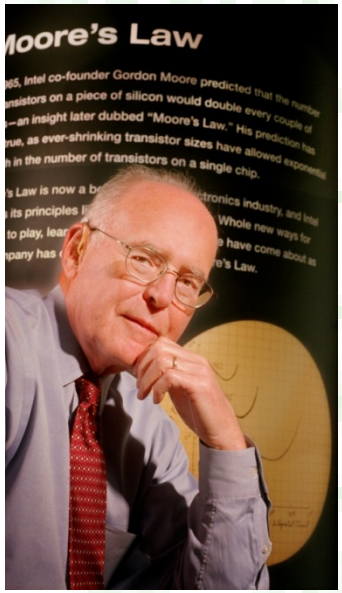
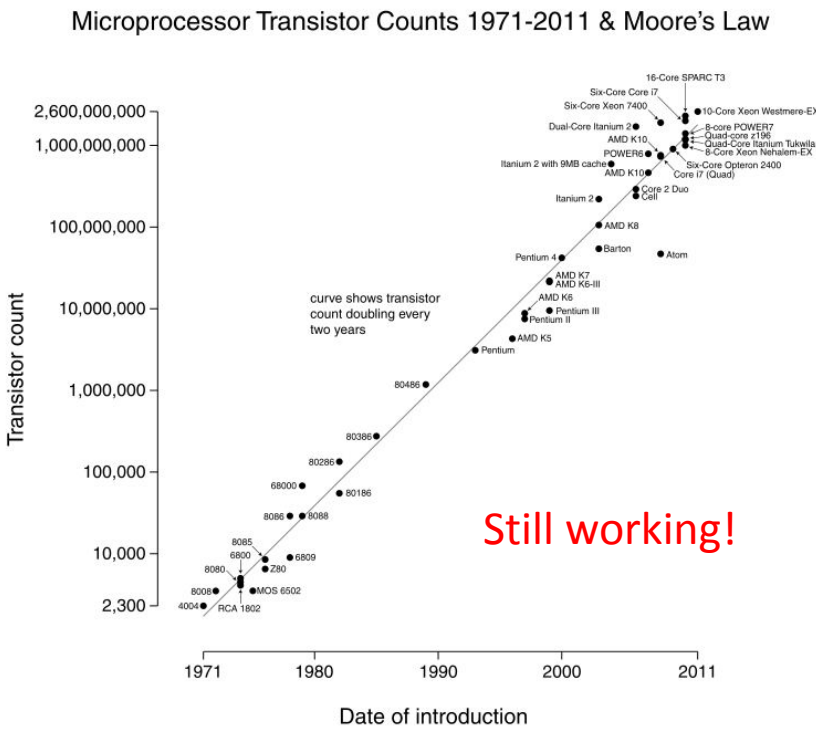


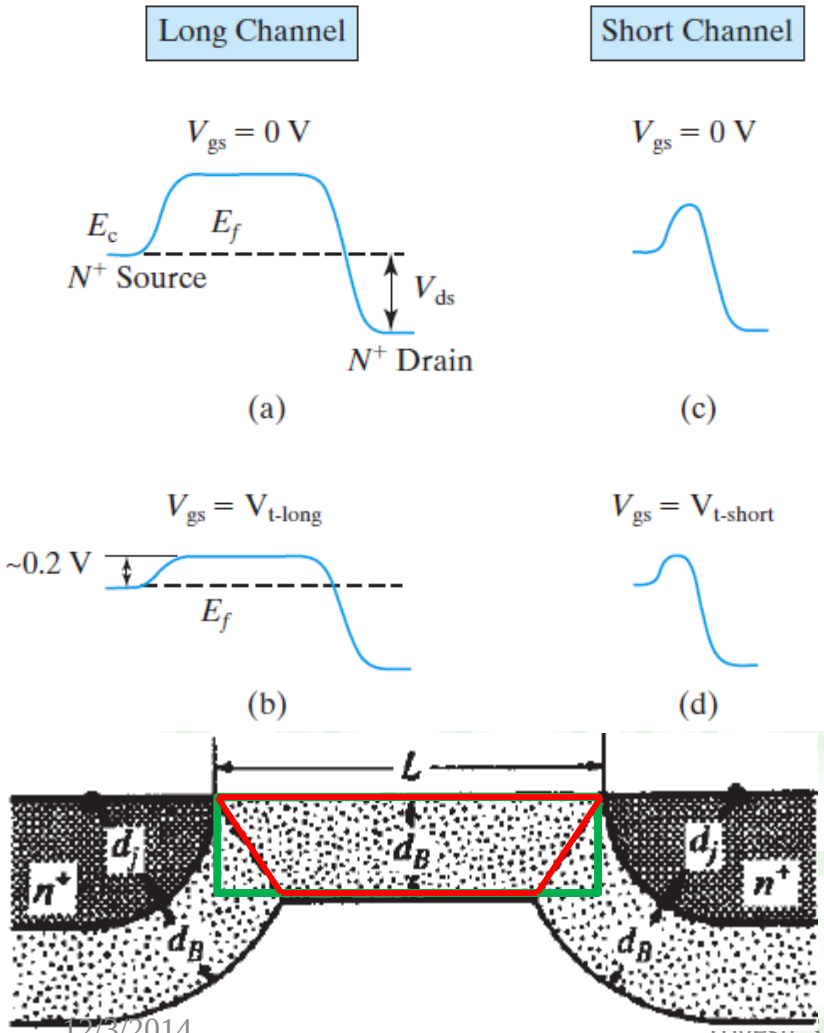

Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.

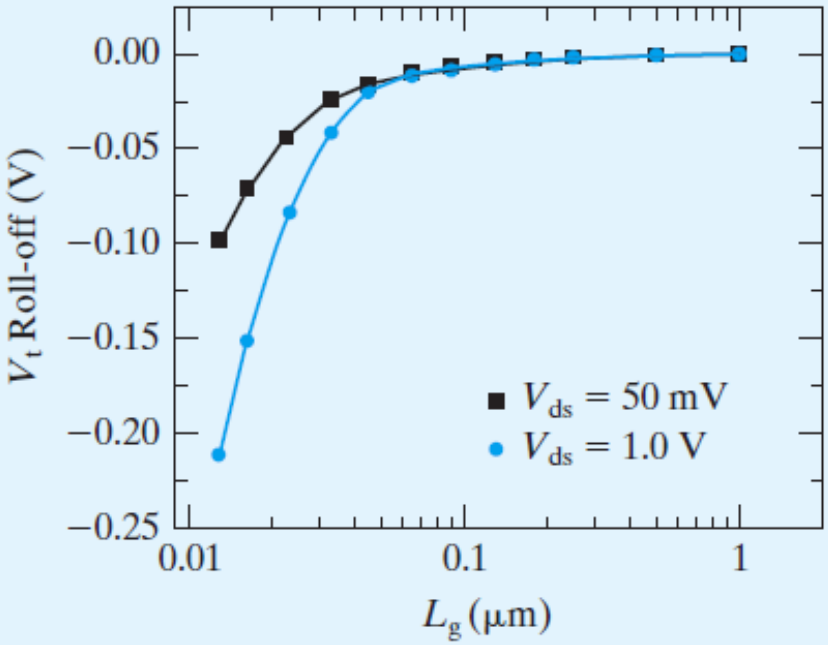



Source: http://www.intel.com/pressroom/kits/events/moores_law_40th/

# Threshold Voltage Roll-Off

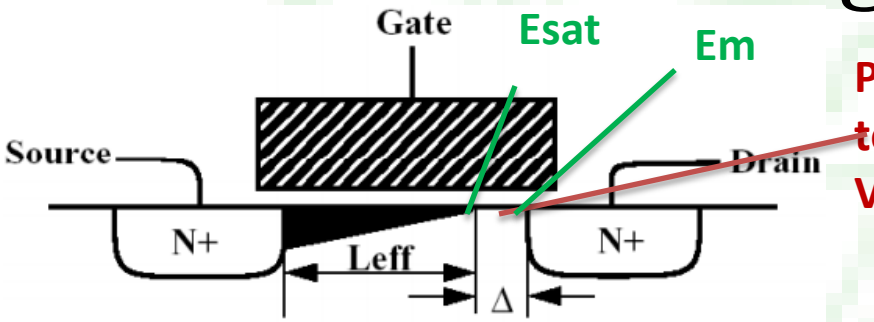



Vt decreases at very small Lg. It determines the minimum acceptable Lg because Ioff is too large when Vt becomes too low or too sensitive to Lg.

Source: Chenming Hu – Modern Semiconductor Devices for Integrated Circuits

# Channel Length Modulation

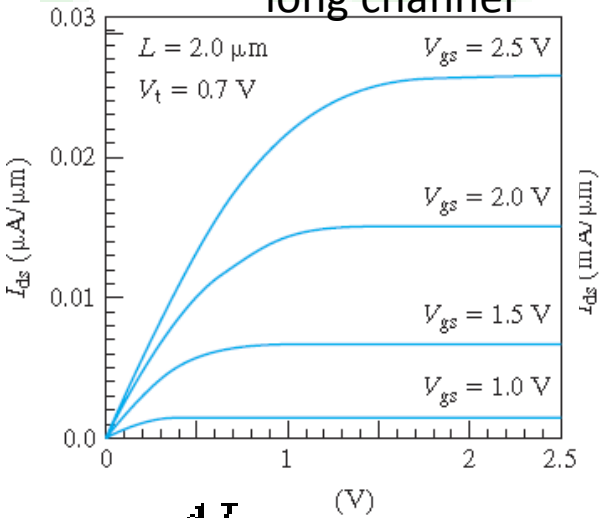



Pinch off point moves towards the source as $V_{ds}$ increases

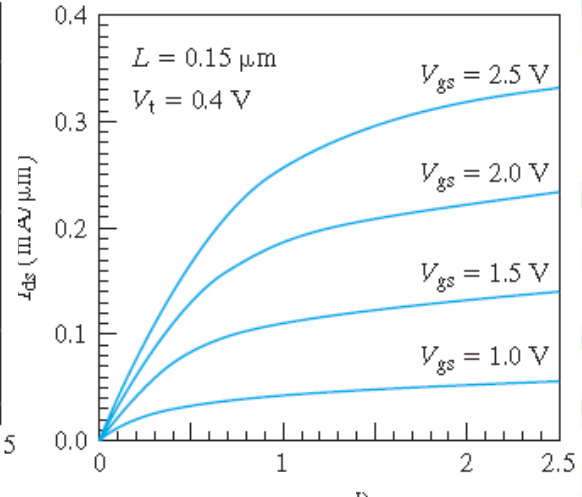
$$\Delta \propto V_{DS} - V_{DSsat}$$

$$I_{Dsat} = I_{Dsat0}\left(1+\frac{\Delta L}{L}\right) = I_{Dsat0}\left(1+\frac{V_{ds}-V_{dsat}}{V_A}\right)$$

long channel




Short channel




$$g_{ds} = \frac{dI_{dsat}}{dV_{ds}}$$

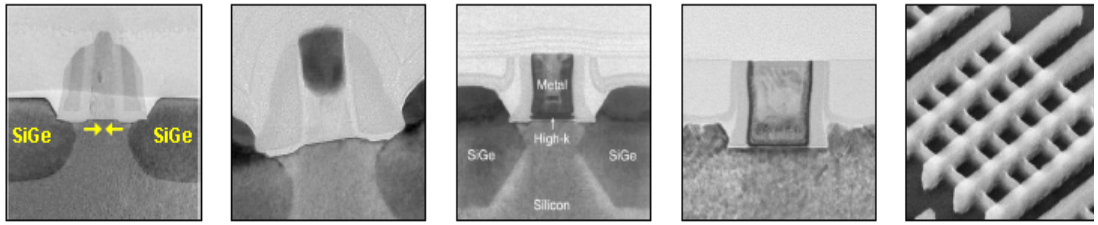
$g_{ds}$ ↑ $A_v$ reduces



C.Hu, Modern Semiconductor Devices for IC, 2009 Prentice Hall

# Technology Trend

| 2003 | 2005 | 2007 | 2009 | 2011 |
|---|---|---|---|---|
| 90 nm | 65 nm | 45 nm | 32 nm | 22 nm |



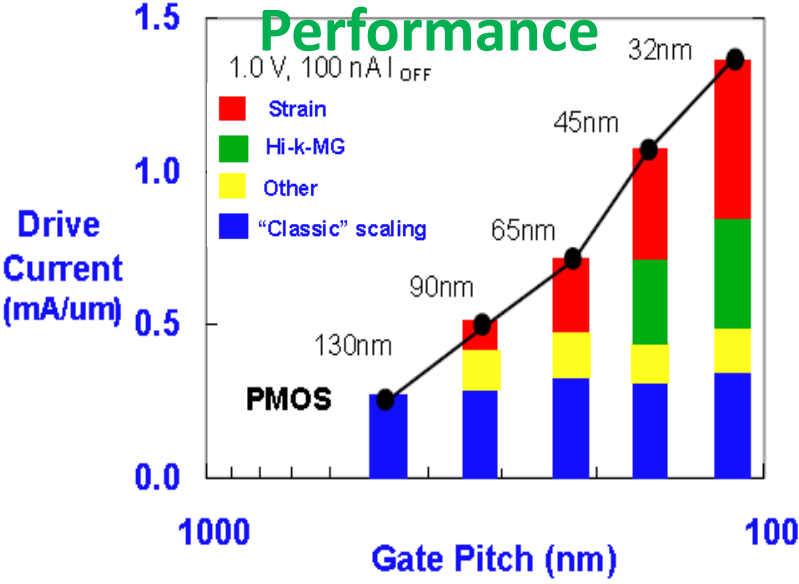

| Invented SiGe Strained Silicon | 2nd Gen. SiGe Strained Silicon | Invented Gate-Last High-k Metal Gate | 2nd Gen. Gate-Last High-k Metal Gate | First to Implement Tri-Gate |
|---|---|---|---|---|

Strained Silicon

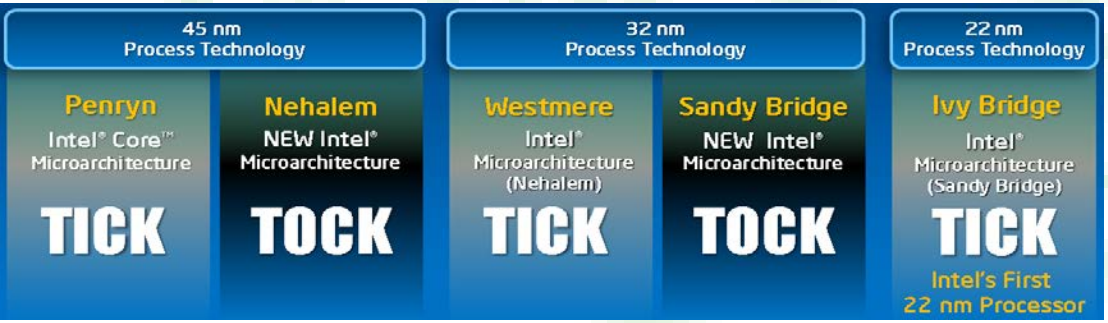
High-k Metal Gate

Tri-Gate

## Performance




## Product






Source : www.intel.com

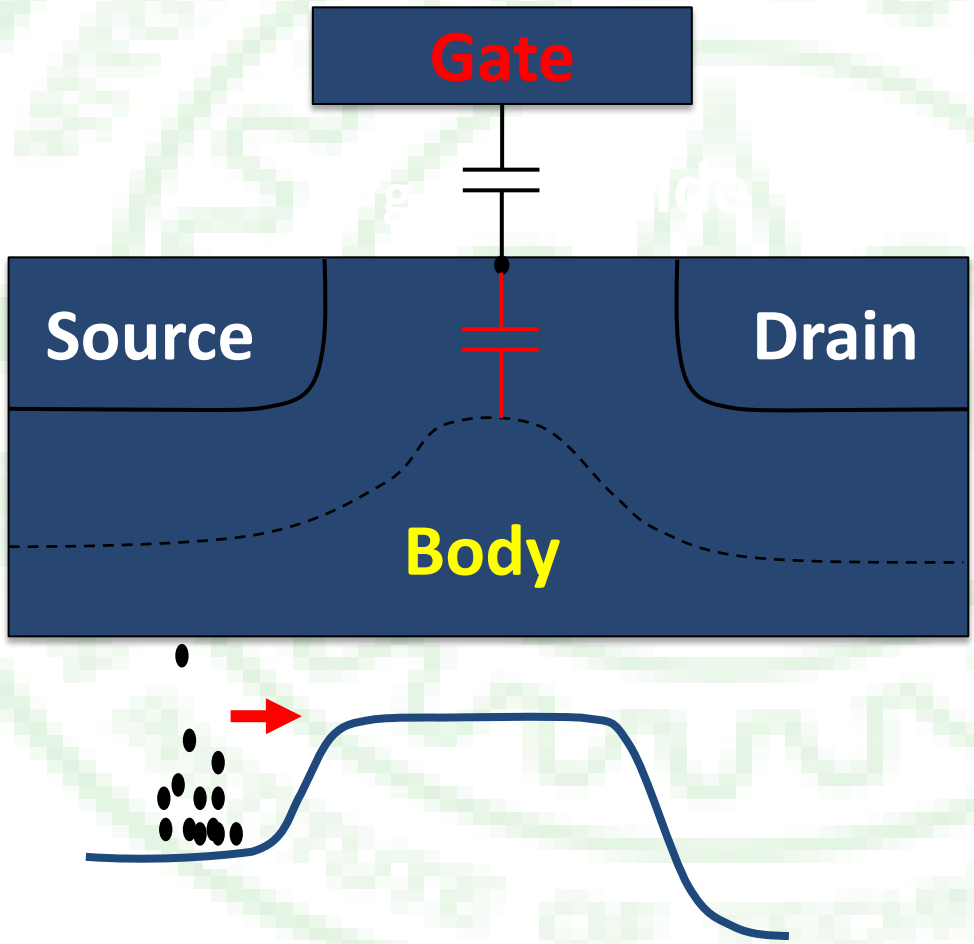
# Wasn't that smooth ride?

- Where is the bottleneck?

$$I_{ON} = \mu \frac{W}{L} C_{ox} (V_{DD} - V_{TH})^2$$

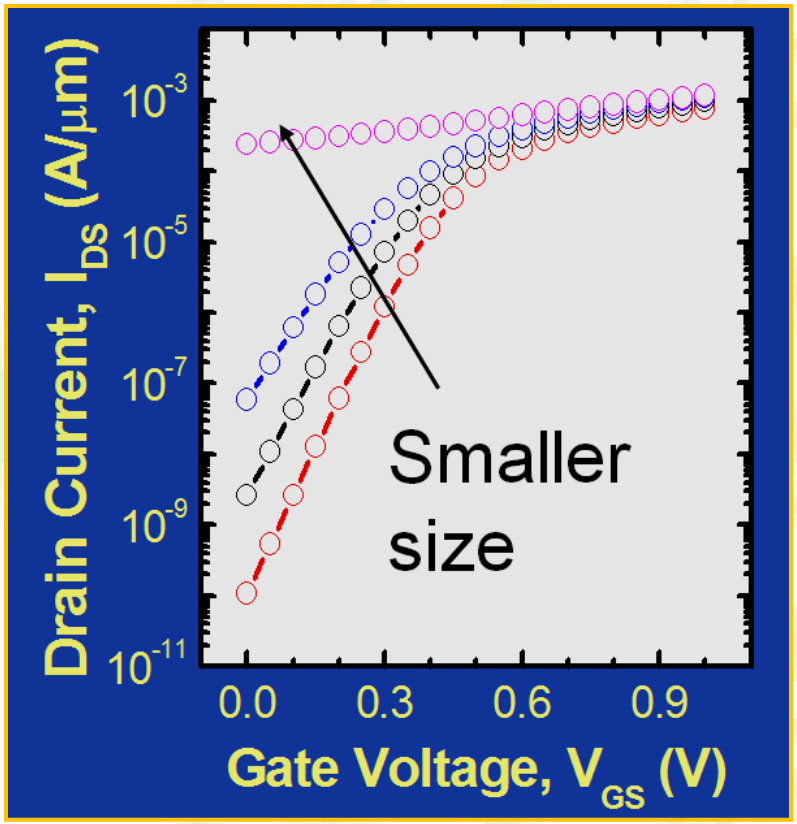
- $V_{TH}$ can't be decreased – why?
- Subthreshold slope gets worse!

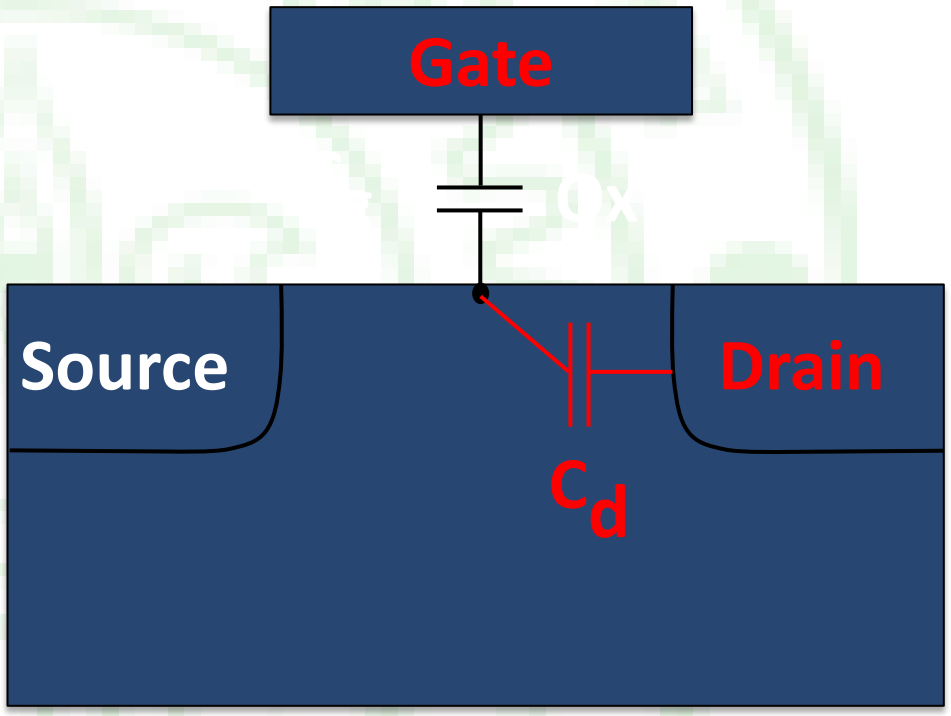
# Thin Depletion Layer - Problem




- $Q_G = Q_i + Q_b$
- Charge sharing

# Short Channel – Big Problem

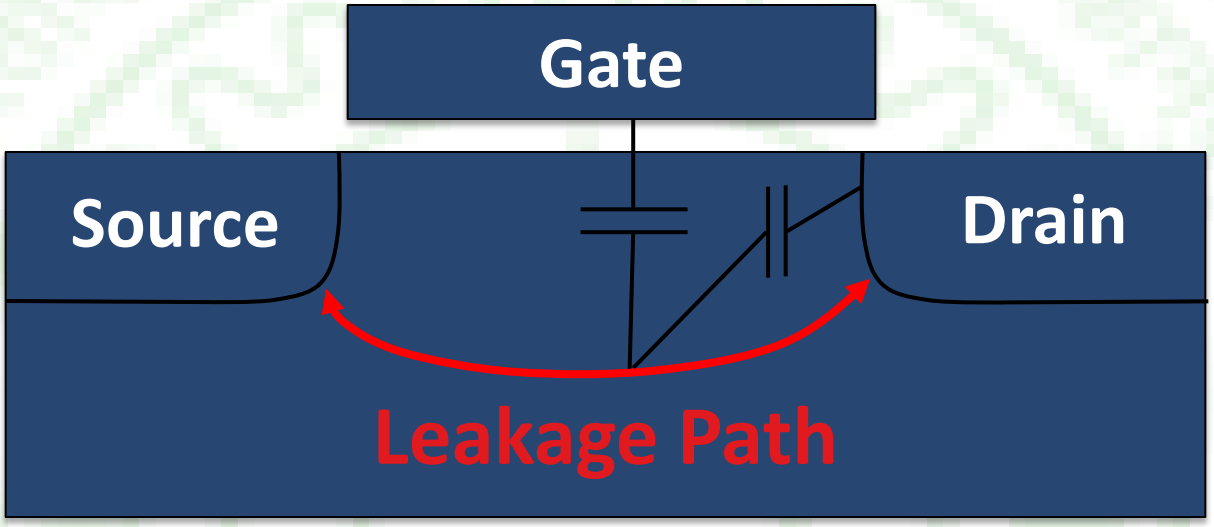



**MOSFET becomes "resistor" at small L.**

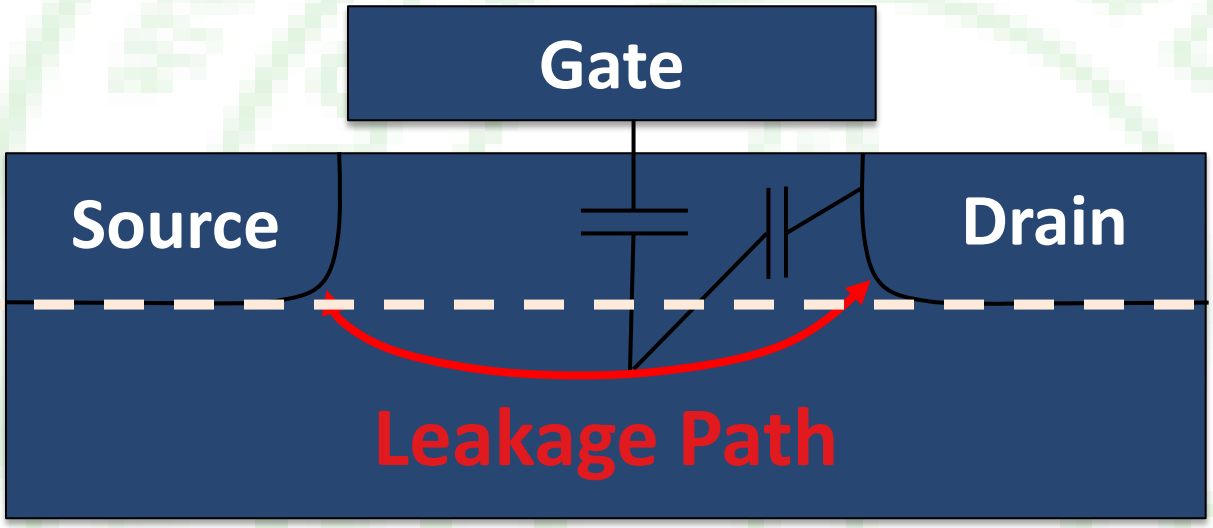
Chenming Hu, "Modern Semiconductor Devices for ICs" 2010, Pearson

# Making Oxide Thin is Not Enough




**Gate cannot control the leakage current paths that are far from the gate.**

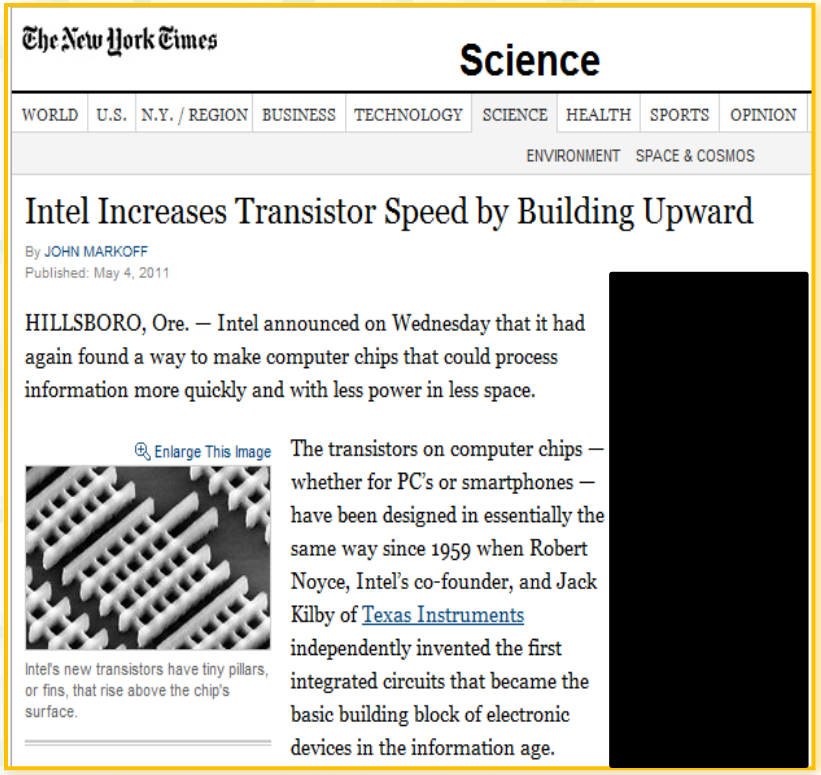
# What can we do?




# May 4, 2011

# The New York Times Front Page

- **Intel will use 3D FinFET at 22nm**
- **Most radical change in decades**
- **There is a competing SOI technology**



The New York Times

Science

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

ENVIRONMENT | SPACE & COSMOS

## Intel Increases Transistor Speed by Building Upward

By JOHN MARKOFF
Published: May 4, 2011

HILLSBORO, Ore. — Intel announced on Wednesday that it had again found a way to make computer chips that could process information more quickly and with less power in less space.

Enlarge This Image

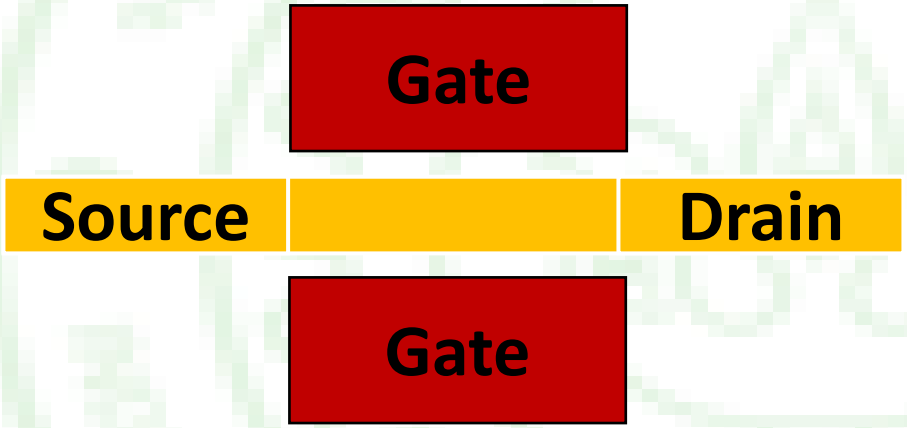
Intel's new transistors have tiny pillars, or fins, that rise above the chip's surface.

The transistors on computer chips — whether for PC's or smartphones — have been designed in essentially the same way since 1959 when Robert Noyce, Intel's co-founder, and Jack Kilby of Texas Instruments independently invented the first integrated circuits that became the basic building block of electronic devices in the information age.

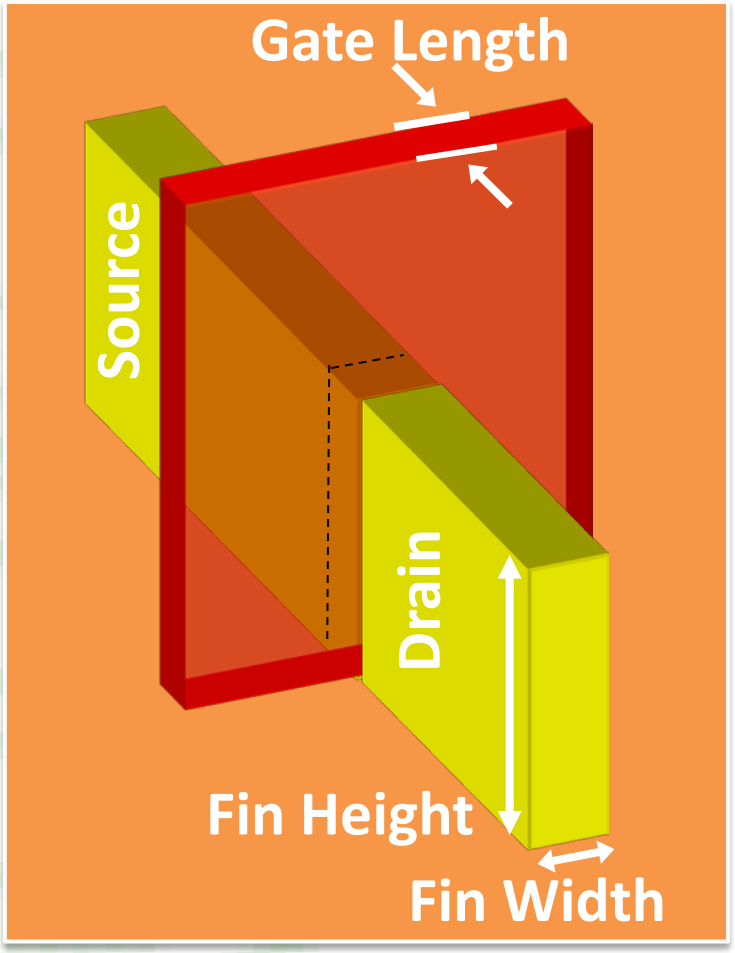
# One Way to Eliminate Si Far from Gate

**Thin body controlled**

**By multiple gates.**



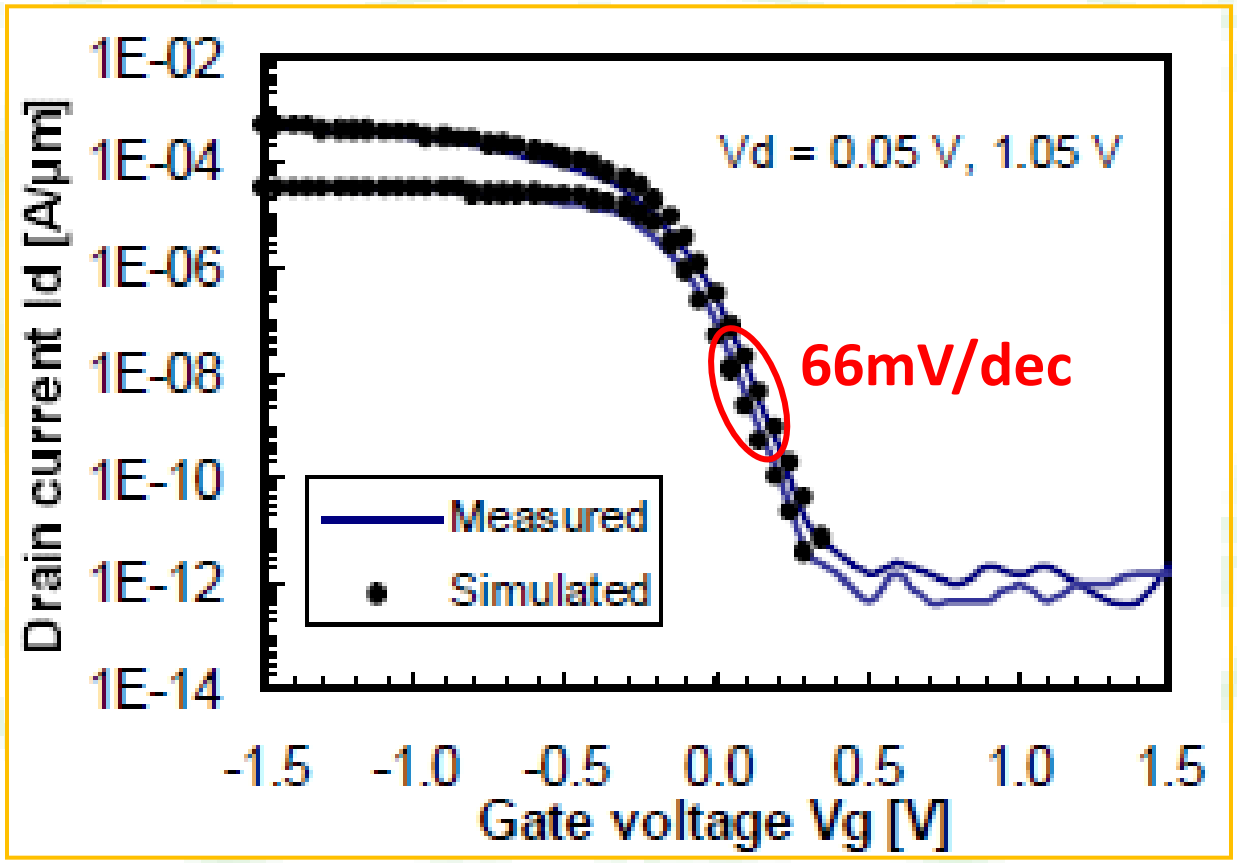

**FinFET body is a thin Fin.**




N. Lindert et al., DRC paper II.A.6, 2001

# 40nm FinFET – 1999

**30nm Fin allows 2.7nm SiO2 & undoped body ridding random dopant fluctuation.**



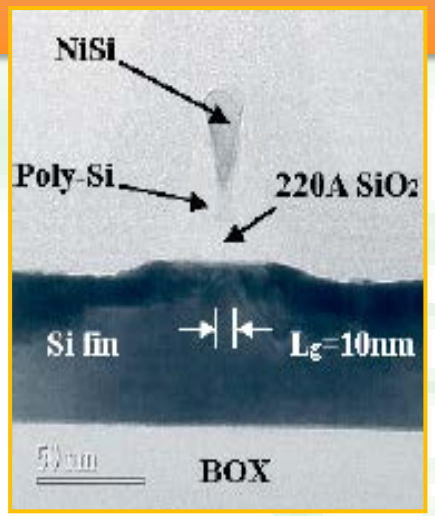

X. Huang et al., IEDM, p. 67, 1999

# Introduced New Scaling Rule

**Leakage is well suppressed if**

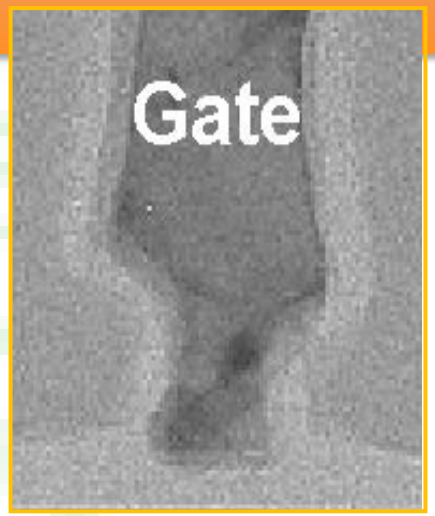
**Fin thickness < Lg**

**10nm Lg AMD**
**2002 IEDM**




**5nm Lg TSMC**
**2004 VLSI**

**3nm Lg KAIST**
**2006 VLSI**



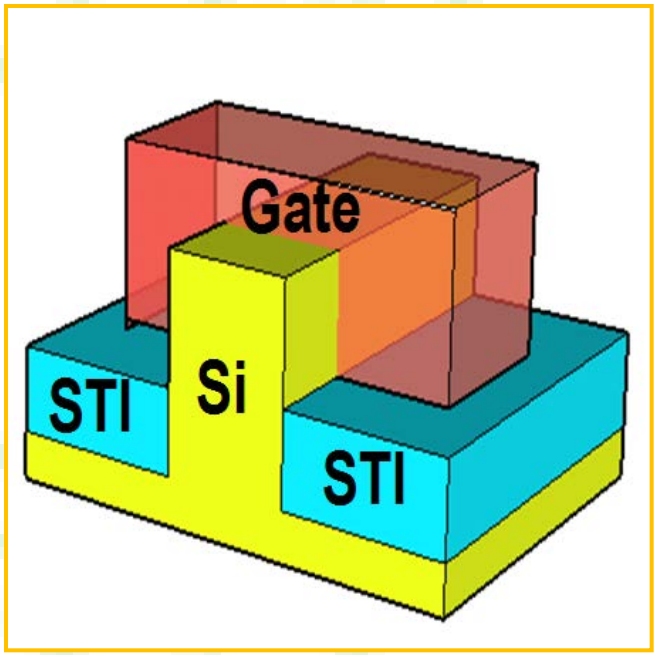
# Two Improvements Since 1999

- **2002 FinFET with thin oxide on Fin top**

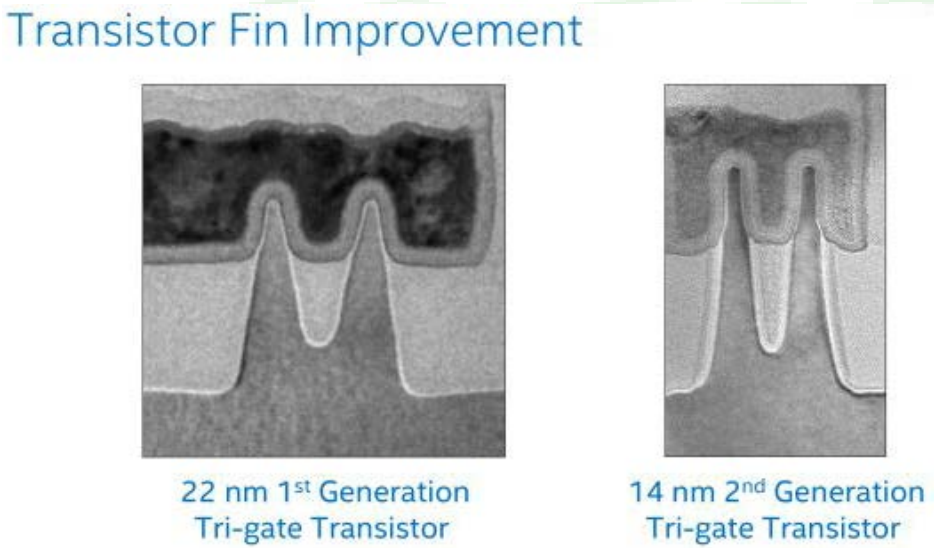
  F.L.Yang et al. (TSMC) 2002 IEDM, p. 225.

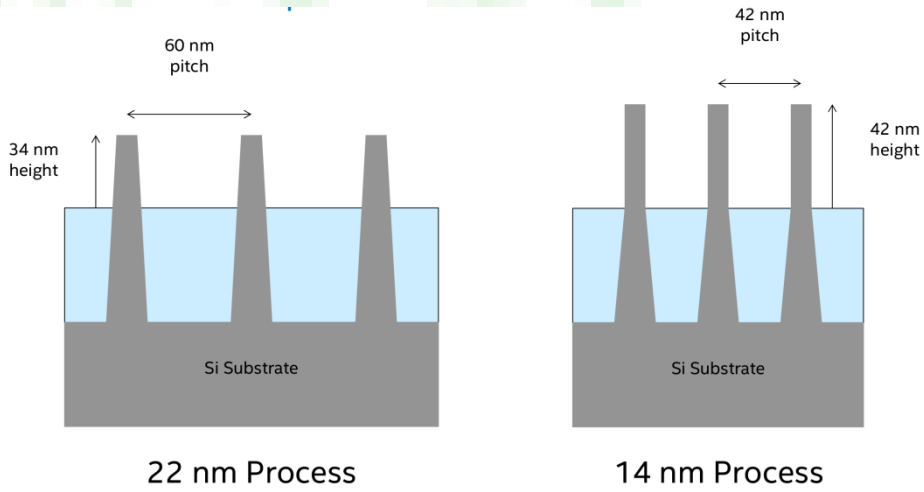
- **2003 FinFET on bulk substrate**

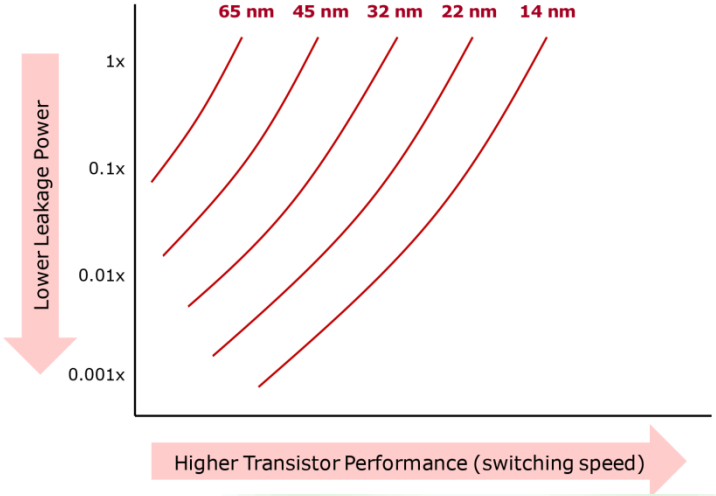
  T. Park et al. (Samsung) 2003 VLSI Symp. p. 135.



# State-of-the-Art **14nm** FinFET



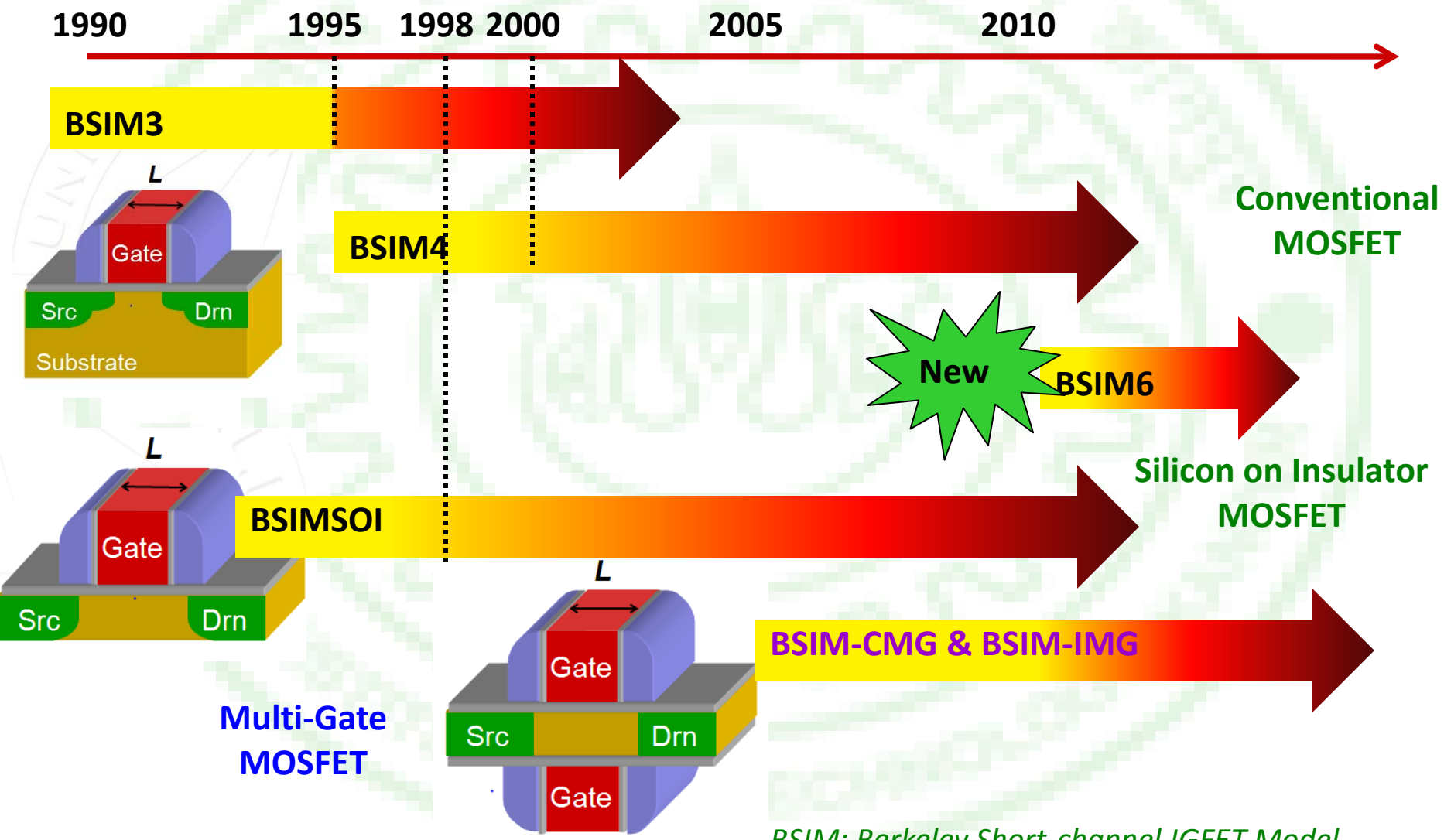




Taller and Thinner Fins for increased drive current and performance

Source: Anandtech

# BSIM Family of Compact Device Models



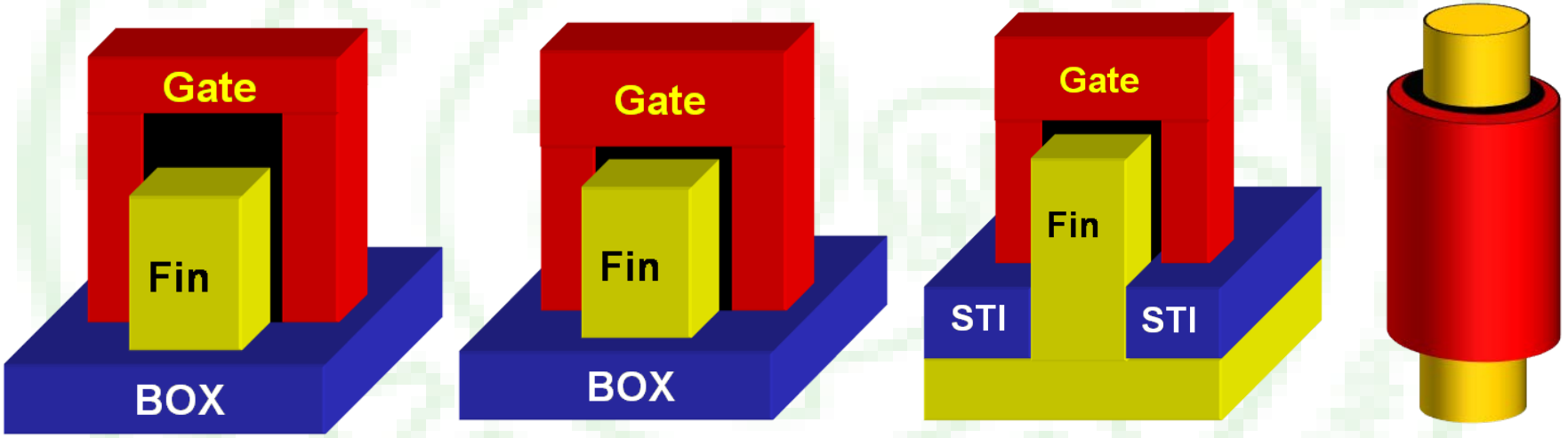

*BSIM: Berkeley Short-channel IGFET Model*

# BSIM-CMG and BSIM-IMG

- **Berkeley Short-channel IGFET Model**
- **First industry standard SPICE model for IC simulation**
- **Used by hundreds of companies for IC design since 1997**
- **BSIM FinFET model became industry standard in March 2012**

# Common-Multi-Gate Modeling

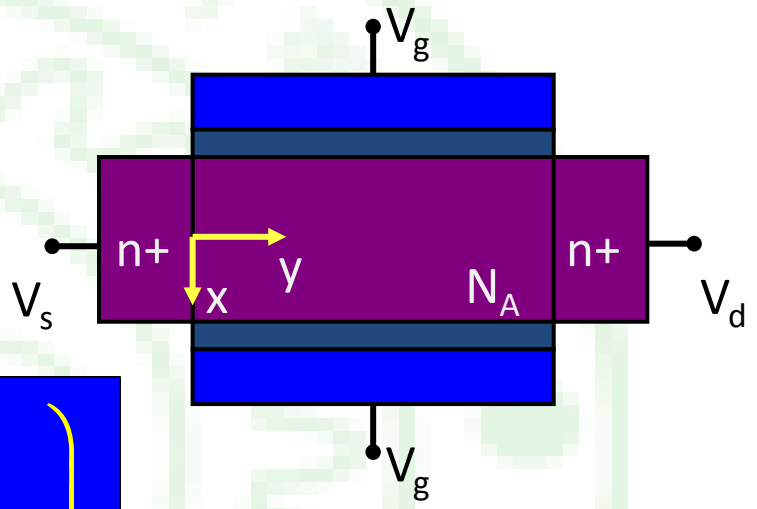
- Common Multi-gate (BSIM-CMG):
  - All gates tied together




  - Surface-potential-based core I-V and C-V model
  - Supports double-gate, triple-gate, quadruple-gate, cylindrical-gate; Bulk and SOI substrates

# BSIM-CMG Model

- Surface potential obtained by solving the 1D Poisson's equation




$$\frac{\partial^2 \psi}{\partial x^2} = \frac{q n_i}{\varepsilon_{Si}} \cdot \left( \underbrace{e^{\frac{q\psi}{kT}} \cdot e^{-\frac{q\varphi_B}{kT}} \cdot e^{-\frac{qV_{ch}}{kT}}}_{\text{Inversion Carriers}} + \underbrace{e^{\frac{q\varphi_B}{kT}}}_{\text{Body Doping}} \right)$$

- A Perturbation approach is used to handle finite body doping

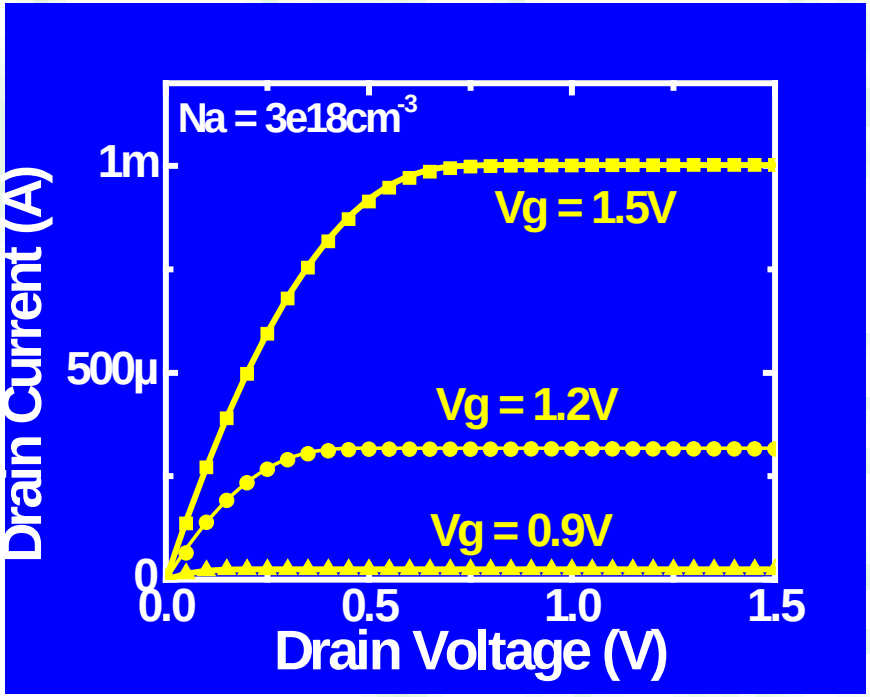
M. V. Dunga et al.,TED 2006

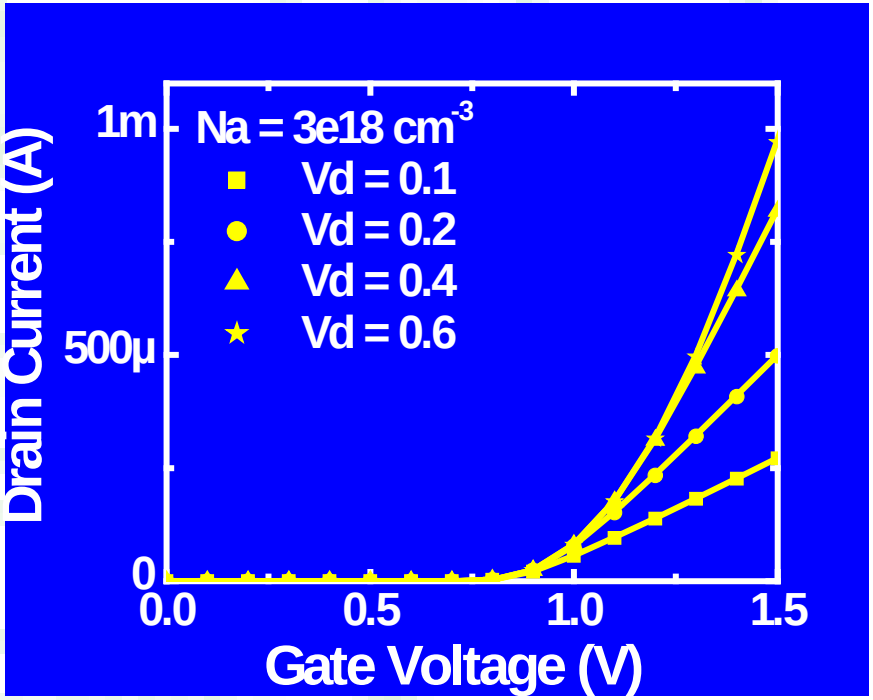
$$\underbrace{\psi}_{\text{Net Surface Potential}} = \underbrace{\psi_{inv}}_{\text{Inversion Carriers only}} + \underbrace{\psi_{pert}}_{\text{Perturbation due to finite doping}}$$

# BSIM-CMG Model

- Drain current derived from drift-diffusion

$$\mathrm{Id} = \mu \frac{W_{eff}}{L} \left[ \frac{Q_i^2}{2C_{ox}} + 2\,\mathrm{V_t}\,\mathrm{Q_i} - \mathrm{V_t} \cdot (5\mathrm{C_{Si}}\mathrm{V_t} + \mathrm{Q_B})\,\ln(5\mathrm{V_t}\mathrm{C_{Si}} + \mathrm{Q_B} + \mathrm{Q_i}) \right]_{\mathrm{d}}^{\mathrm{s}}$$

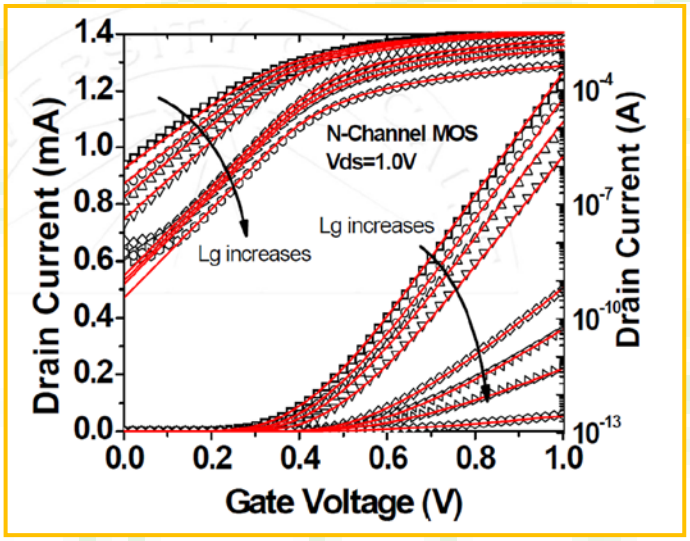


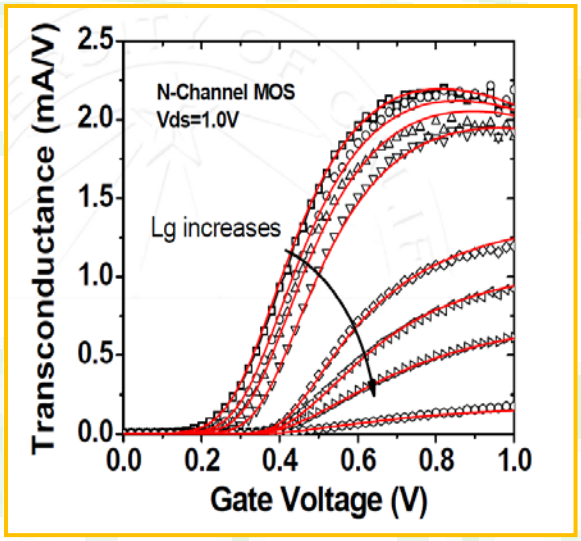



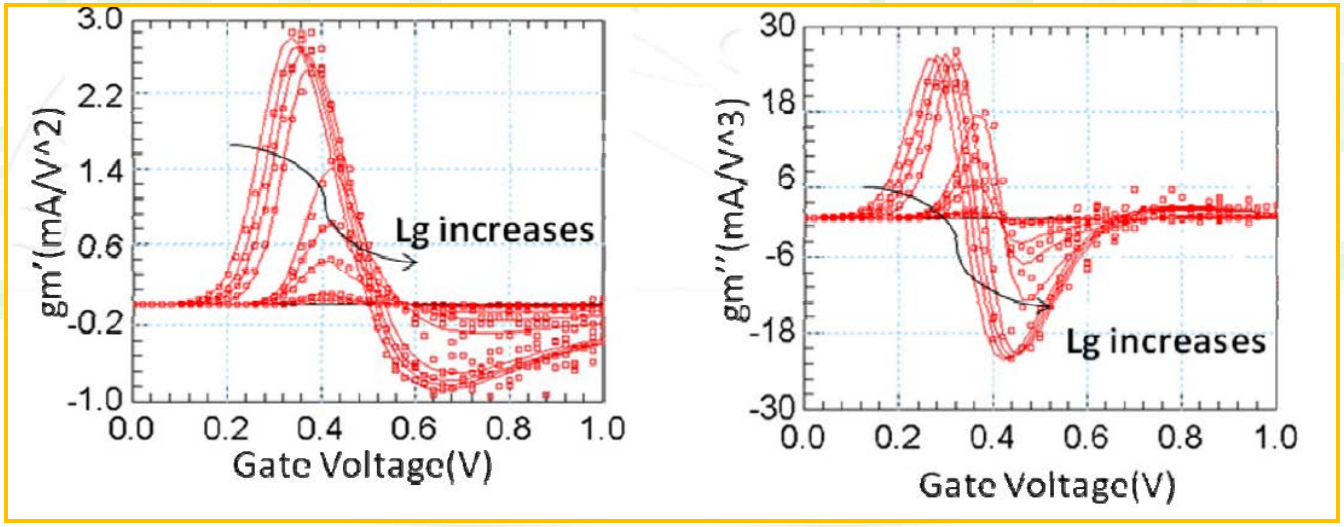

M. V. Dunga, UCB Ph.D. Thesis

# BSIM-CMG

# Global fitting with 30nm–10μm FinFETs



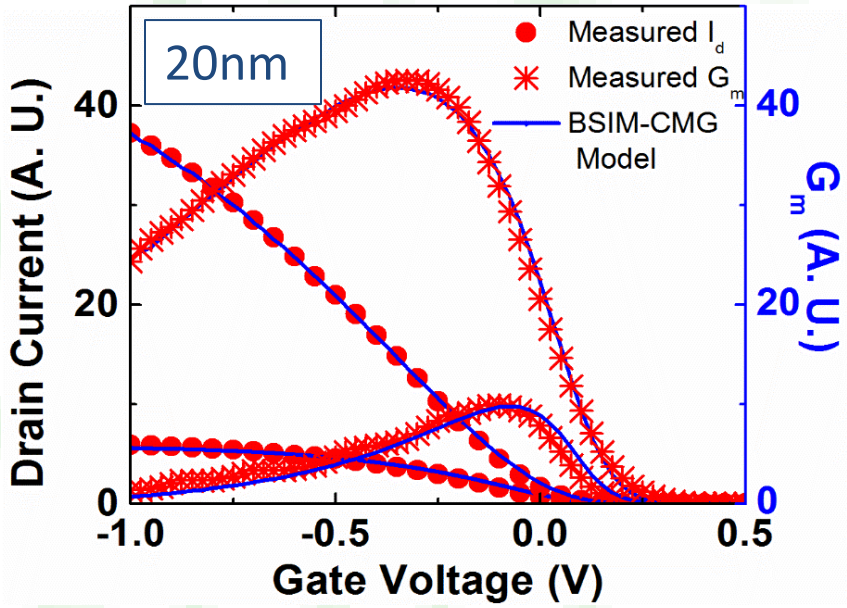



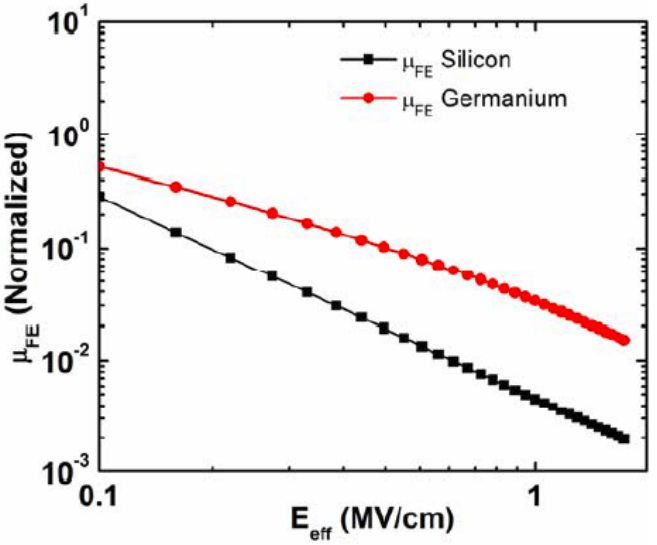



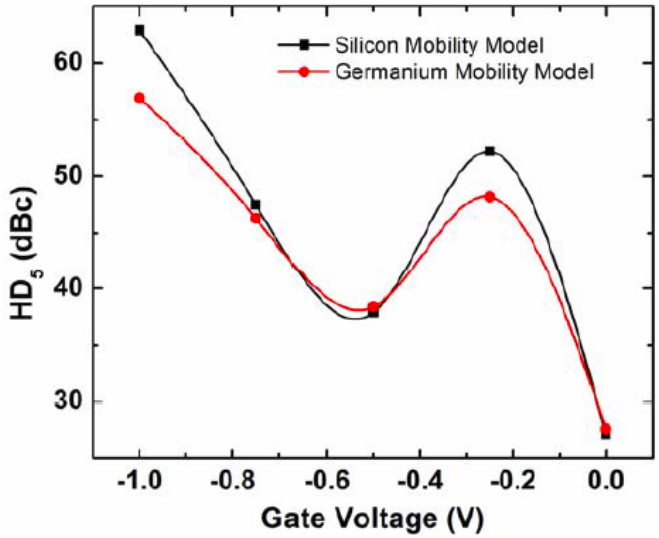
# Modeling of Germanium FinFETs @10nm

- Ge FinFET may be used in 10nm node for better P-FinFET.
- Industry standard BSIM FinFET model can now model Ge FinFET.
- Early availability of a unified Si/Ge FinFET model facilitates technology-circuits co-development.

# Modeling of Germanium FinFETs @10nm



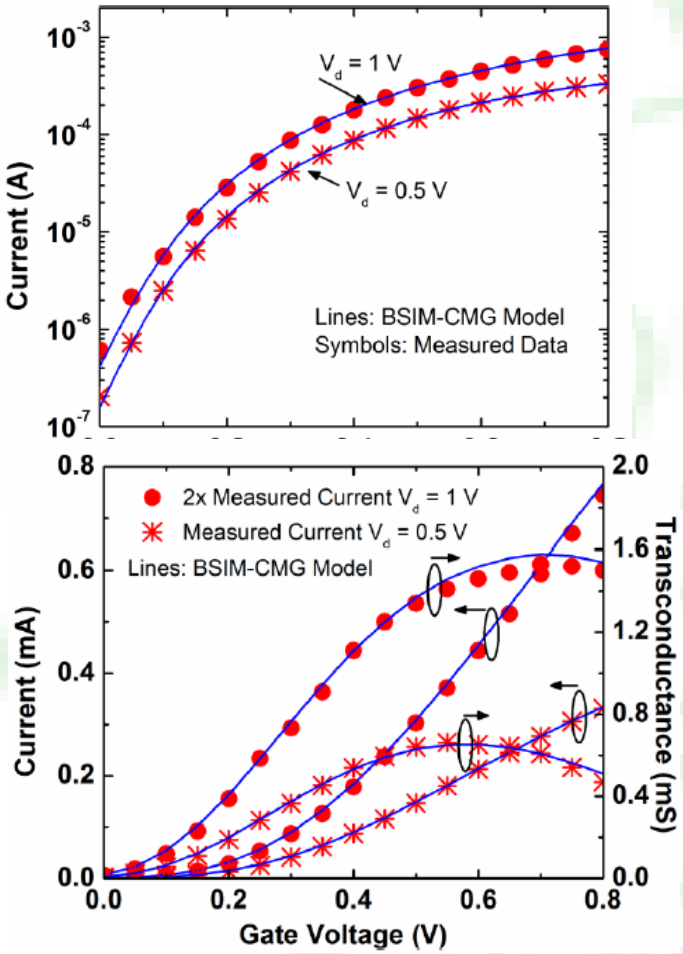

- Due to the lower *m** of holes in Ge the charge-centroid is farther away from the oxide interface resulting in a weaker SR scattering.
- Ge mobility has a weaker dependence on $E_{eff}$ up-to ~0.5 MV/cm as the impact of SR scattering is only seen at much higher $E_{eff}$ in Ge as compared to Si.

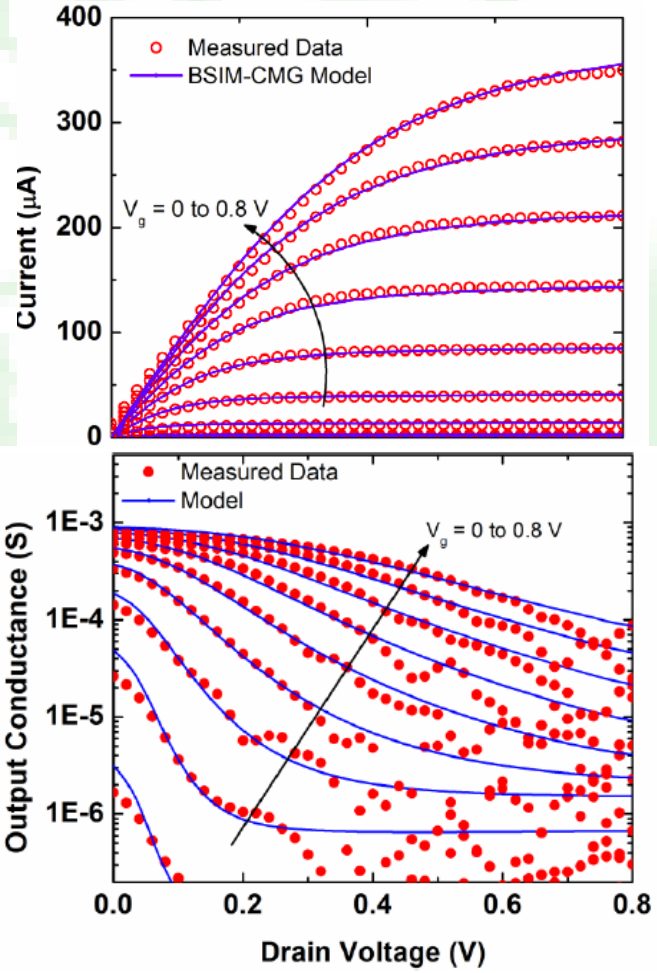
S. Khandelwal et. al., "Modeling 20nm Germanium FinFET with the Industry Standard FinFET Model", IEEE Electron Device Letters, July 2014.

# Modeling of InGaAs FinFET @10nm

L = 20 nm, H = 30 nm, W = 20 nm, Nfin = 4.

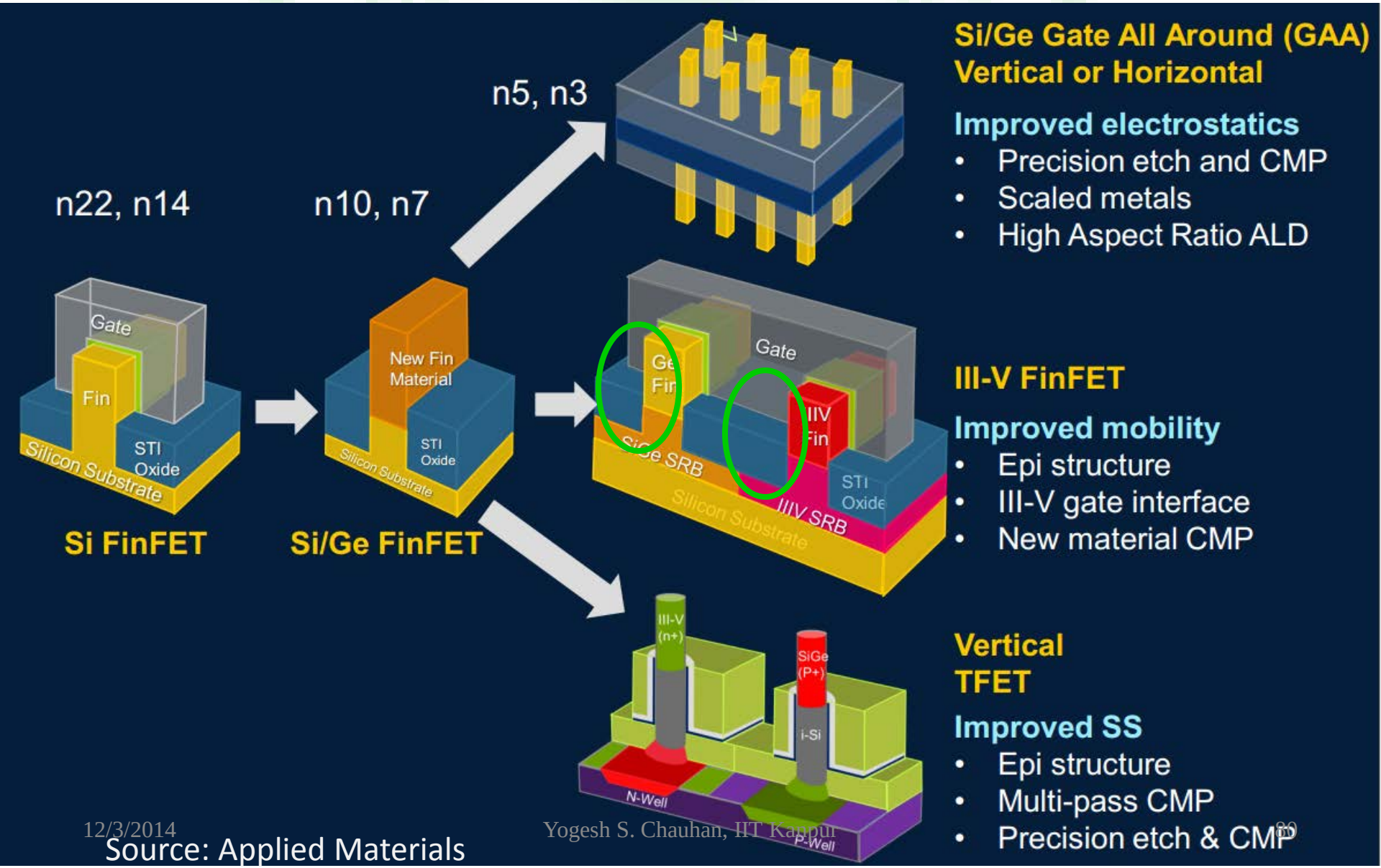
Data from: J. J. Gu et al. IEDM 2012



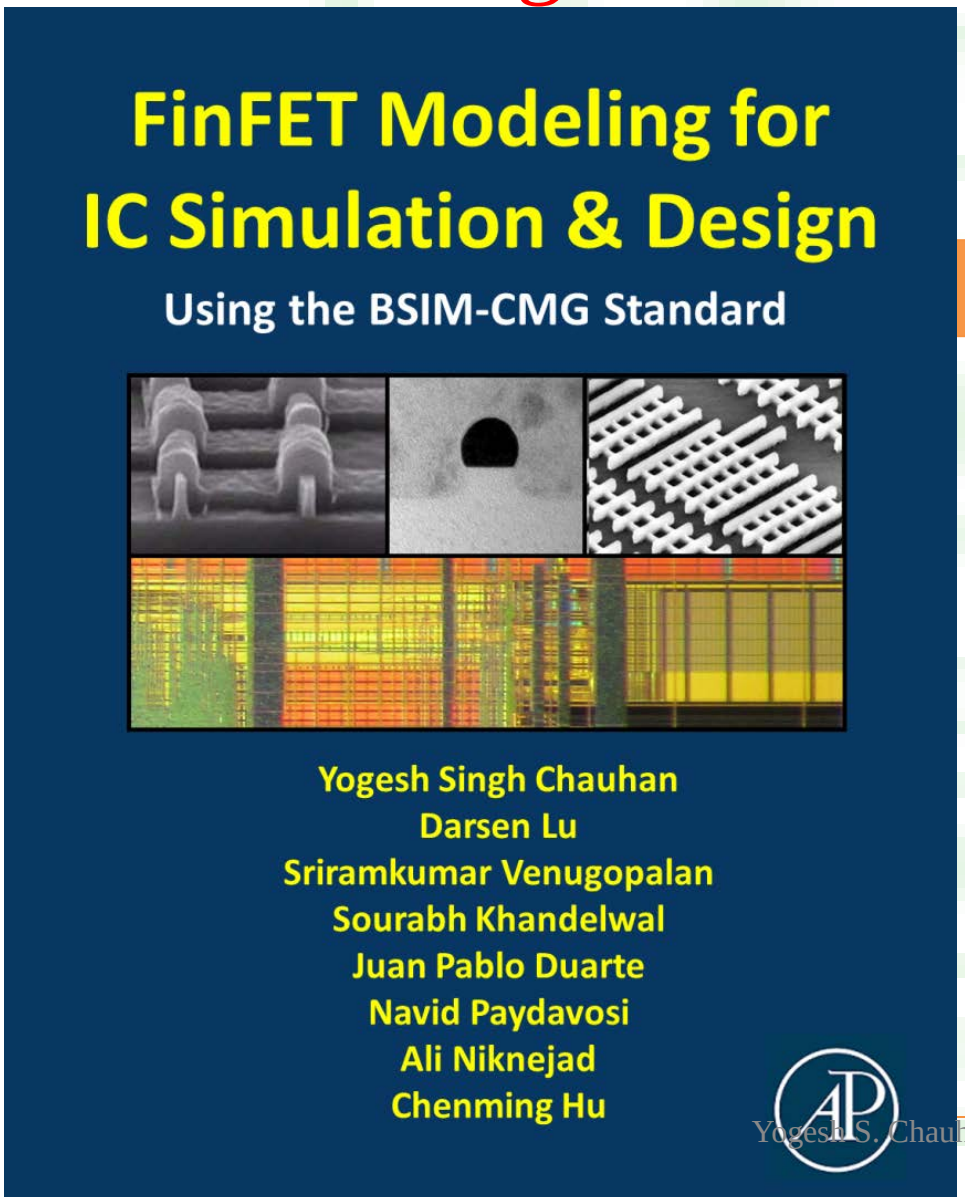

S. Khandelwal et. al., "InGaAs FinFET Modeling Using Industry Standard Compact Model BSIM-CMG", Workshop on Compact Modeling, Washington D.C., USA, June 2014.

# Transistor Pathway




Source: Applied Materials

# FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard




**Book**

## Chapters

1. FinFET- from Device Concept to Standard Compact Model
2. Analog/RF behavior of FinFET
3. Core Model for FinFETs
4. Channel Current and Real Device Effects
5. Leakage Currents
6. Charge, Capacitance and Non-Quasi-Static Effect
7. Parasitic Resistances and Capacitances
8. Noise
9. Junction Diode Current and Capacitance
10. Benchmark tests for Compact Models
11. BSIM-CMG Model Parameter Extraction
12. Temperature Effects

# Acknowledgement

- My students
- BSIM team
- CMC members