ECE520 – VLSI Design

Lecture 2: Basic MOS Physics

Payman Zarkesh-Ha

Office: ECE Bldg. 230B
Office hours: Wednesday 2:00-3:00PM or by appointment
E-mail: pzarkesh@unm.edu

Review of Last Lecture

- **Semiconductor technology trend and Moor's law**
- **Benefits of transistor scaling:**
  - **More functionality in the same foot print**
  - **Faster device**
  - **Devices with less switching energy**
  - **Less cost/function**
- **Challenges of transistor scaling:**
  - **Device size reaching quantum level**
  - **Power dissipation and heat removal concerns**
  - **Interconnect worsen by scaling**
  - **Manufacturing yield issues**

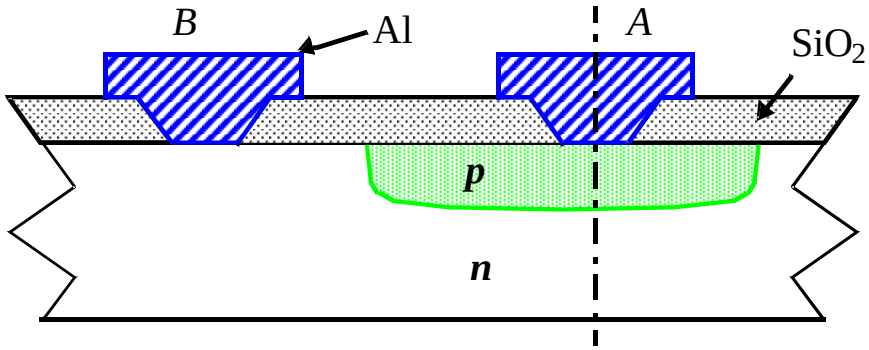
Today's Lecture

- **Overview of Diode Physics**
- **BASIC MOS Physics:**
  - **Understanding of device operation**
  - **Basic device equations for long channel MOSFET**
  - **Long channel MOS models for manual analysis**

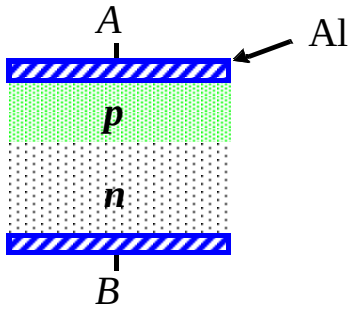
Reading Assignment

- **Today we will review Chapter 3 (MOS Physics)**
  - **Skim through Diodes but focus on Section 3.2.3 (diode transient behavior)**
  - **Study Section 3.3 (MOS transistor) *thoroughly***

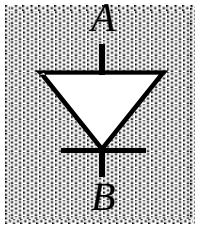
The Diode



Cross-section of *pn* -junction in an IC process

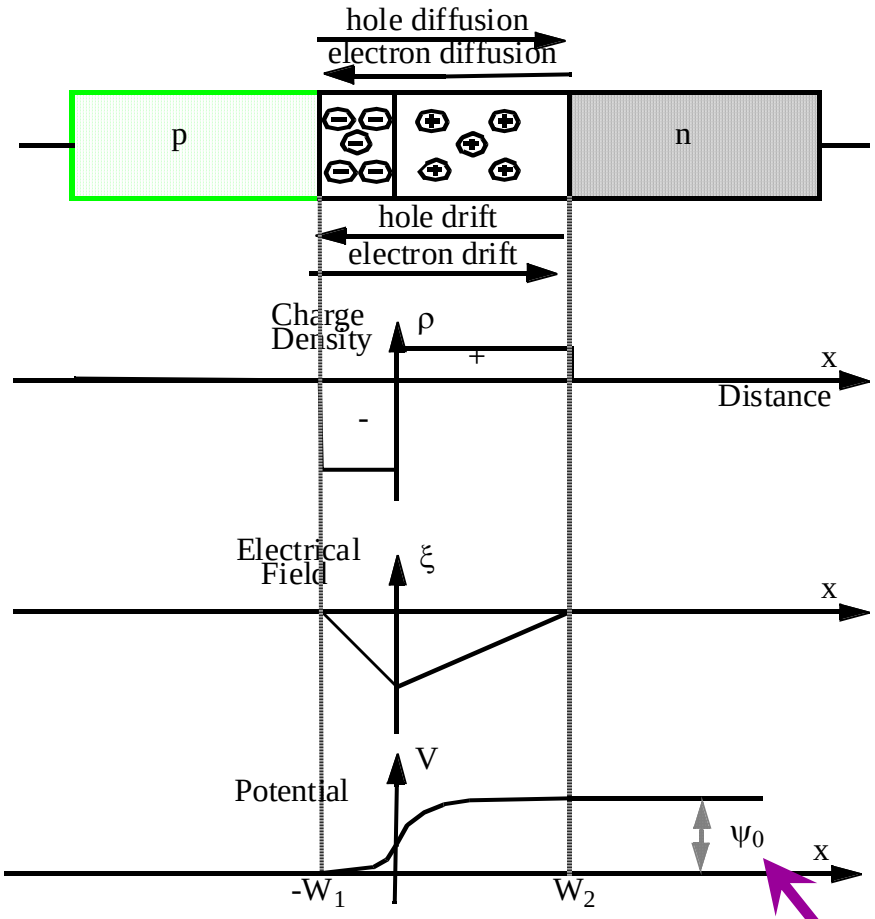


One-dimensional representation



diode symbol

Depletion Region



(a) Current flow.

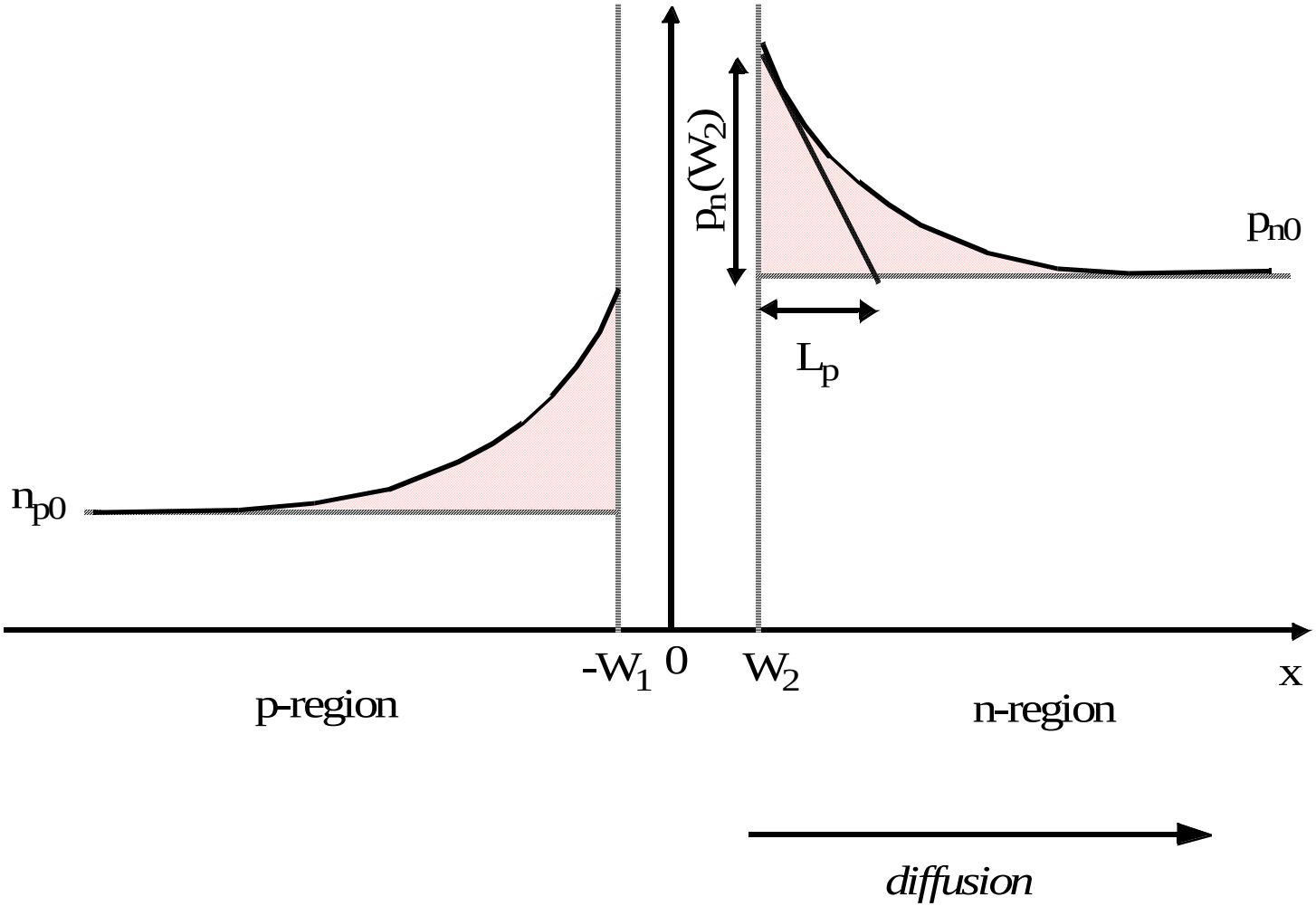
(b) Charge density.

(c) Electric field.

(d) Electrostatic potential.

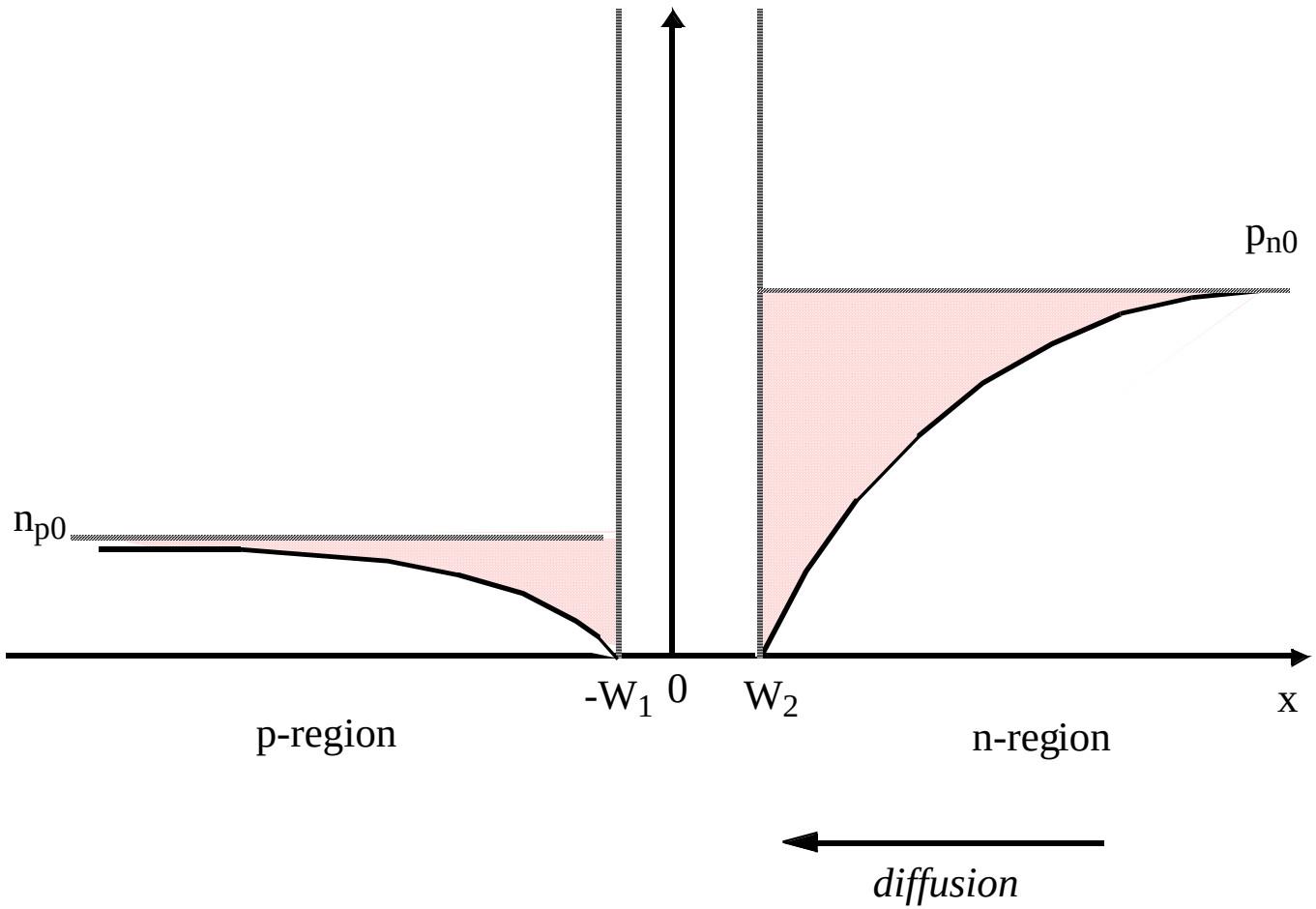
Built-in potential

Forward Bias Diode



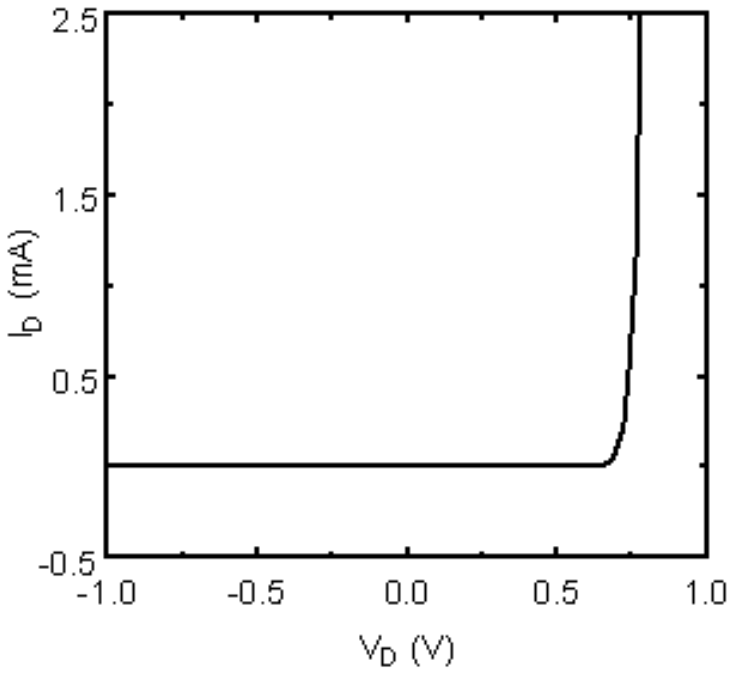
Typically avoided in Digital ICs

Reverse Bias Diode

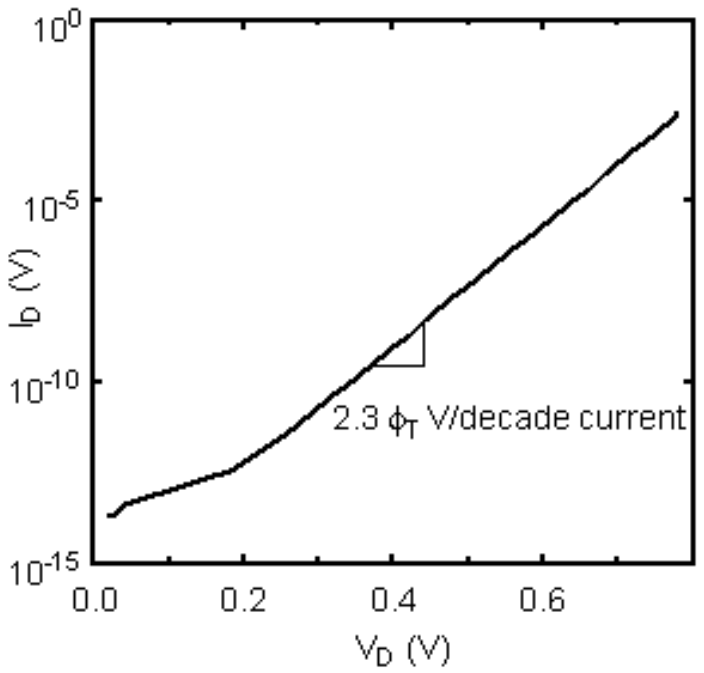


The Dominant Operation Mode

Diode IV Curve



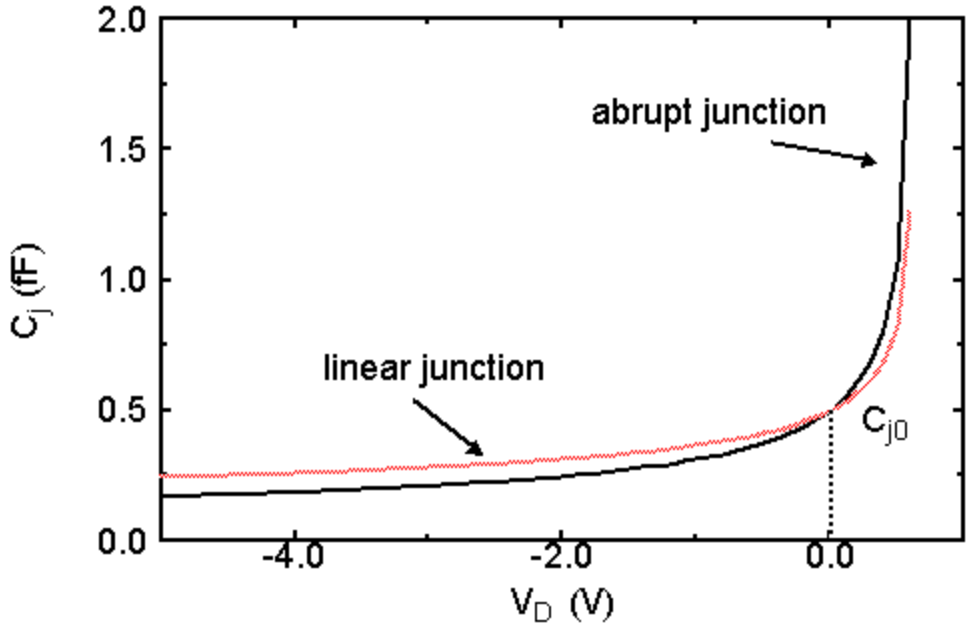
(a) On a linear scale.



(b) On a logarithmic scale (forward bias).

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

Junction Capacitance



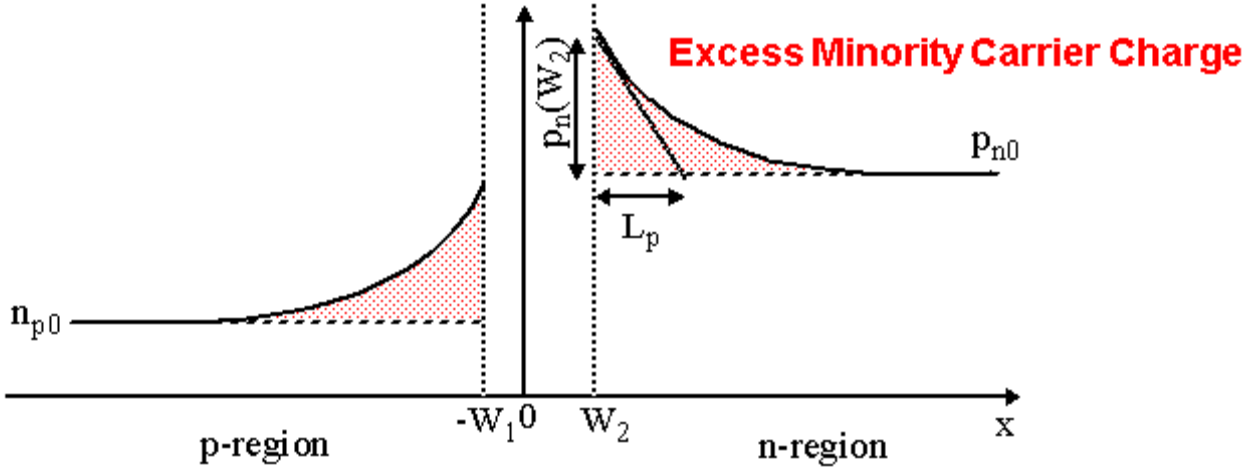
$$C_j = \frac{C_{j0}}{(1 - V_D/\phi_0)^m}$$

m = 0.5: abrupt junction
m = 0.33: linear junction

$$C_{j0} = A_D\sqrt{\left(\frac{\varepsilon_{si} q}{2}\frac{N_A N_D}{N_A + N_D}\right)\phi_0^{-1}} \qquad \phi_0 = \frac{KT}{q}\mathrm{Ln}\left(\frac{N_A N_D}{n_i^2}\right)$$

Built-in potential

Diffusion Capacitance



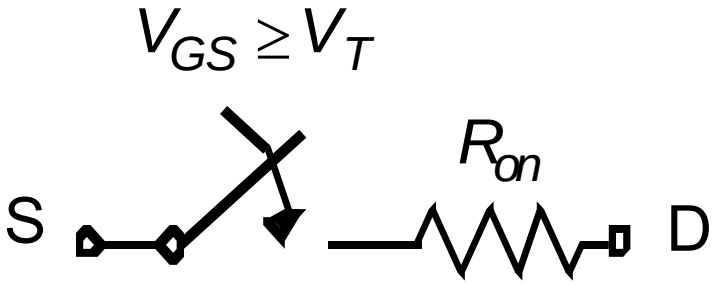
$$C_d = \frac{dQ_D}{dV_D} = \tau_T \frac{dI_D}{dV_D} \approx \frac{\tau_T I_D}{\phi_T}$$

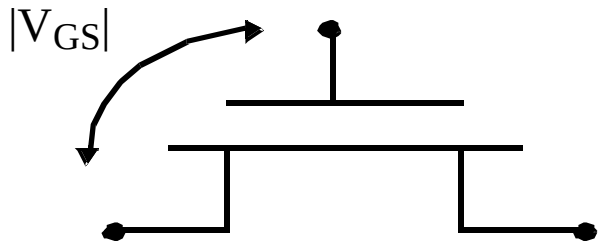
$$\phi_T = \frac{KT}{q}$$

Thermal Potential

What is a Transistor?

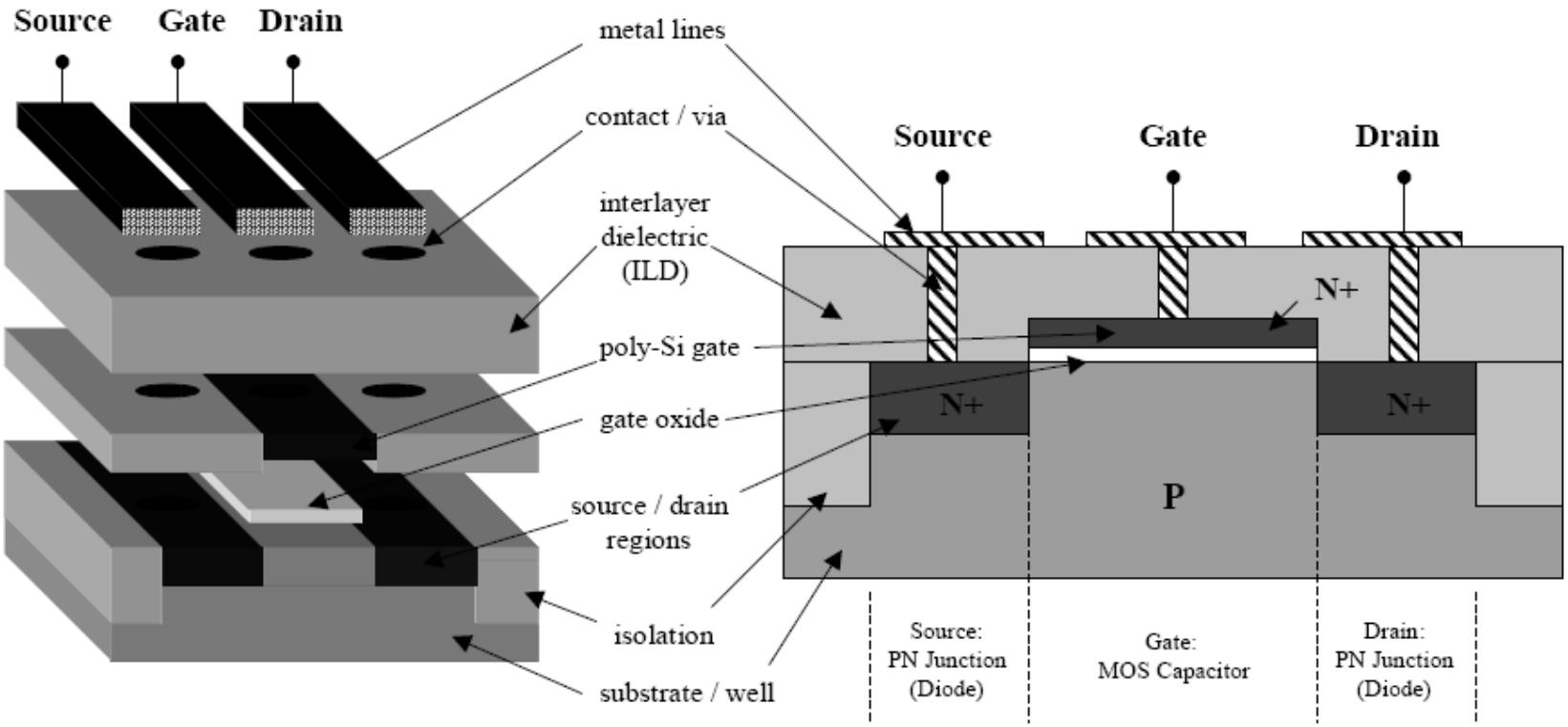
A Switch! ⟷ An MOS Transistor





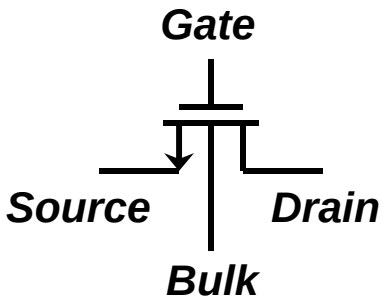
MOSFET Top & Cross Section View

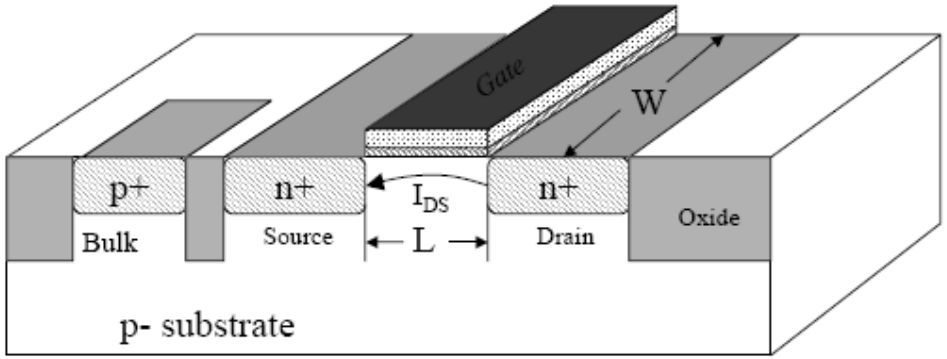
Metal **O**xide **S**emiconductor **F**ield **E**ffect **T**ransistor



NMOS Device Cross-Section

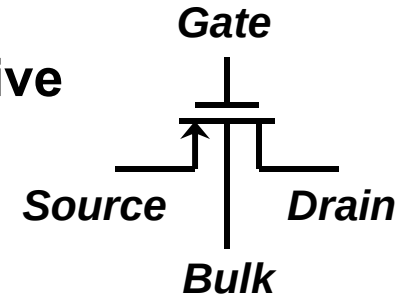
- I_{DS} is Defined as "from Drain to Source" Current
  - Majority carriers are electrons
  - NMOS device conducts when "gate-to-source" voltage is positive
- I_{DS} is as a function of:
  - Channel width (W)
  - Inverse of channel length (1/L)
  - Gate-to-source potential (V_{GS})

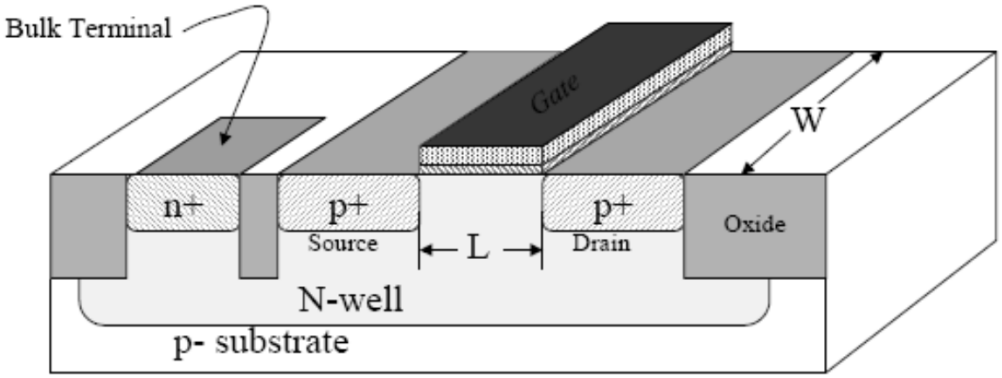




PMOS Device Cross-Section

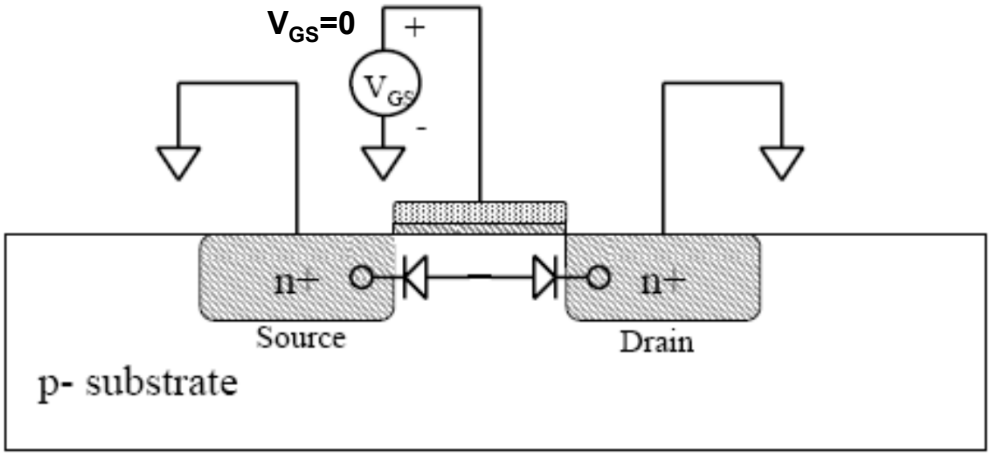
- Complement of NMOS
- Built inside an N-well implant in substrate
- Majority carriers are holes, not electrons
- Conducts when gate-source voltage is negative





Device Operation: Cutoff

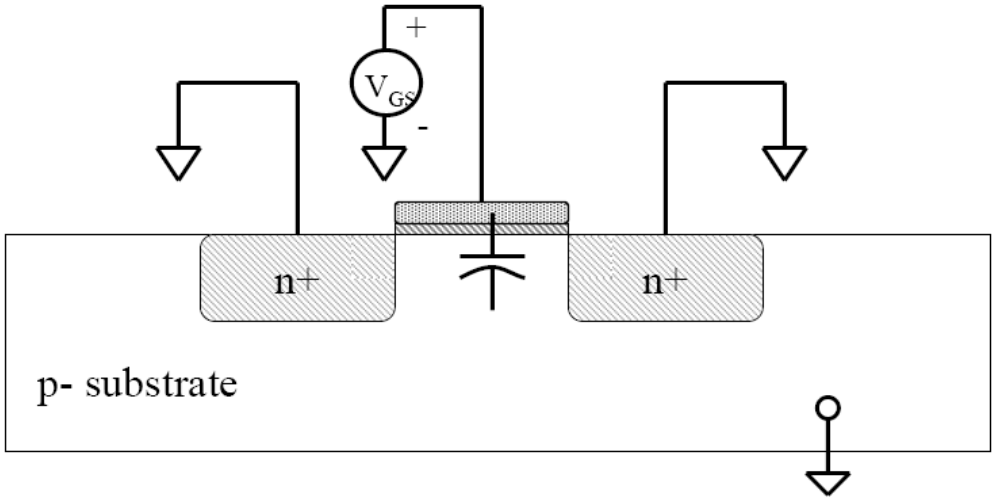
- **Cutoff region (V_{GS} = 0)**
  - The Source to Drain connection looks like two back to back series connected diode
- **Therefore ideally I_{DS} = 0**
  - 1st order approximation only



Gate Oxide Capacitance

- **Polysilicon gate forms a conductive top plate of a capacitor**
  - **Gate oxide forms the dielectric of a parallel plate capacitor**
  - **P-doped substrate forms the conductive bottom plate of a capacitor**

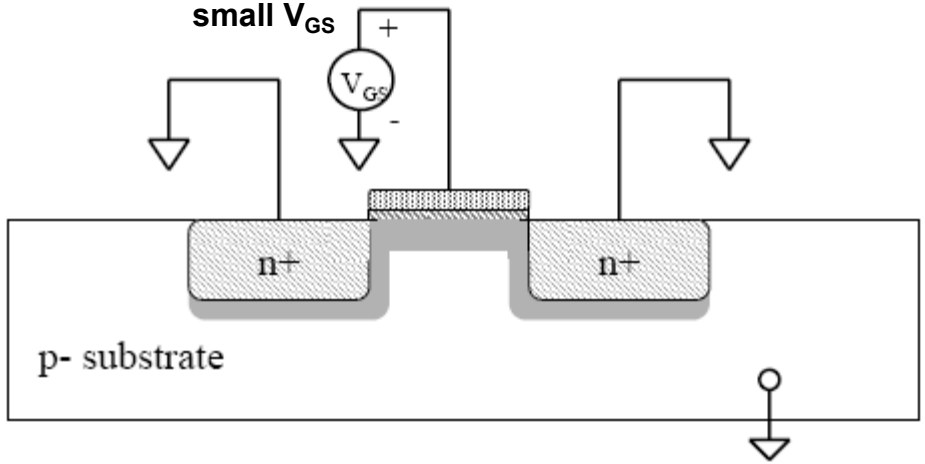
$$C_{ox} = \frac{\varepsilon A}{t_{ox}} \qquad C'_{ox} = \frac{C}{A} \equiv \frac{\varepsilon_{ox}}{t_{ox}}$$



Device Operation: Depletion

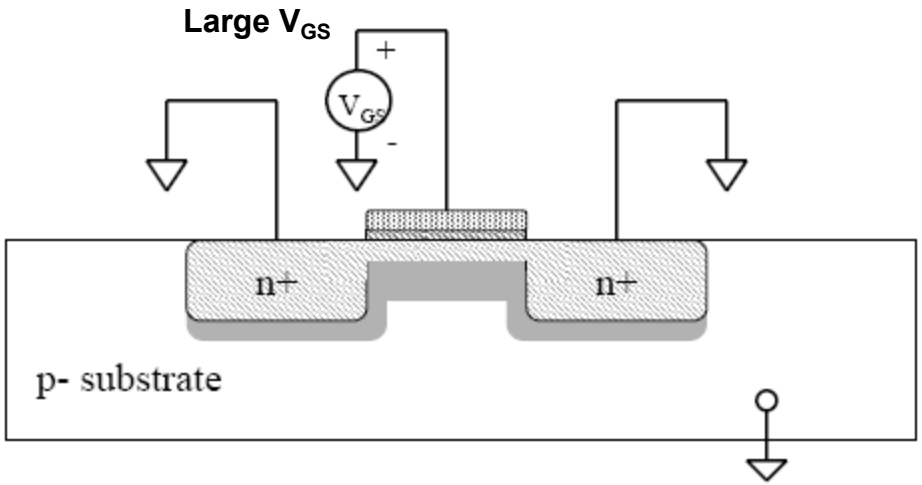
- **As gate potential increases**
  - **Positive majority carriers (holes) in the substrate repelled from the surface (depleting the material of carriers)**
  - **A depletion region is formed under the surface of the gate**
  - **This depletion region is formed as potential at the silicon surface underneath the gate reaches φ_F**

$$\phi_F = \frac{KT}{q} Ln\left(\frac{N_A}{n_i}\right)$$



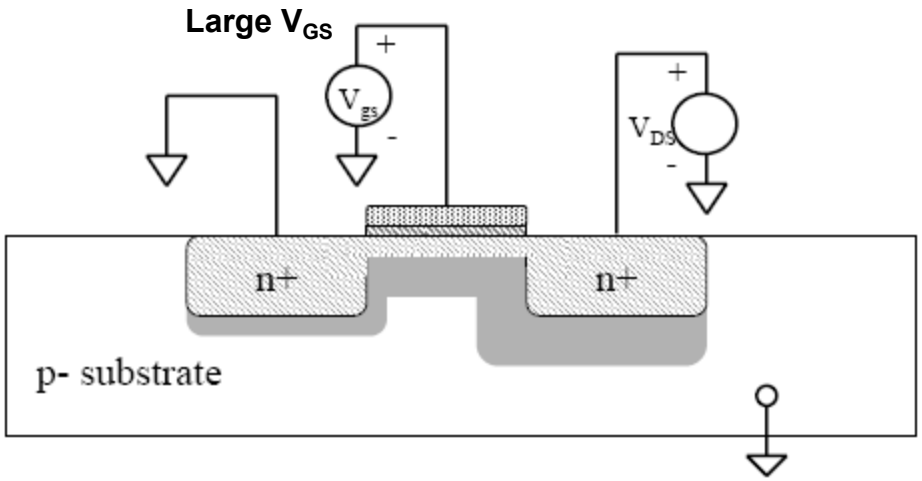
Device Operation: Inversion

- As the surface potential beneath the gate increases beyond φ_F
  - Electrons from heavily doped source and drain are attracted to the gate and move into the channel
  - When the surface potential reaches $2\varphi_F$ the charge density of electrons in the channel equals the original doping density of the P-substrate
  - At this time the channel is inverted
  - Therefore, a conductive path is formed between source and drain

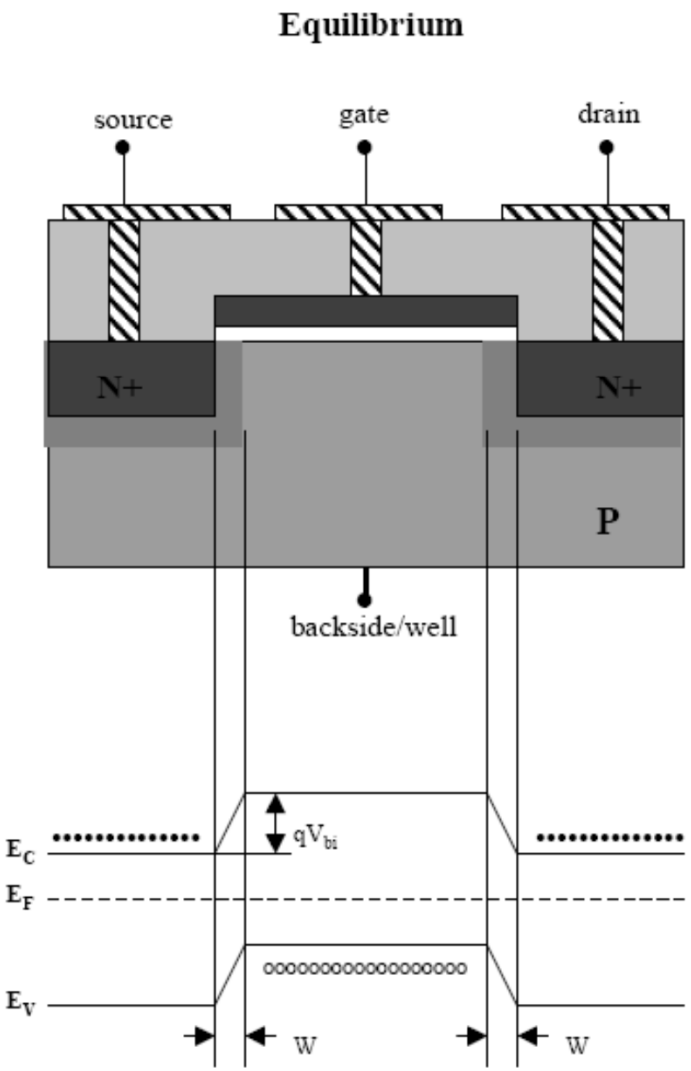


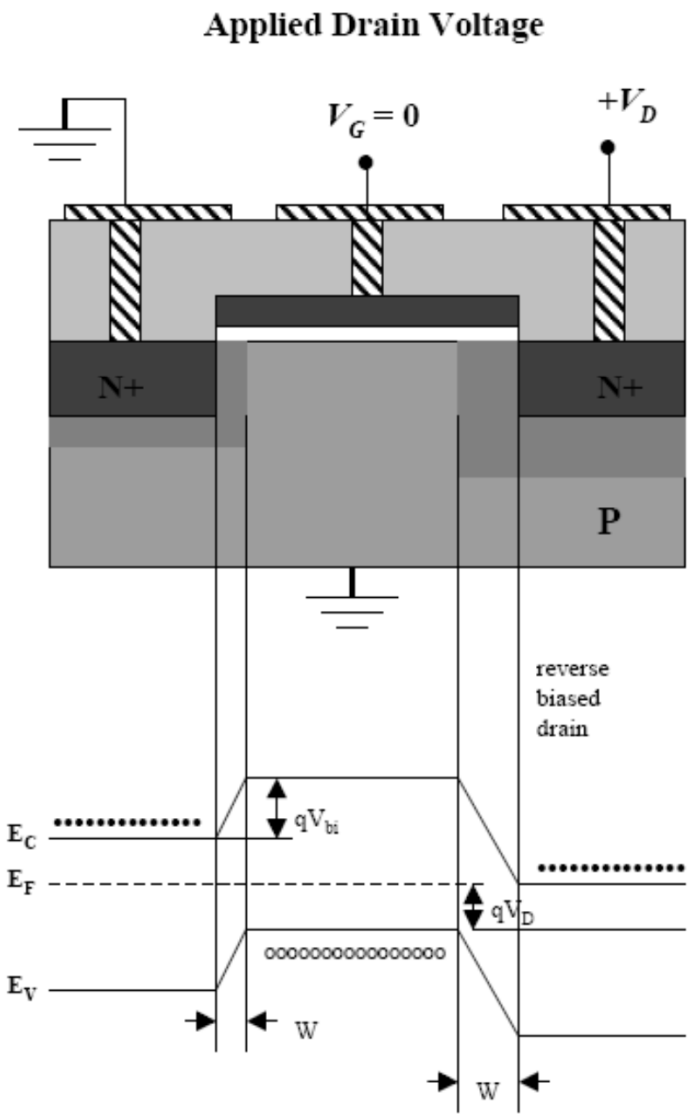
Device Operation: Inversion

- Inversion region is simply a resistor
- We need an applied V_{DS} to get current flow
- When drain voltage is applied the depletion region grows at the drain junction



MOSFET Band Diagram

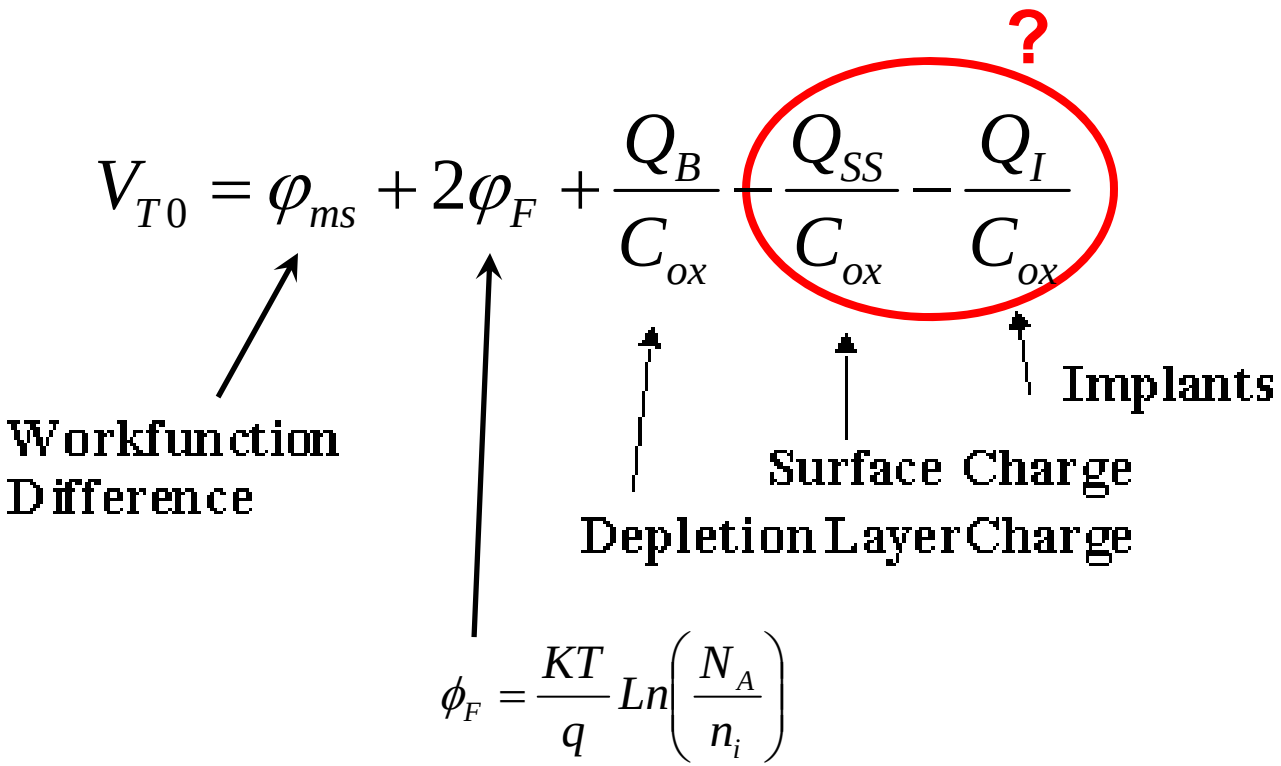




MOSFET Threshold Voltage

- The gate potential at which the channel inverts is called the threshold voltage (V_T)
- V_T is always referenced in relation to the gate to source potential V_{GS} (this is because the surface potential needs to exceed the source to "lure" electrons away into the channel)
- V_T is comprised of five main components:
  - Work function difference between the gate and substrate ϕ_F(substrate) – ϕ_F(gate)
  - V_{GS} component required to change the surface potential of $2\phi_F$
  - V_{GS} needed to offset the depletion region charge
  - V_{GS} needed to offset charges trapped in the gate oxide
  - V_{GS} component accounted for threshold adjustment implant

MOSFET Threshold Voltage Components

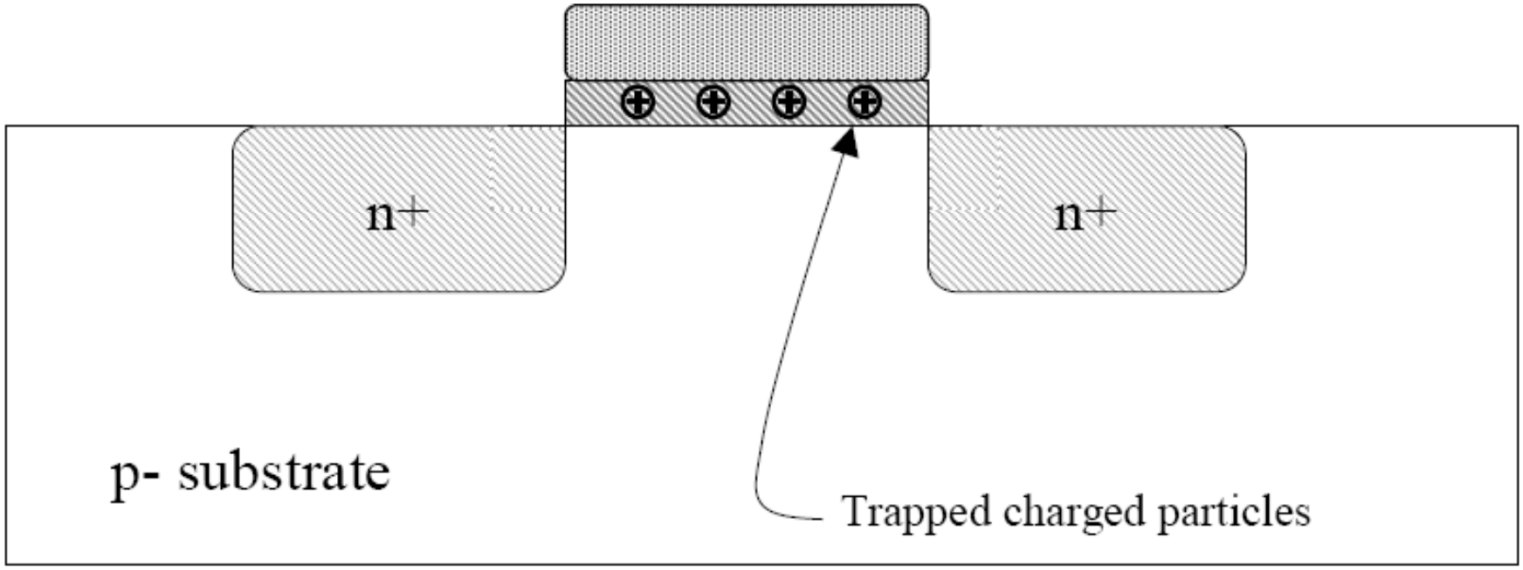


MOSFET Threshold Voltage Components

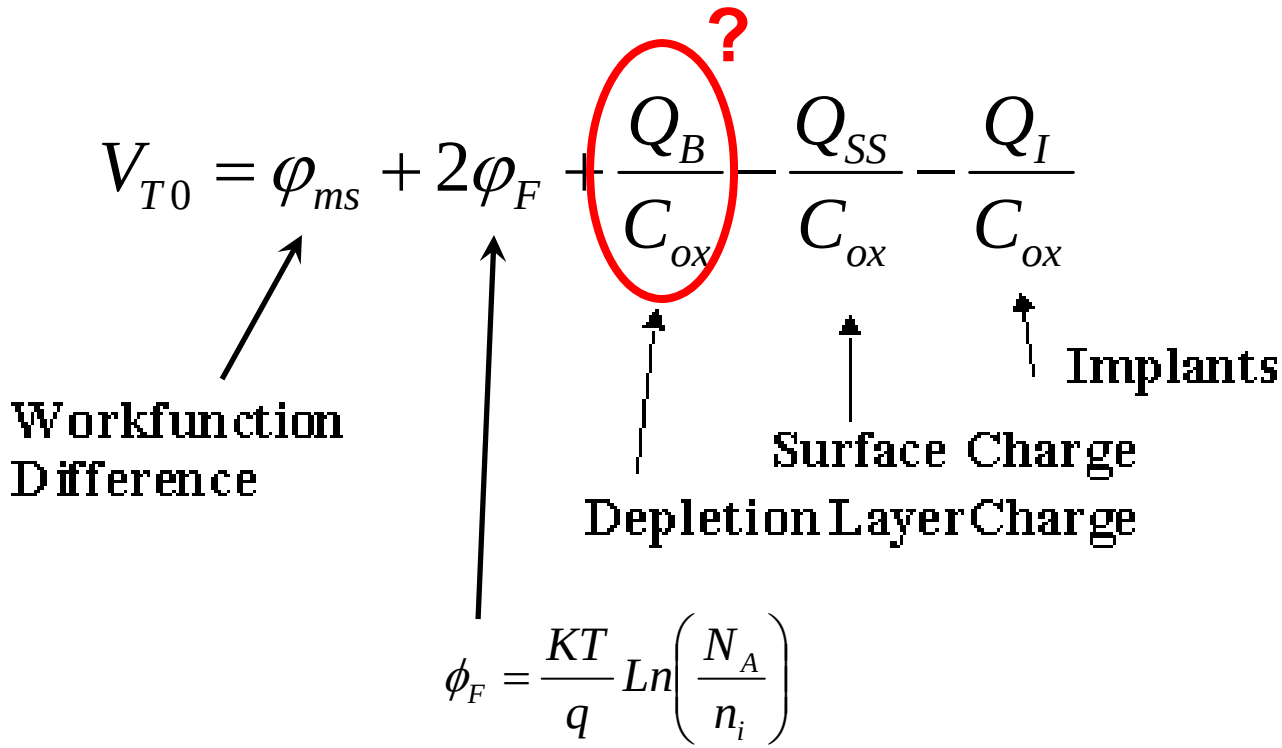
- Charged particles can get trapped in the gate oxide.
- These particles increase the V_T of the device by:

$$\frac{Q_{ox}}{C_{ox}}$$

where Q_{ox} = quantity of charge trapped in gate ox.



MOSFET Threshold Voltage Components



MOSFET Threshold Voltage Components

- The depletion region thickness (the thickness of the displaced charge):

$$X_{dm} = \sqrt{\frac{2\varepsilon_{Si}\left|-2\phi_F\right|}{qN_A}}$$

- Thus the quantity of charge per unit gate area displaced is:

$$Q_B = X_{dm} \times qN_A = \sqrt{2qN_A\varepsilon_{Si}\left|-2\phi_F\right|}$$

- This quantity of charge is being displaced by the gate potential with the gate oxide acting as the dielectric of a capacitor. Let C_{ox} be the capacitance per unit area of the gate.

$$Q = CV \qquad V = \frac{Q}{C_{ox}}$$

- Therefore the component of V_T due to displaced depletion charge (V_B) is:

$$V_B = \frac{\sqrt{2qN_A\varepsilon_{si}\left|-2\phi_F\right|}}{C_{ox}}$$

MOSFET Threshold Voltage Components

$$V_{T0} = \varphi_{ms} + 2\varphi_F + \frac{\sqrt{2qN_A\varepsilon_{si}\left|2\varphi_F\right|}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}}$$

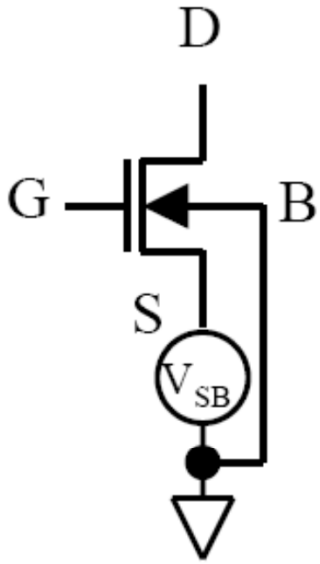
Where: $\varphi_F = \frac{KT}{q} Ln\left(\frac{N_A}{n_i}\right)$ **and** $C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$

But what happen Bulk and Source are at different potential?

Body Effect

- The body effect in a MOSFET occurs when the bulk is at a different potential than the source. For an N-device this would be when $V_{SB}>0$
- Consider the case were the source is at V_{SB}, but the bulk is held at 0V.
  - For the channel to invert V_G has to exceed $V_{SB} + V_T$.
  - This means the surface potential under the gate is:

$$-2\phi_F + V_{SB}$$



- Thus the depletion depth is:

$$X_{dm} = \sqrt{\frac{2\varepsilon_{Si}\left|-2\phi_F + V_{SB}\right|}{qN_A}}$$

- Thus the V_B component of V_T increases α to V_{SB}

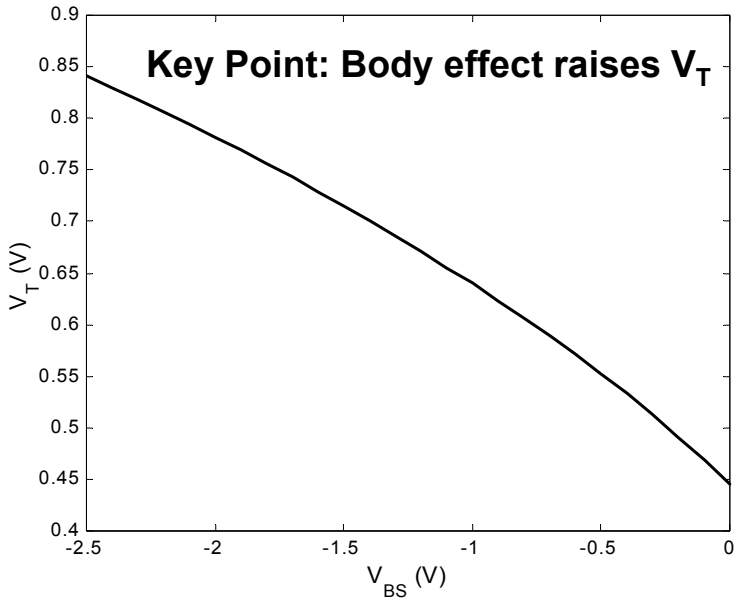
$$V_B = \frac{\sqrt{2qN_A\varepsilon_{si}\left|-2\phi_F + V_{SB}\right|}}{C_{ox}}$$

Body Effect

- Since most terms of V_T are purely a function of the device fabrication it is customary to to write the equation in a form that emphasizes the V_{SB} component.

$$V_T = V_{T0} + \gamma\left(\sqrt{|2\varphi_F - V_{BS}|} - \sqrt{|2\varphi_F|}\right)$$

Where $\gamma = \dfrac{\sqrt{2qN_A\varepsilon_{Si}}}{C_{OX}}$ and is called the body effect coefficient.



❑ Why does this matter?

- **In a stack (such as NMOS in a NAND gate) the sources of higher devices in the stack do not equal 0V due to resistance of the lower transistors - this results in lower current drive (lower Ids) due to higher apparent V_T**
- **Single polarity pass gates can only bring the drain to V_{DD}-V_T**
- **Body bias can be purposely created to lower standby power by modulating I_{OFF}**

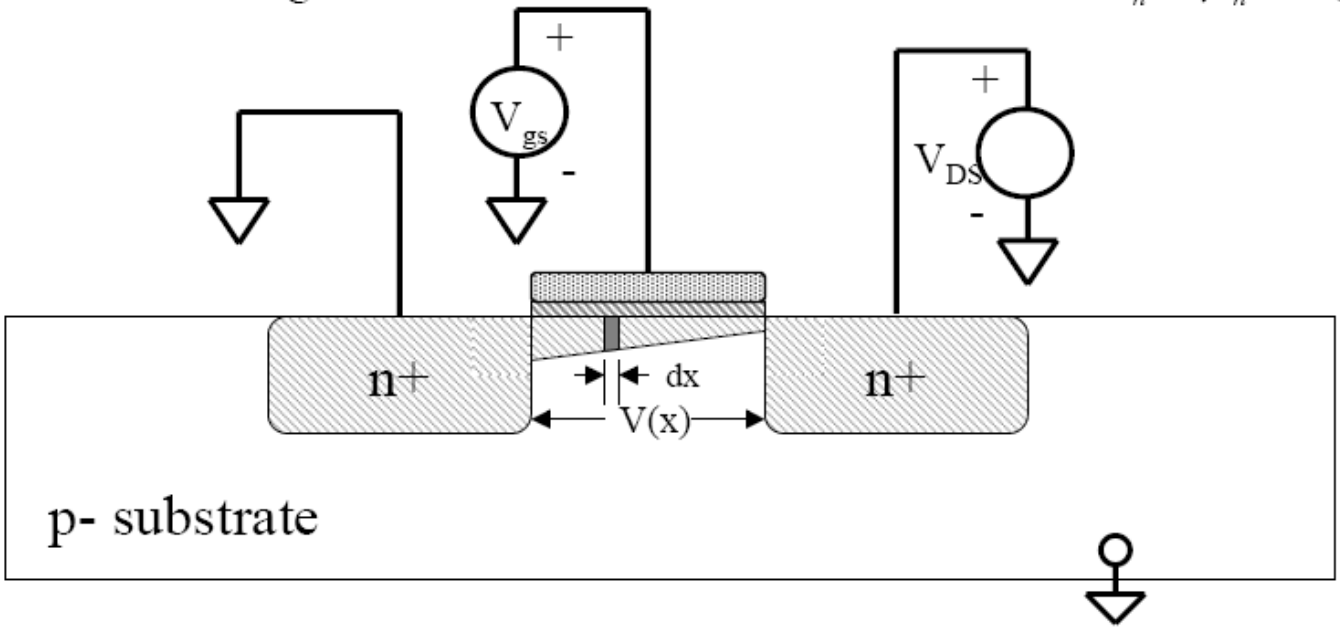
Current Voltage Relation

- Let V(x) represent the voltage of the inverted channel as a function of x (position in channel length from source) V(0) = 0V, V(L) = V_{DS}

- Then the charge density in the channel as a function of x is:

$$Q(x) = C_{ox}\left(V_{GS} - V(x) - V_T\right)$$

- Current = Charges/unit time = $I_D = -v_n Q(x) W$ where $v_n = \mu_n \times E(x)$



Current Voltage Relation

- However: $E(x) = \frac{dV}{dx}$ and $v = \mu_n \frac{dV}{dx}$

- Therefore: $I_D dx = \mu_n C_{ox} W (V_{GS} - V - V_T) dV$

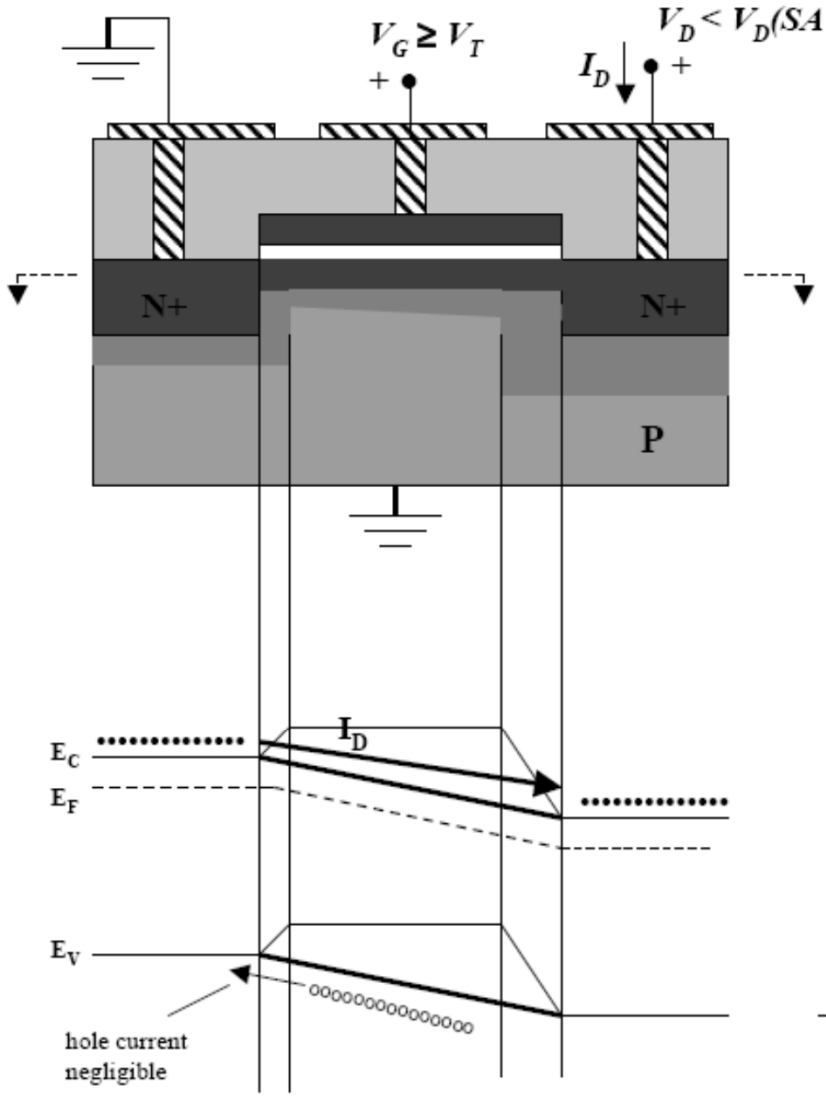
$$\int_0^L I_D dx = \int_0^{V_{DS}} \mu_n C_{ox} W (V_{GS} - V - V_T) dV$$

- Which yields: $I_D = \mu_n C_{ox} \frac{W}{L} \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right]$

L = channel length
W = channel width (perpendicular to L)
μ_{eff} = surface mobility [~ 500 $cm^2/V \cdot s$ for electrons in Si at 300K, ~ 150 $cm^2/V \cdot s$ for holes]
C_{OX} = gate capacitance = $\varepsilon_{OX} A / t_{OX}$

Which is valid for values of $V_{DS} < V_{GS} - V_T$ *(i.e. Linear Region)*

Device Operation: Linear Region



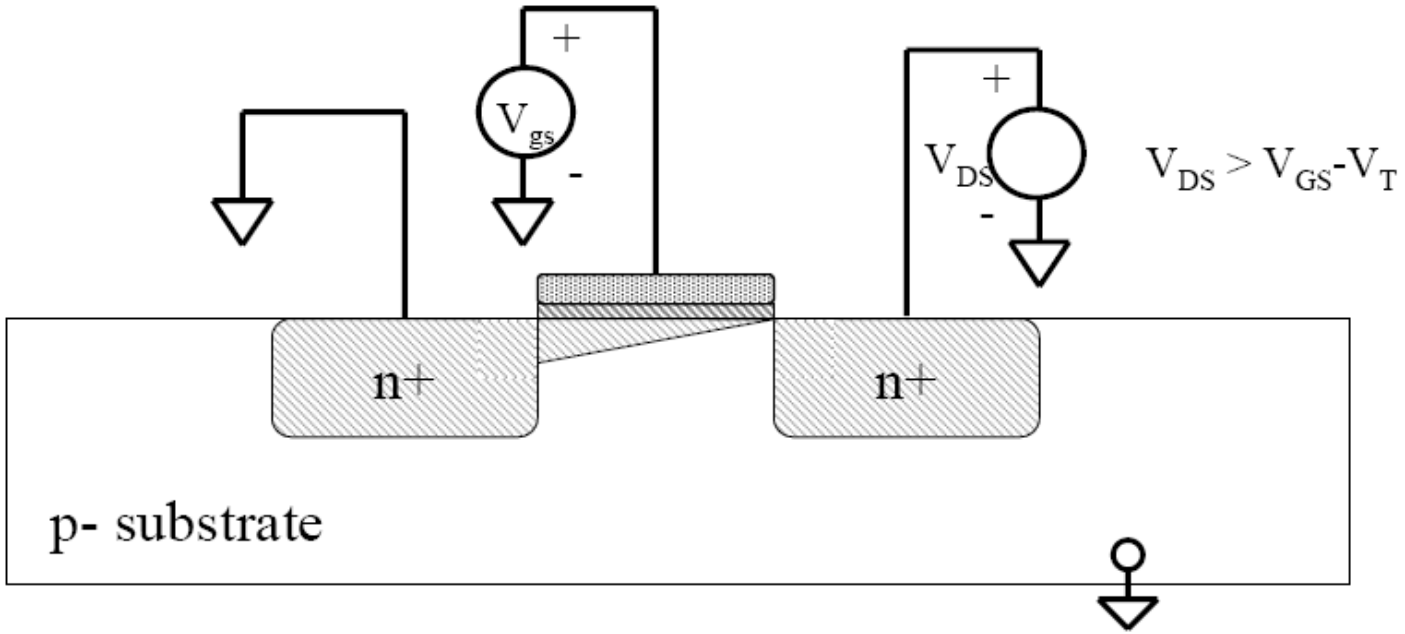
N-type channel, behaves like a resistor (ohmic)

Device Operation: Saturation

- $V_{GD} = V_{GS} - V_{DS}$; so as V_{DS} increases V_{GD} will no longer exceed V_T, thus the charge density in the channel near the drain will decrease.

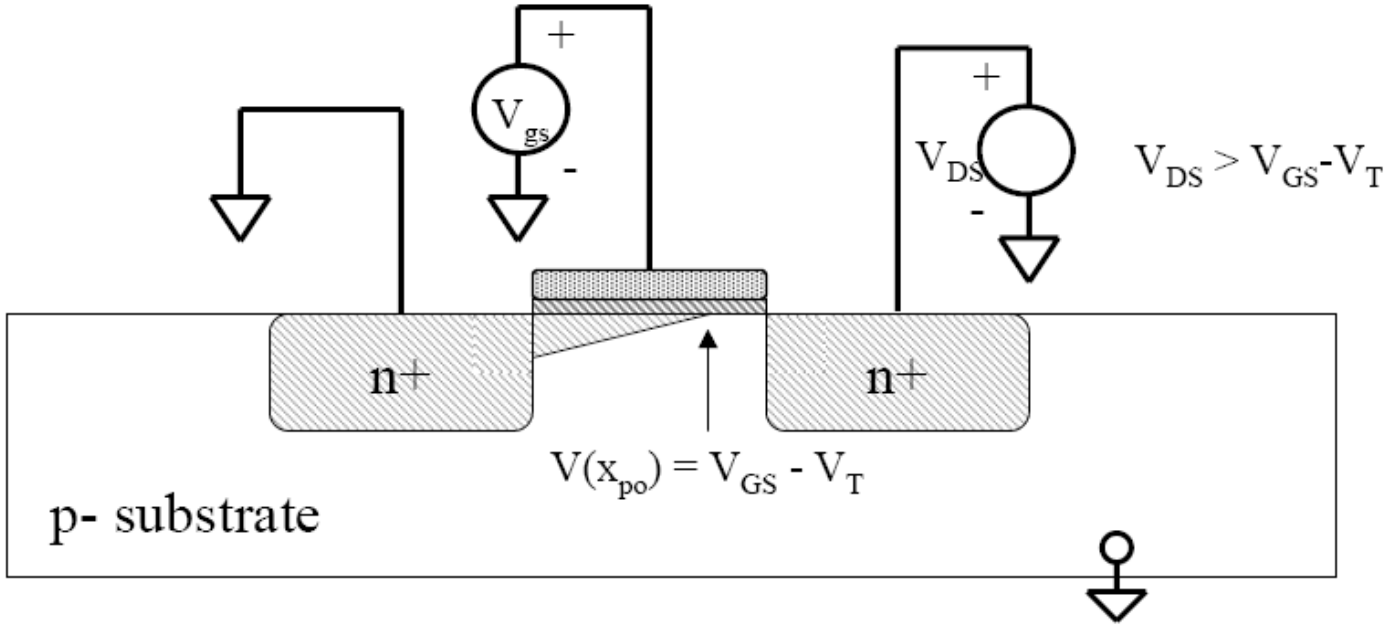
- If $V_{DS} = V_{GS} - V_T$ then $V_{GD} = V_T$. At this operating point the charge density in the channel would diminish to zero right at the drain.

- When $V_{DS} = V_{GS} - V_T$ the device is transitioning to saturation mode.



Device Operation: Saturation

- As V_{DS} increases beyond $V_{GS} - V_T$ the charge density in the channel reaches zero prior to reaching the drain. At this point mobile charges are injected into the depletion region and swept to the drain.
- The early termination of the channel is termed "pinch off".
- I_{DS} stops increasing with V_{DS}, and the device is said to be "saturated".

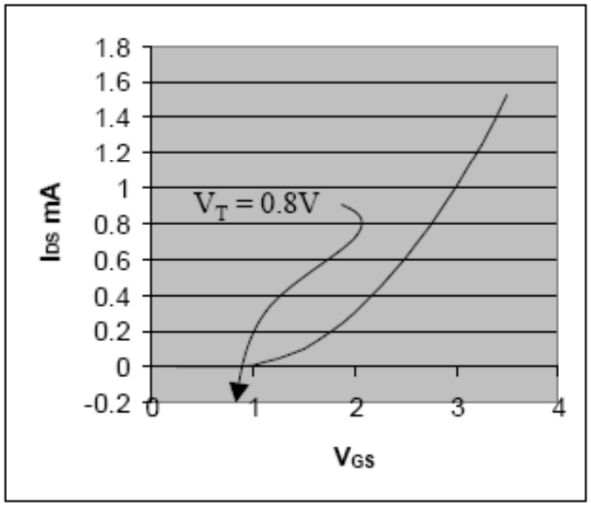


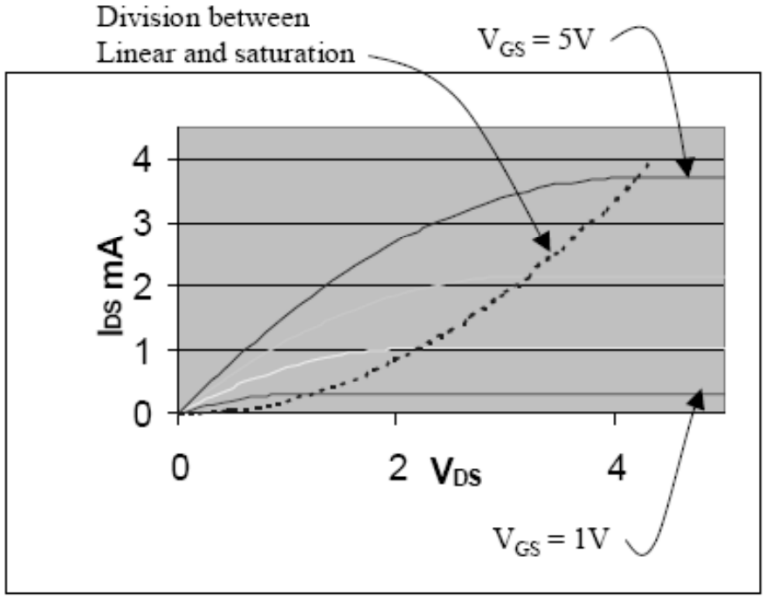
Device Operation: Saturation

- Since I_{DS} does not increase with increasing V_{DS} beyond $V_{DS} = V_{GS}$-V_T one can find the equation for I_{DS} in saturation by substituting $V_{DS} = V_{GS}$-V_T into the I_{DS} equation for linear mode:

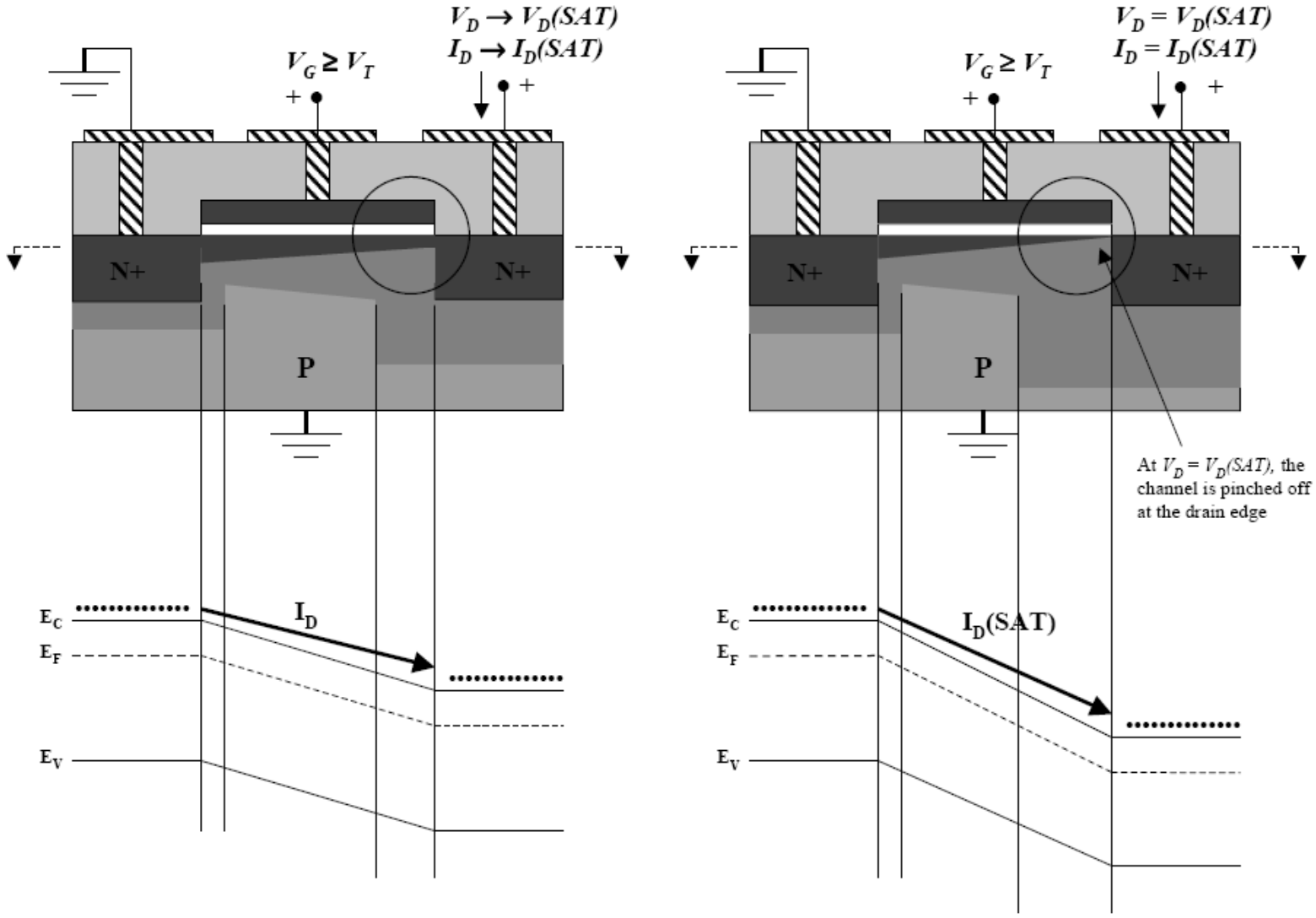
$$I_{DS} = \mu_n C_{ox} \frac{W}{L}\left[(V_{GS} - V_T)(V_{GS} - V_T) - \frac{(V_{GS} - V_T)^2}{2}\right]$$

$$I_{DS} = \mu_n C_{ox} \frac{W}{2L}(V_{GS} - V_T)^2$$

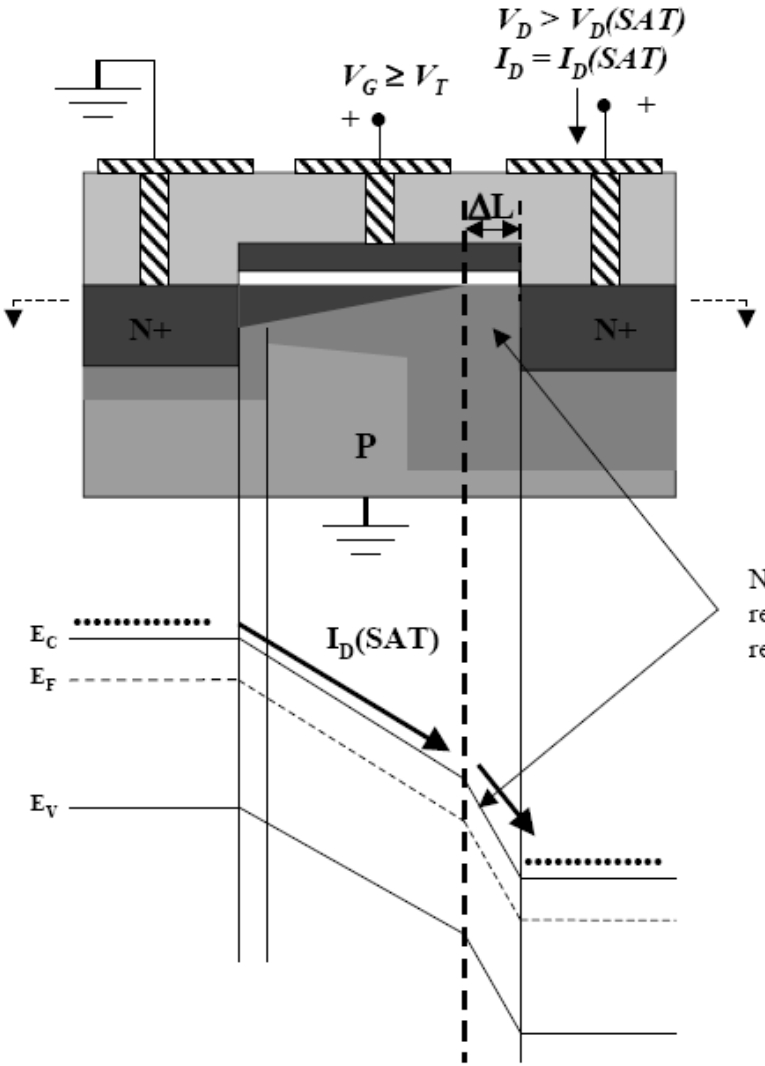




Linear into Saturation



Saturation Region



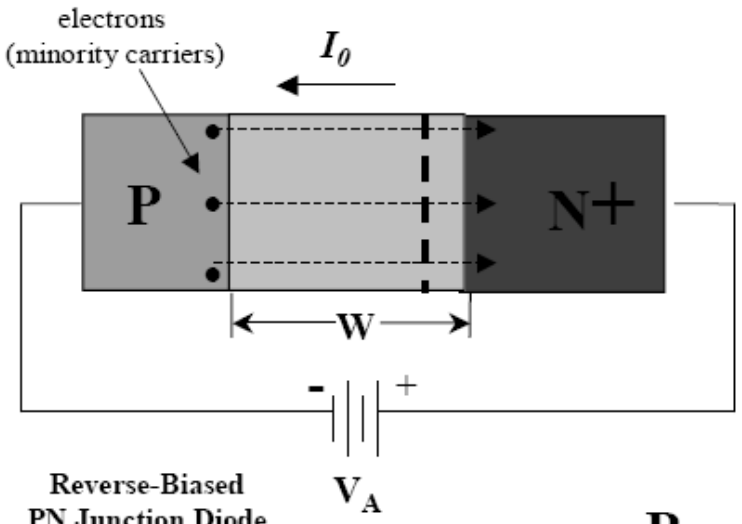
$V_D(SAT) \approx V_G - V_T$

No longer behaving like a simple resistor. Now looks like a depletion region being fed by the channel.

Defined by:

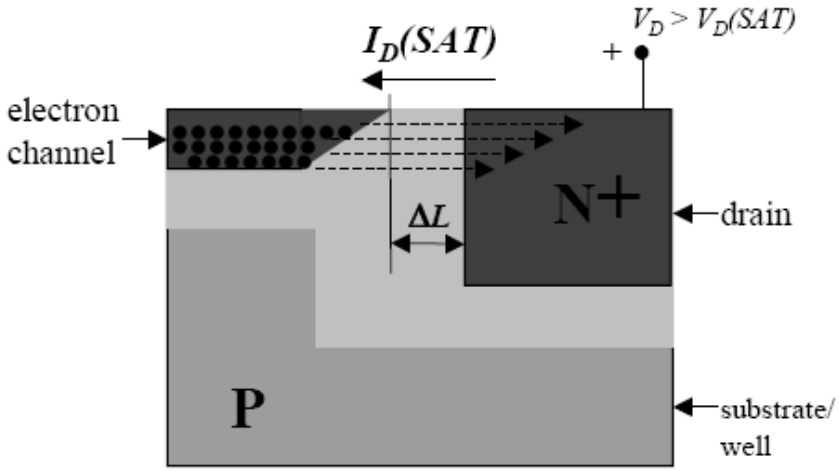
1. **Pinch-off of inversion region due to high drain field**
2. **Carriers no longer confined to channel region in the pinch-off region**

Saturation Region

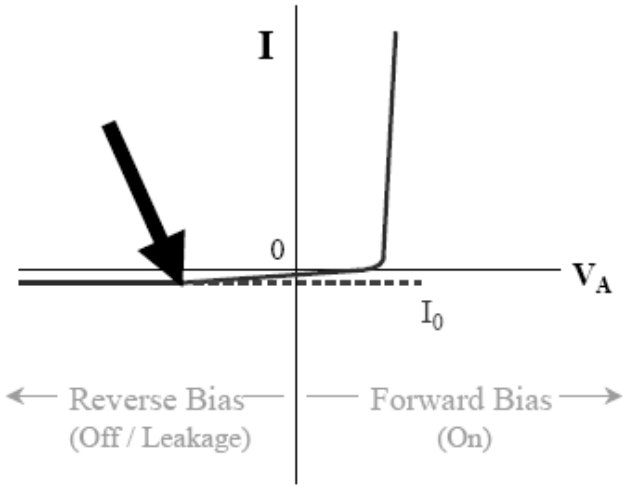


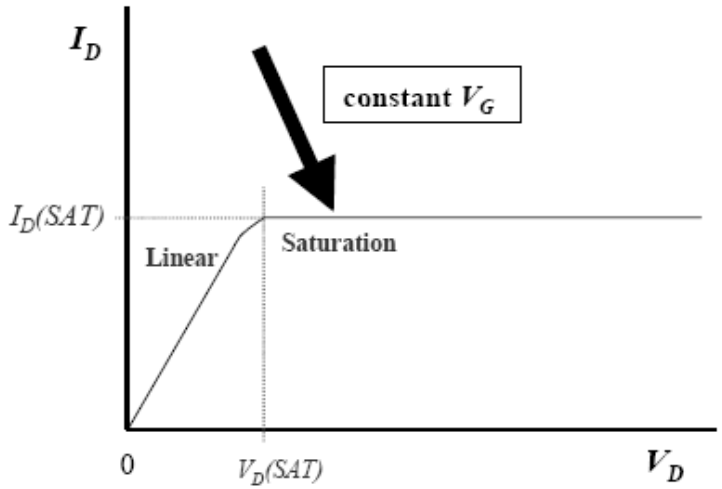
Reverse-Biased
PN Junction Diode

P

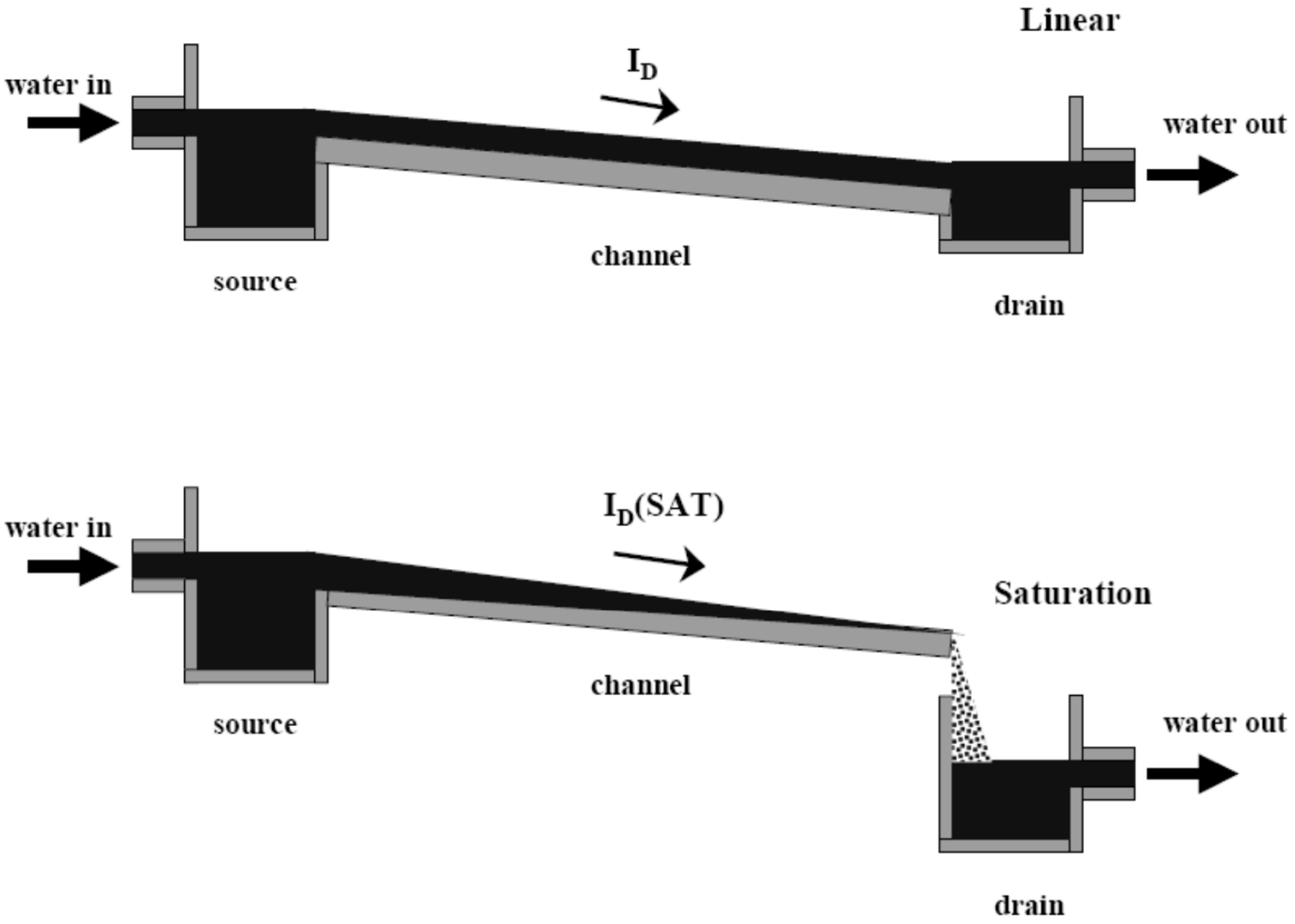


MOSFET in
Saturation



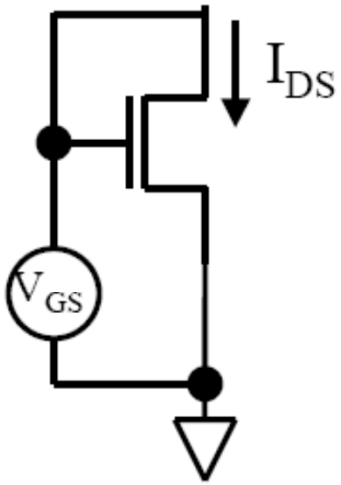


Saturation Region Analogy



MOSFET Parameter Measurement

- Consider the following configuration:
  - $V_{DS} > V_{GS} - V_T$ (device always in cutoff or saturation).
- For $V_{GS} > V_T$, I_{DS} will equal:



$$I_{DS} = \mu_n C_{ox} \frac{W}{2L}(V_{GS} - V_T)^2 \quad \rightarrow \quad I_{DS} = \frac{k_n}{2}(V_{GS} - V_T)^2$$

where

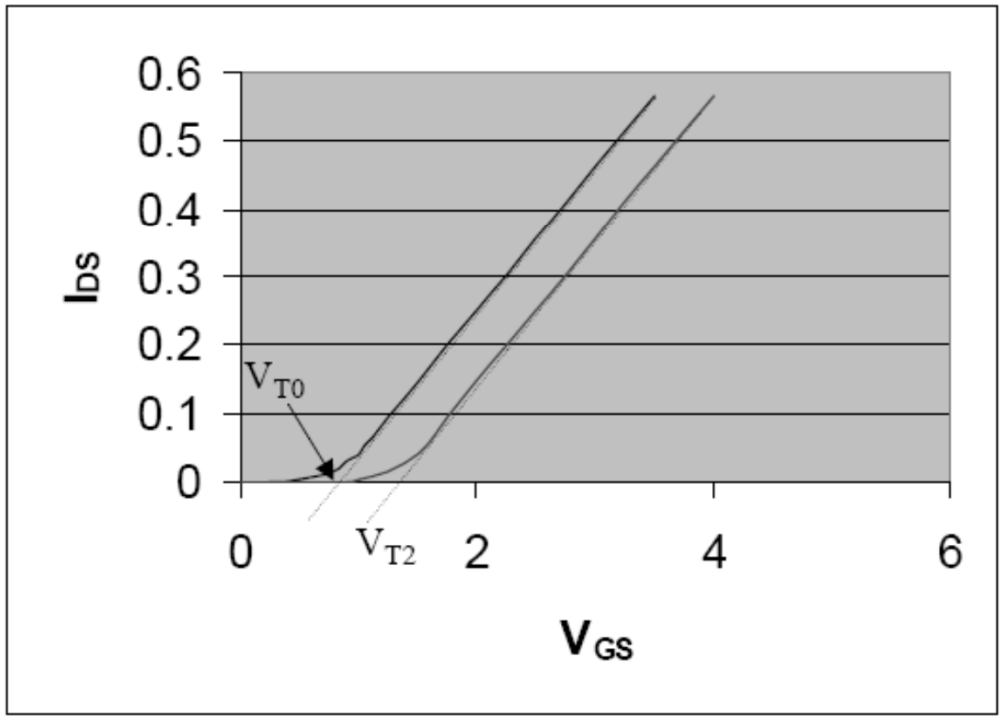
$$k_n = \mu_n C_{ox} \frac{W}{L}$$

- Taking the square root of each side yields

$$\sqrt{I_{DS}} = \sqrt{\frac{k_n}{2}}(V_{GS} - V_T)$$

MOSFET Parameter Measurement

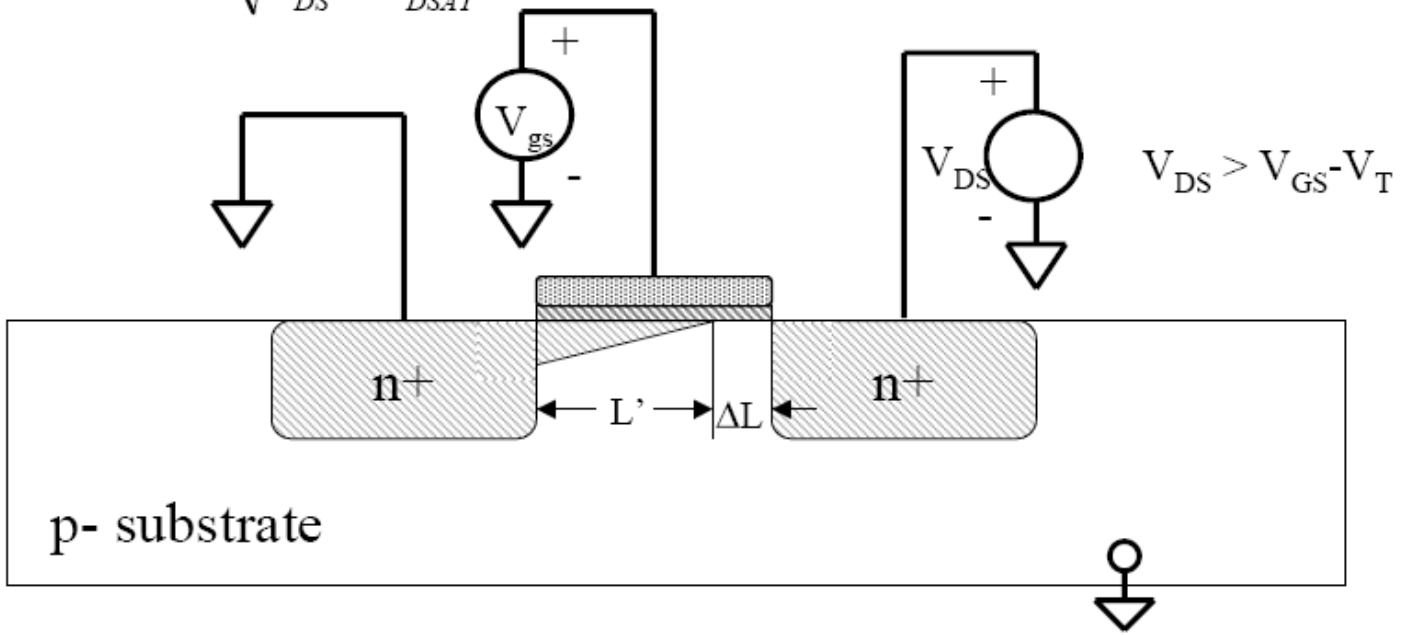
- Now the sqrt(I_{DS}) can be plotted vs V_{GS}:
- Two curves are shown here. One for a V_{SB} of 0V, and one for a V_{SB} of 2V.
- Slope of line = $\sqrt{\frac{k_n}{2}}$
- V_T can be determined where the extension of curve intersects x-axis
- Difference in V_T's can tell you the body-effect coefficient



$$\gamma = \frac{V_{T2} - V_{T0}}{\sqrt{|2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}}$$

Channel Length Modulation

- Our previous view of saturation is too simple. I_{DS} *will* still have some V_{DS} dependence for V_{DS} values greater than V_{GS}-V_T
- As V_{DS} increases beyond V_{GS}-V_T more and more of the channel becomes "pinched off". Thus the effective channel length (L') is reduced by ΔL.
- This ΔL is proportional to: $\sqrt{V_{DS} - V_{DSAT}}$; However one will discover that $\frac{1}{L - k\sqrt{V_{DS} - V_{DSAT}}}$ is a fairly linear function. Therefore …



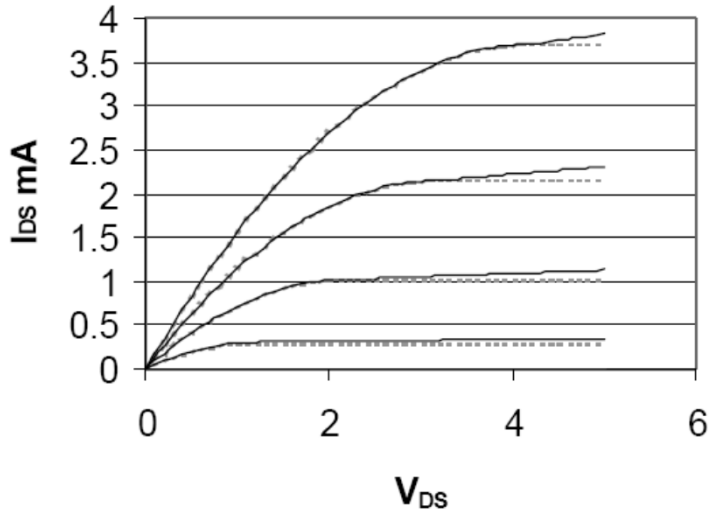
Channel Length Modulation

- The effect of channel length modulation is typically modeled with an empirical linear factor λ.

- Thus the equation for I_{DS} in saturation becomes:

$$I_{DS} = \mu_n C_{ox} \frac{W}{2L}(V_{GS} - V_T)^2(1 + \lambda V_{DS})$$

where λ = "channel length modulation factor"



Device Operation: I-V curves

$$I_{DS} = K_n' \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right](1 + \lambda V_{DS})$$

$$I_{DS} = \frac{K_n'}{2}\frac{W}{L}(V_{GS} - V_T)^2(1 + \lambda V_{DS})$$

$V_{DS} < V_{GS} - V_T$

$V_{DS} > V_{GS} - V_T$