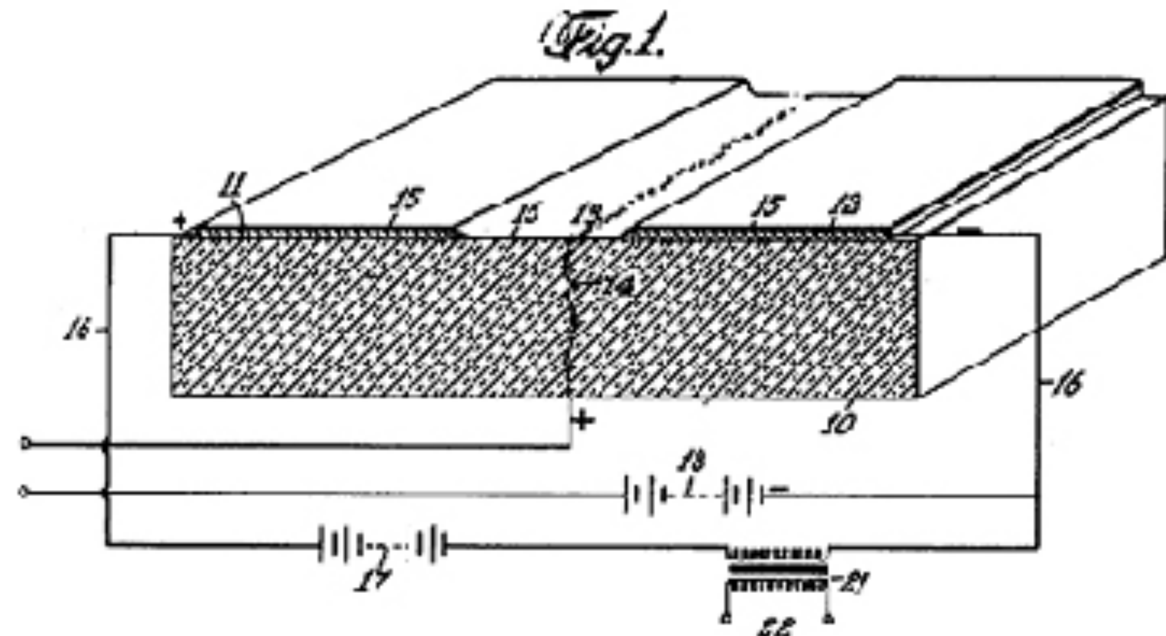
MOS Transistors

Prof. Krishna Saraswat

Department of Electrical Engineering
Stanford University
Stanford, CA 94305
saraswat@stanford.edu

1930: Patent on the Field-Effect Transistor



Julius Lilienfeld filed a patent describing a three-electrode amplifying device based on the semiconducting properties of copper sulfide. He did not demonstrate the device experimentally

1960 - MOS Transistor Demonstrated

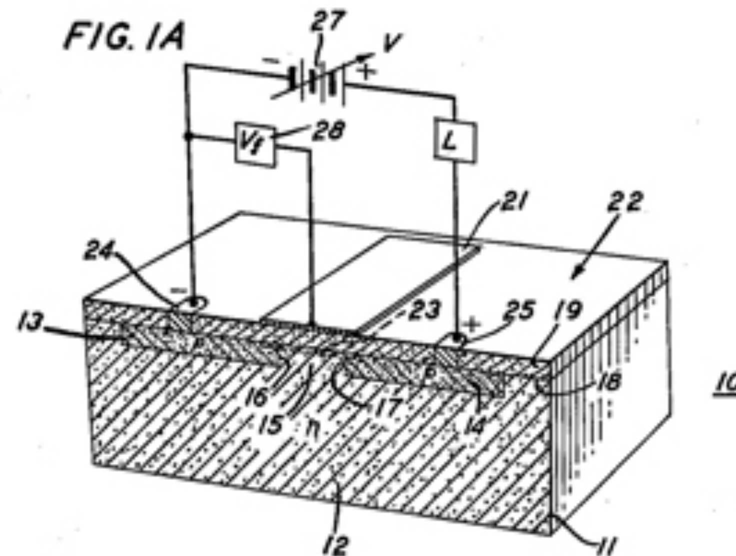
Dawon Kahng

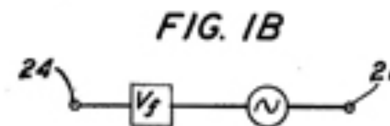
John Atalla

Aug. 27, 1963 DAWON KAHNG **3,102,230**

ELECTRIC FIELD CONTROLLED SEMICONDUCTOR DEVICE

Filed May 31, 1960





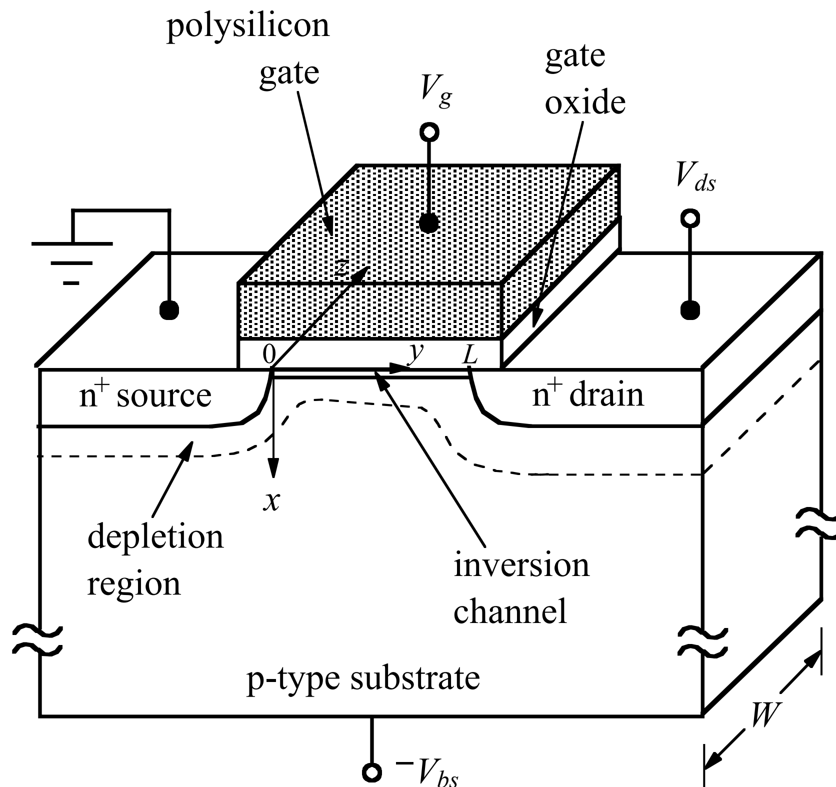
John Atalla and Dawon Kahng at Bell demonstrate the first successful MOS field-effect amplifier.

Outline

- Current-voltage characteristics
- Scaling and short channel behavior
- Future MOS technologies

MOS Transistor

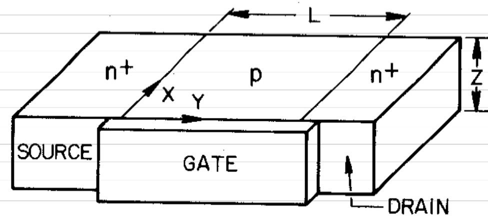
The theory developed for the MOS capacitor can be extended directly to the MOS Field-Effect-Transistor (MOSFET) by considering the following structure.

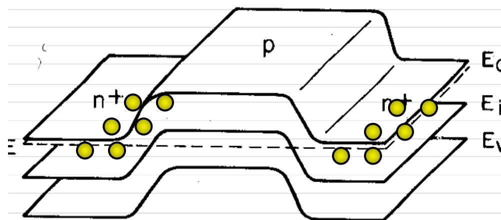


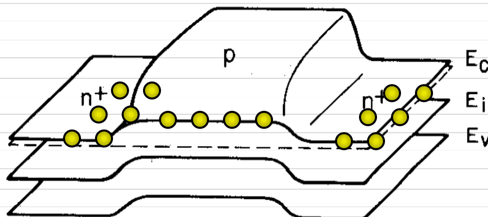
Enhancement mode MOSFET

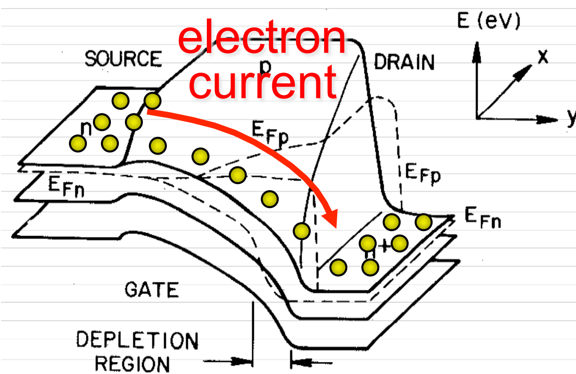
- V_G provides control of surface carrier densities: Q=CV
- For V_G << V_T, the structure consists of two diodes back to back and only leakage currents flow.
- When V_G is only slightly below V_T a depletion region will be formed.
- For V_G > V_T, an inversion layer, i.e., a conducting channel, exists between source and drain and current will flow.
- For any further increase in V_G the excess potential will result in an increase in the electron density in the channel

NMOS Transistor 3D Band Diagram









Source: Sze (1981)

N-channel enhancement mode MOSFET, $V_T > 0$

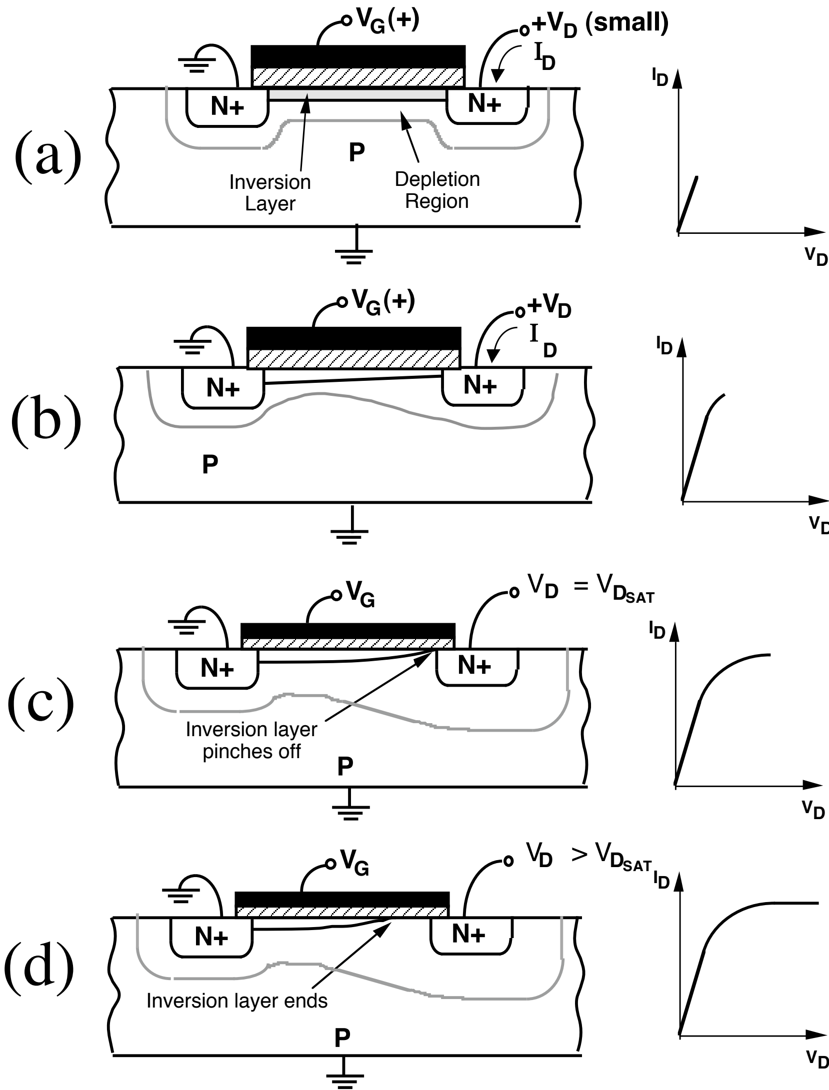
P-channel enhancement mode MOSFET, $V_T < 0$

$\mathbf{V_G = V_D = 0}$ no carriers in the channel

$\mathbf{V_G > 0,\ V_D = 0}$ carriers in the channel but no movement between source and drain

$\mathbf{V_G > V_T,\ V_D > 0}$ electrons flow from source to drain

Variation of Drain Current with V_D



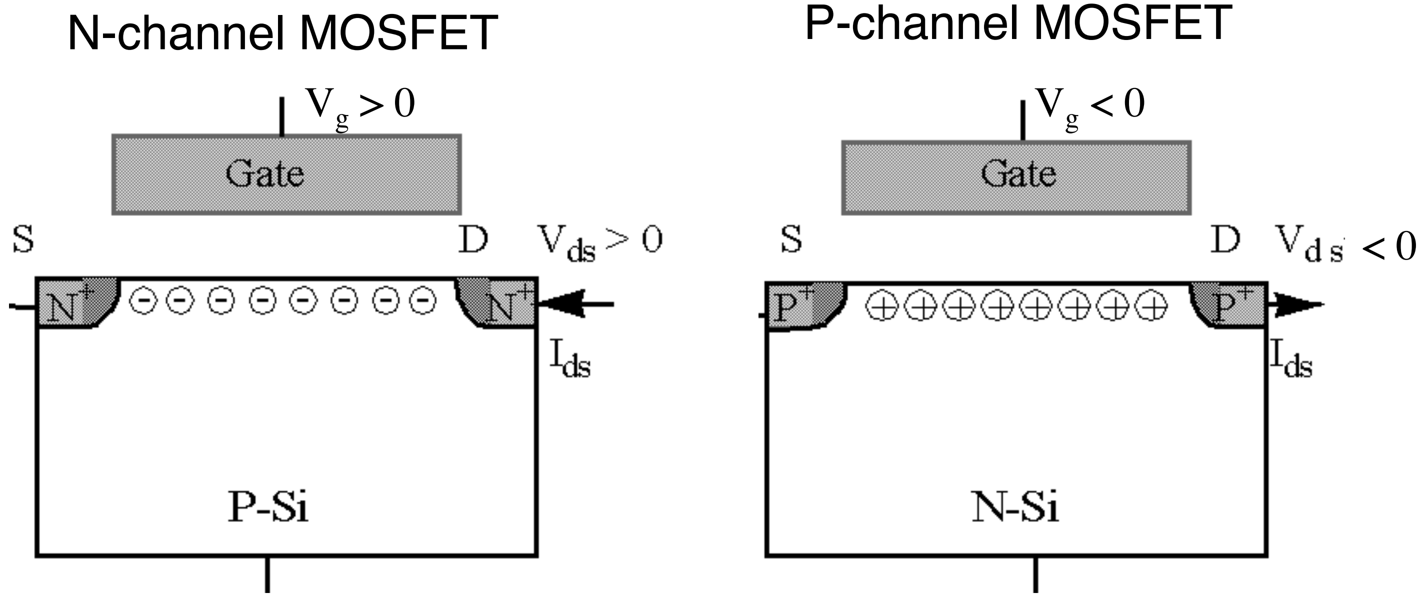
Linear region
$I_D \alpha V_D$

Before pinch off

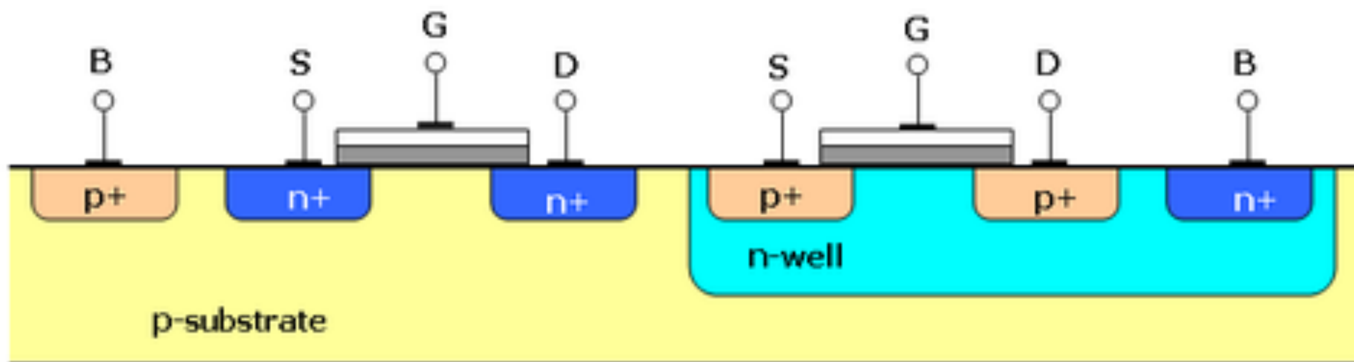
Inversion layer pinch off
Onset of saturation

Saturation region I_D constant with V_D. (Assumption valid only for long channel)

Complementary MOS (CMOS) Technology



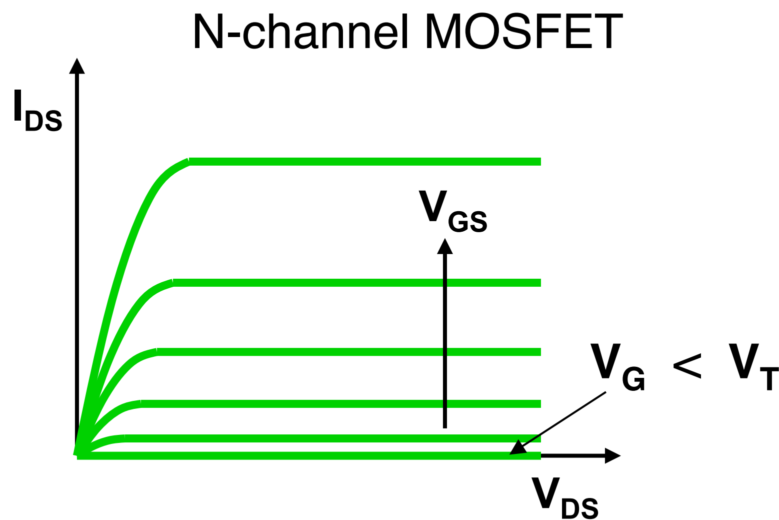
All the polarities for P-channel MOSFET are opposite to that of N-channel MOSFET



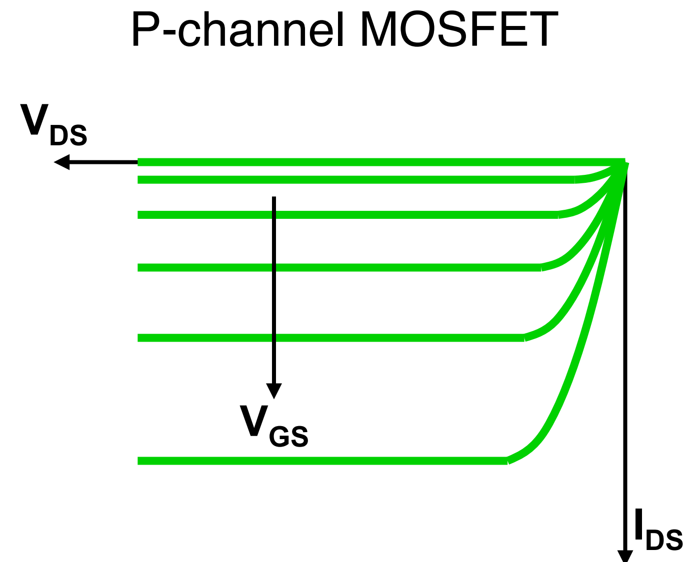
Current-voltage characteristics

- Increase in V_G will result in an increase in the electron density in the channel and thus the drain current.
- After pinch off drain current saturates.

N-channel MOSFET

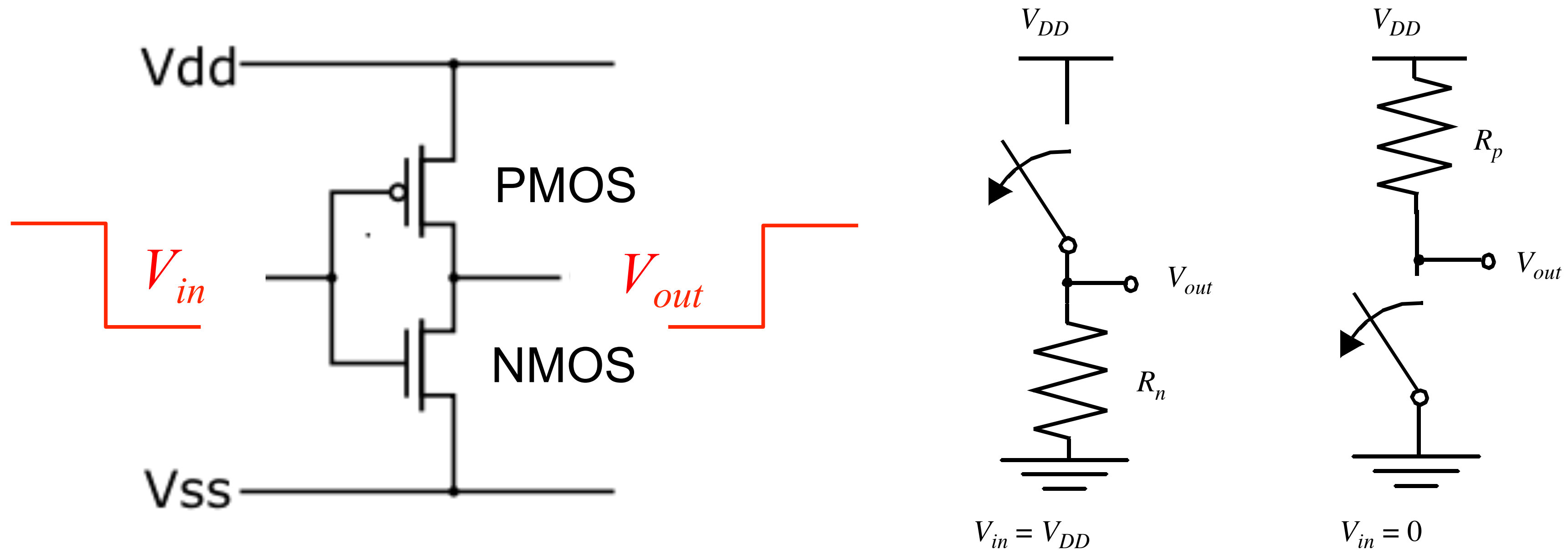


P-channel MOSFET



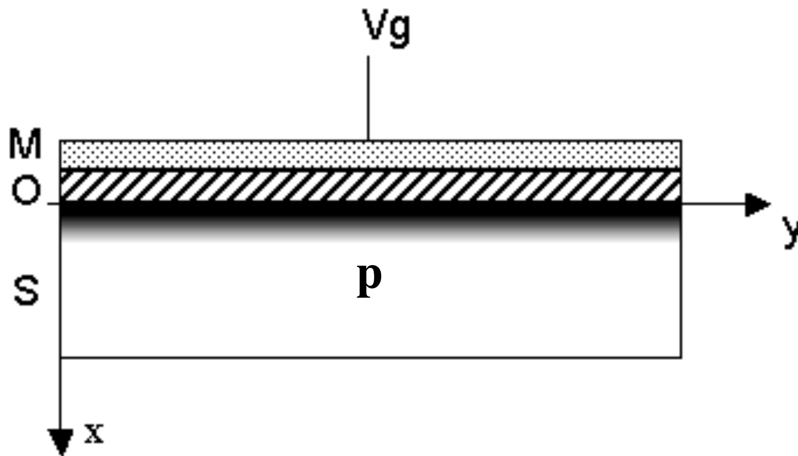
- For P-channel MOSFET, all of the polarities are reversed and the inversion layer exists for $V_G < V_T$

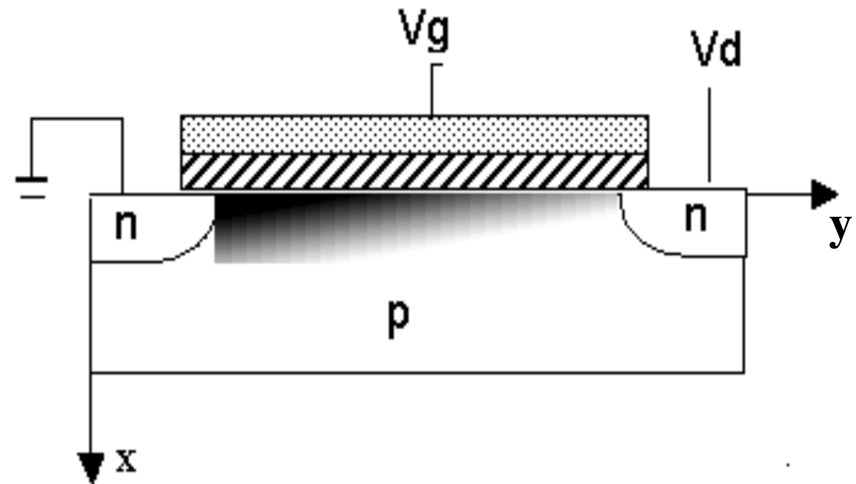
CMOS Inverter



- For $|V_g| < |V_t|$ the transistor is off represented by open circuit
- For $|V_g| > |V_t|$ the transistor is on represented by a resistor
- Output is an inverted form of input waveform
- CMOS inverter is the most important building block of modern logic circuits
- What is the power dissipation in this circuit?

Gradual Channel Approximation





Vertical field E_x → inversion layer charge

Lateral field E_y → flow of carriers from source to drain

$$\frac{\partial E_x}{\partial x} = \frac{\rho(x)}{\varepsilon_{si}}$$

$$C_{ox}V_g = -(Q_i + Q_d)$$

inversion (→ Q_i), depletion (→ Q_d)

$$Q_i = -C_{ox}(V_g - V_t)$$

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} = \frac{\rho(x,y)}{\varepsilon_{si}}$$

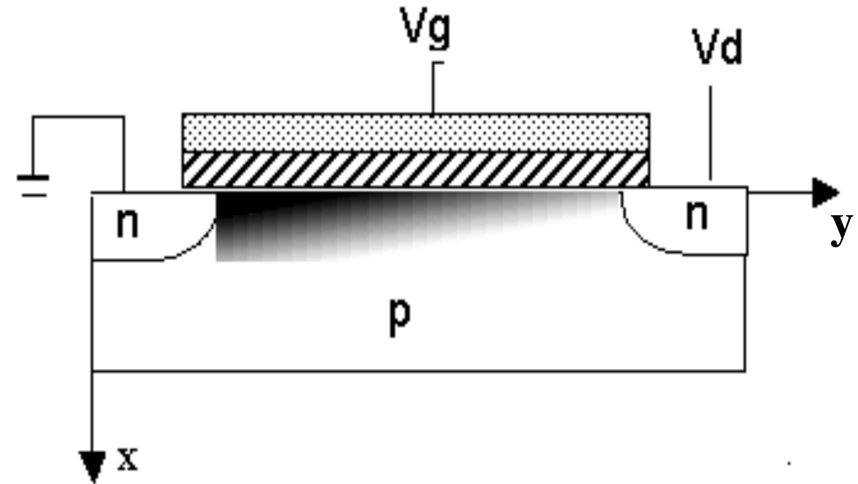
$$\frac{\partial E_x}{\partial x} >> \frac{\partial E_y}{\partial y}$$

$$Q_i(y) \approx -C_{ox}(V_g - V(y) - V_t)$$

Current – Voltage Dependence

$$J_e(y) = qD_n \frac{dn(y)}{dy} + q\mu_n n(y)\mathrm{E}_y$$

Diffusion Drift



Charge/Area

$$Q_i(y) = -qn_s(y) \approx -C_{ox}(V_g - V(y) - V_t)$$

$$\frac{dQ_i(y)}{dy} \approx C_{ox}\frac{dV(y)}{dy}$$

$$n_s(y) = \int_0^\infty n(x,y)dx$$

Sheet Charge density

$$J_e = q\left[D_n \frac{d\left(-Q_i(y)/q\right)}{dy} + \mu_n\left(-Q_i(y)/q\right)\left(-\frac{dV(y)}{dy}\right)\right]$$

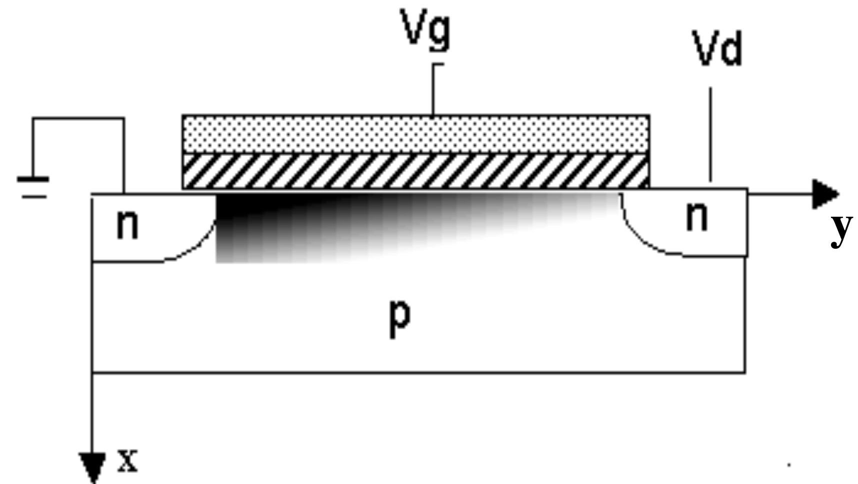
Current / Width (z-direction)

$$J_e = -D_n\left[\frac{dQ_i(y)}{dy}\right] + \mu_n Q_i(y)\left[\frac{dV(y)}{dy}\right] \qquad (1)$$

Current vs Voltage

$$J_e = -D_n \left[\frac{dQ_i(y)}{dy} \right] + \mu_n Q_i(y) \left[\frac{dV(y)}{dy} \right]$$

Diffusion   Drift



Note that E_y and $Q_i(y)$ are negative

When $V_{GS}>V_T$ & $V_{DS}>V_T$ diffusion current is negligible

$$J_e = \mu_n Q_i(y) \frac{dV(y)}{dy} \qquad Q_i(y) \approx -C_{ox}(V_g - V(y) - V_t)$$

$$J_e \int_0^L dy = -\mu_n C_{ox} \int_0^{V_{DS}} (V_{GS} - V - V_t) dV$$

$$J_D = -\frac{W}{L} \mu_n C_{ox} \left[(V_{GS} - V_t) V_{DS} - \frac{1}{2} V_{DS}^2 \right] \qquad (2)$$

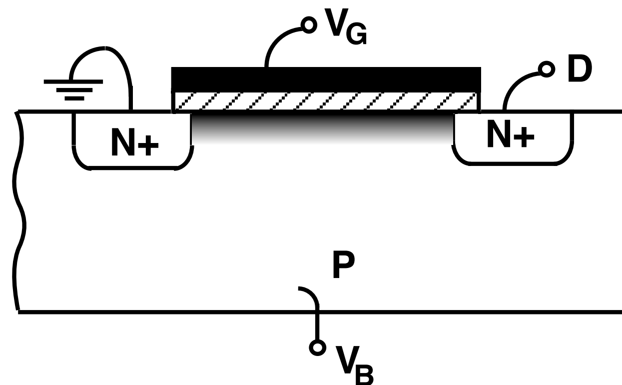
Threshold Voltage

- For the case where backside is grounded (V_B = 0) V_T is given by the equation,

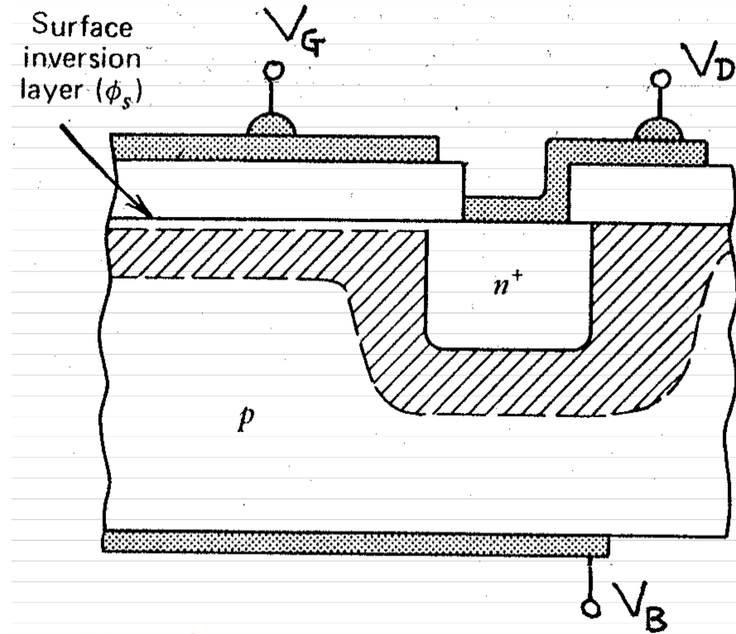
$$V_T = V_{FB} + \frac{t_{ox}}{\varepsilon_{ox}} \sqrt{2\varepsilon_s q N_a \left(-2\phi_p\right)} - 2\phi_p \qquad (3)$$

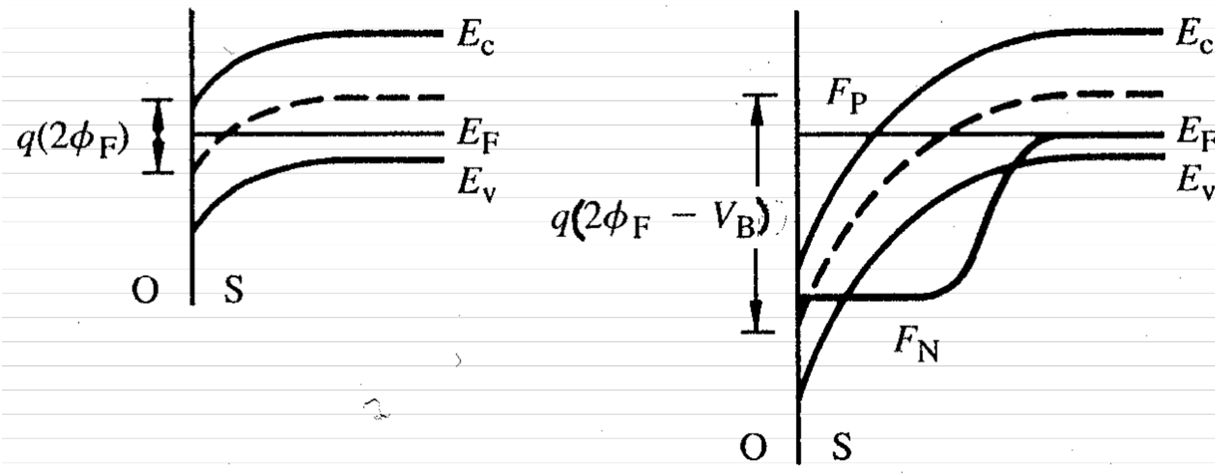
- In many circuit applications backside is biased. For finite value of V_B

$$V_G = V_{FB} + V_{ox} + \phi_s - \phi_p + V_B$$



Effect of Back Bias





Effect of Back Bias

- In a normal MOS capacitor, application of V_B will result mostly in change in V_{ox} as ϕ_s is fixed at $-\phi_p$.
- If there is a nearby n-type region (drain) which contacts with the inverted surface layer the situation changes. When the surface is inverted, there is basically a P-N junction at the surface. A reverse bias can be applied across the P-N junction.
- If V_B is zero, inversion occurs when $\phi_s = -\phi_p$.
- If $V_B < 0$, the semiconductor still attempts to invert when ϕ_s reaches $-\phi_p$. However, with $V_B < 0$ any inversion-layer carriers that do appear at the semiconductor surface migrate laterally into the source and drain because these regions are at a lower potential. Not until $\phi_s = -\phi_p - V_B$, will the surface invert and normal transistor action begin. In essence, back biasing changes the inversion point in the semiconductor from $-\phi_p$ to $-\phi_p - V_B$.

Effect of Back Bias

- An applied reverse bias between the induced surface n-region and the bulk increases the charge Q_d in the depletion region.
- Since the negative charge induced by V_G - V_B is shared between the depletion and inversion layers, an increase of the charge in the depletion layer means that there is less charge available to form the inversion layer for a given gate voltage.
- Looked at another way, more gate voltage must be applied to induce the same number of electrons in the inversion layer when there is a reverse bias.
- With reverse bias present, the surface potential at the onset of strong inversion becomes $\phi_s = -\phi_p + (V_D - V_B)$ rather than $\phi_s = -\phi_p$.

With $\boldsymbol{V_D}$ and $\boldsymbol{V_B}$ applied:

$$x_{d_{\max}} = \sqrt{\frac{2\kappa_s \varepsilon_o \left(-2\phi_p + V_D - V_B\right)}{qN_a}}$$

$$V_T = V_D + V_{FB} + \frac{t_{ox}}{\varepsilon_{ox}} \sqrt{2\, K_s \varepsilon_o q N_a \left(V_D - 2\phi_p - V_B\right)} - 2\,\phi_p \qquad (4)$$

where $C'_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} A = C_{ox} A$

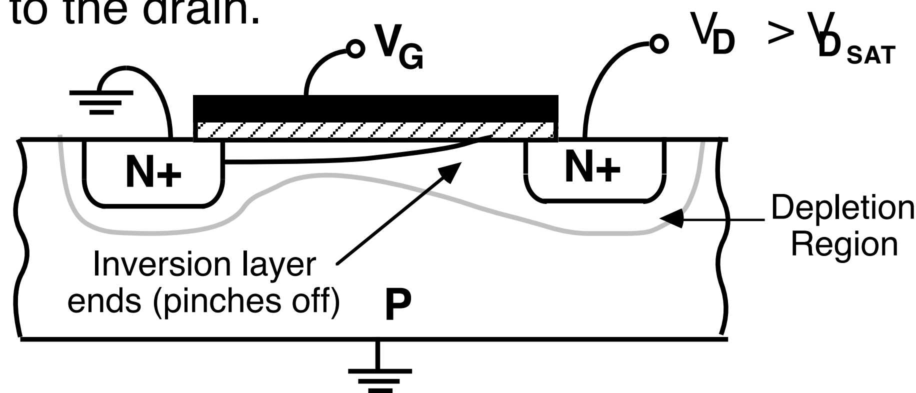
For small values of V_D and $V_B = 0$, Expression for I_D can be approximated by

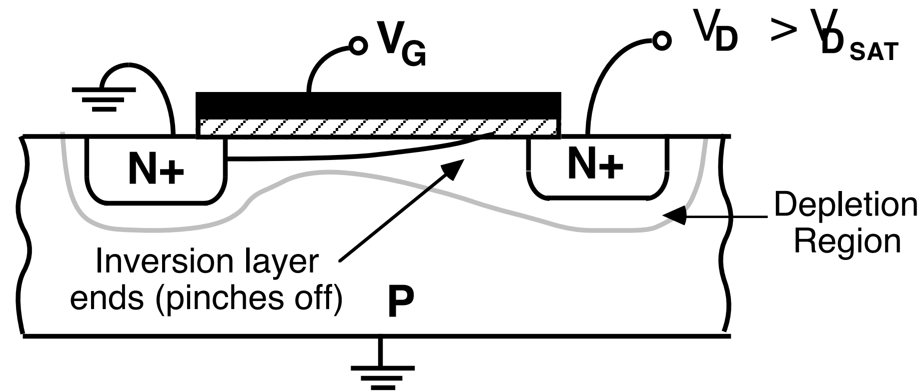
$$I_D = \frac{W}{L}\mu_n \frac{\varepsilon_{ox}}{t_{ox}}\left[(V_{GS} - V_t)V_{DS} - \frac{1}{2}V_{DS}^2\right] \approx \frac{W}{L}\mu_n \frac{\varepsilon_{ox}}{t_{ox}}(V_{GS} - V_t)V_{DS} \qquad (5)$$

This is known as LINEAR REGION.

These equations are valid only as long as an inversion layer exists all the way from source to drain (LINEAR REGION).

As $V_D\uparrow$, the effective voltage between the gate and the channel near the drain will become less than V_T. This happens when $V_G - V_T = V_D$. This value of drain voltage is called the saturation voltage V_{DSAT} (or pinch off voltage because the channel is pinched off at the drain), and for higher drain voltages, a channel will not exist all the way to the drain.





Electrons drift along in the inversion layer and are injected into the depletion region. There the high electric field pulls them into the drain.

Further increase in V_D does not change I_D (to first order), I_D is constant for $V_D > V_{DSAT}$ (SATURATION REGION).

Exact V_{DSAT} can be derived by letting $Q_I(y=L) = 0$ in equation

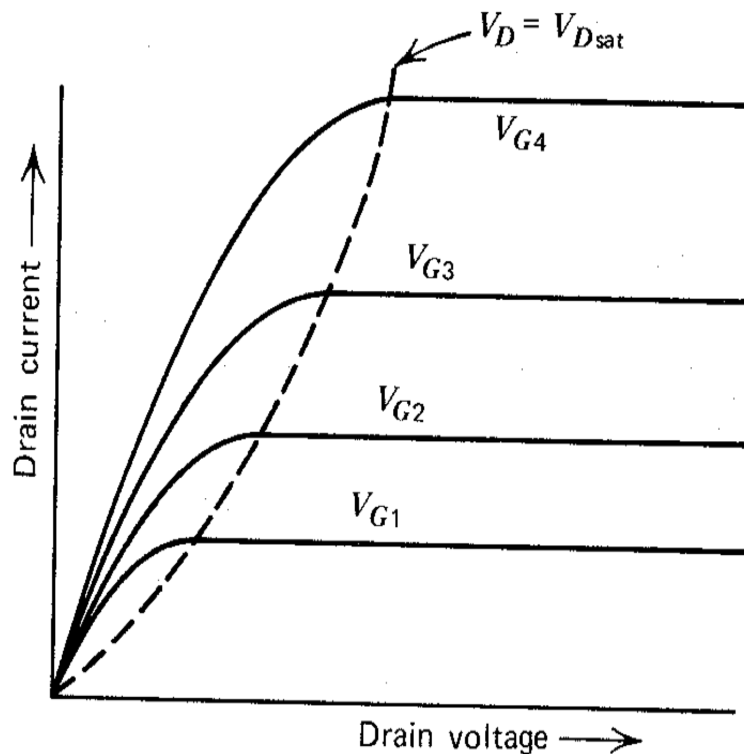
$$Q_i(y) \approx -C_{ox}(V_g - V(y) - V_t)$$

$$Q_i(L) \approx -C_{ox}(V_g - V_{D,Sat} - V_t) = 0$$

$$V_{D,Sat} = (V_g - V_t) \quad (6)$$

If (6) is substituted into (5), the saturation current is

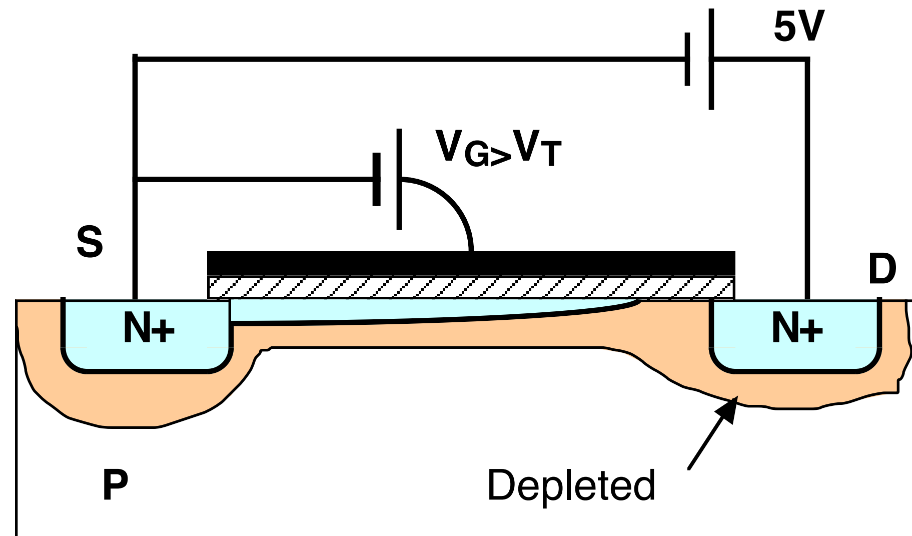
$$I_{D_{SAT}} \approx \frac{W}{2L} \mu_n \frac{\varepsilon_{ox}}{t_{ox}} \left(V_G - V_T\right)^2 \qquad (7)$$



Further increase in $\boldsymbol{V_D}$ does not change $\boldsymbol{I_D}$ (to first order), $\boldsymbol{I_D} \approx$ constant for $\boldsymbol{V_D} > \boldsymbol{V_{DSAT}}$

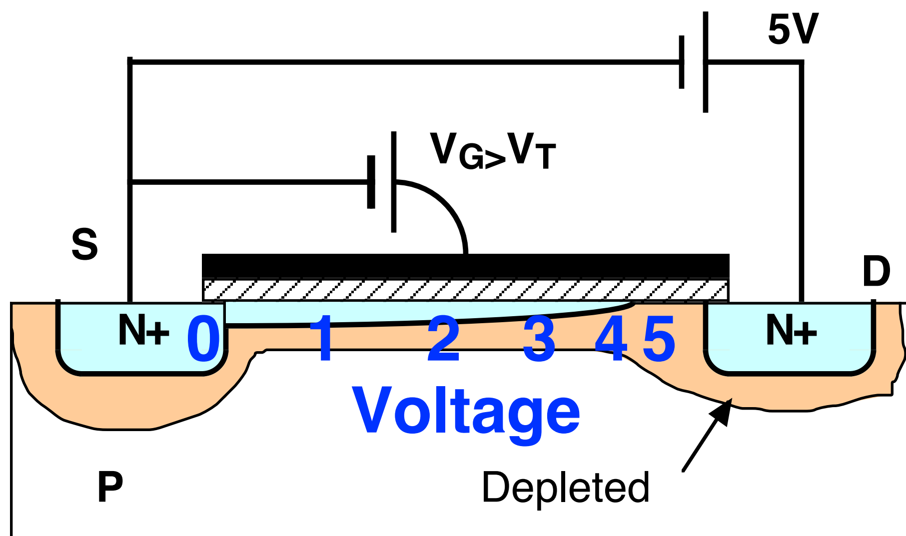
SATURATION REGION.

Why does the current remain constant past pinchoff (saturation):



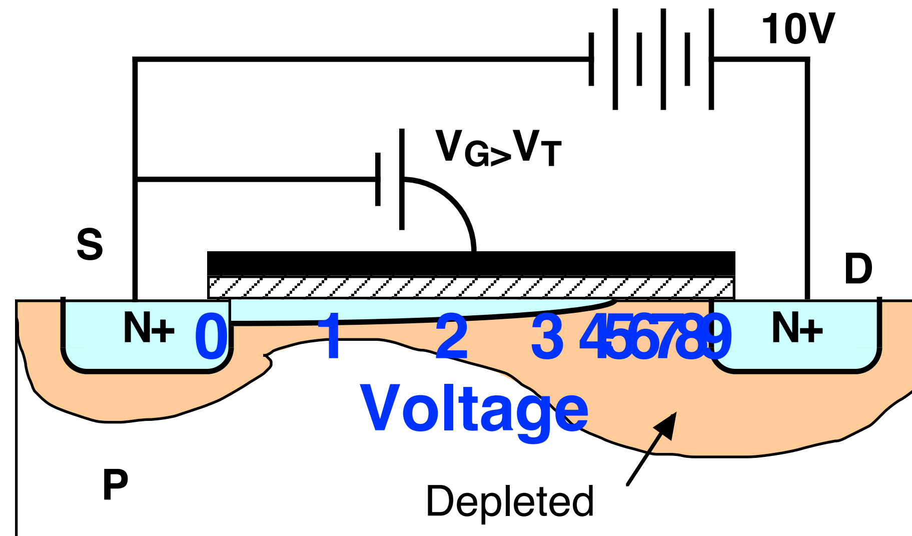
Why doesn't increasing V_D also increase I? If the current is constant, then the electric field, E, must also be constant along most of the channel.

Why the current remains constant past pinchoff (saturation):



Near pinchoff, the voltage is decreasing approximately linearly, hence the E field is “relatively” constant throughout.

Why the current remains constant past pinchoff (saturation):



- At <u>higher</u> V_d, there is a larger depleted region, hence, greater voltage drop.
- $\mathbf{E}$ is still linear in the channel but very large in depletion region.
- Carriers rapidly get swept out of depletion region

Exact Expressions

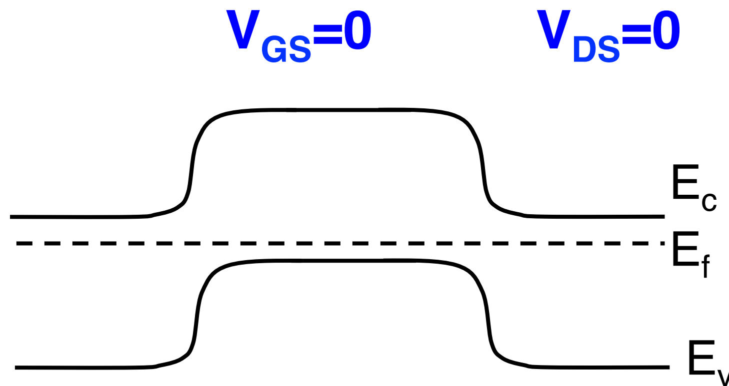
$$V_T(y) = V_{FB} + \frac{1}{C'_{ox}}\sqrt{2\varepsilon_s q N_a\left[-2\phi_p - \left(V_B - V(y)\right)\right]} - 2\phi_p + V(y)$$

$$Q_I(y) = -C'_{ox}\left[V_G - V_{FB} + 2\phi_p - V(y)\right] + \sqrt{2\varepsilon_s q N_a\left[2\,\phi_p - V_B + V(y)\right]}$$

$$I_D = \frac{W}{L}\mu_n\left\{C'_{ox}\left[V_G - V_{FB} + 2\phi_p - \frac{V_D}{2}\right]V_D - \frac{2}{3}\sqrt{2q\varepsilon_s N_a}\left[\left(V_D - 2\,\phi_p - V_B\right)^{\frac{3}{2}} - \left(-2\phi_p - V_B\right)^{\frac{3}{2}}\right]\right\}$$

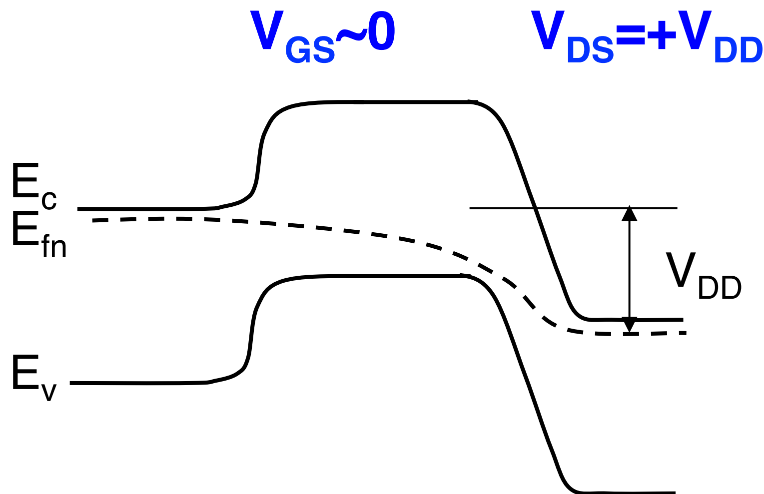
$$V_{D_{SAT}} = V_G - V_{FB} + 2\phi_p + \frac{q\;\varepsilon_s\,N_a}{C'^{\,2}_{ox}}\left[1 - \sqrt{1 + \frac{2\;C'^{\,2}_{ox}\left(V_G - V_{FB} - V_B\right)}{q\;\varepsilon_s\,N_a}}\right]$$

What about the Diffusion Current?



$$J_e = -D_n\left[\frac{dQ_i(y)}{dy}\right] + \mu_n Q_i(y)\left[\frac{dV(y)}{dy}\right]$$

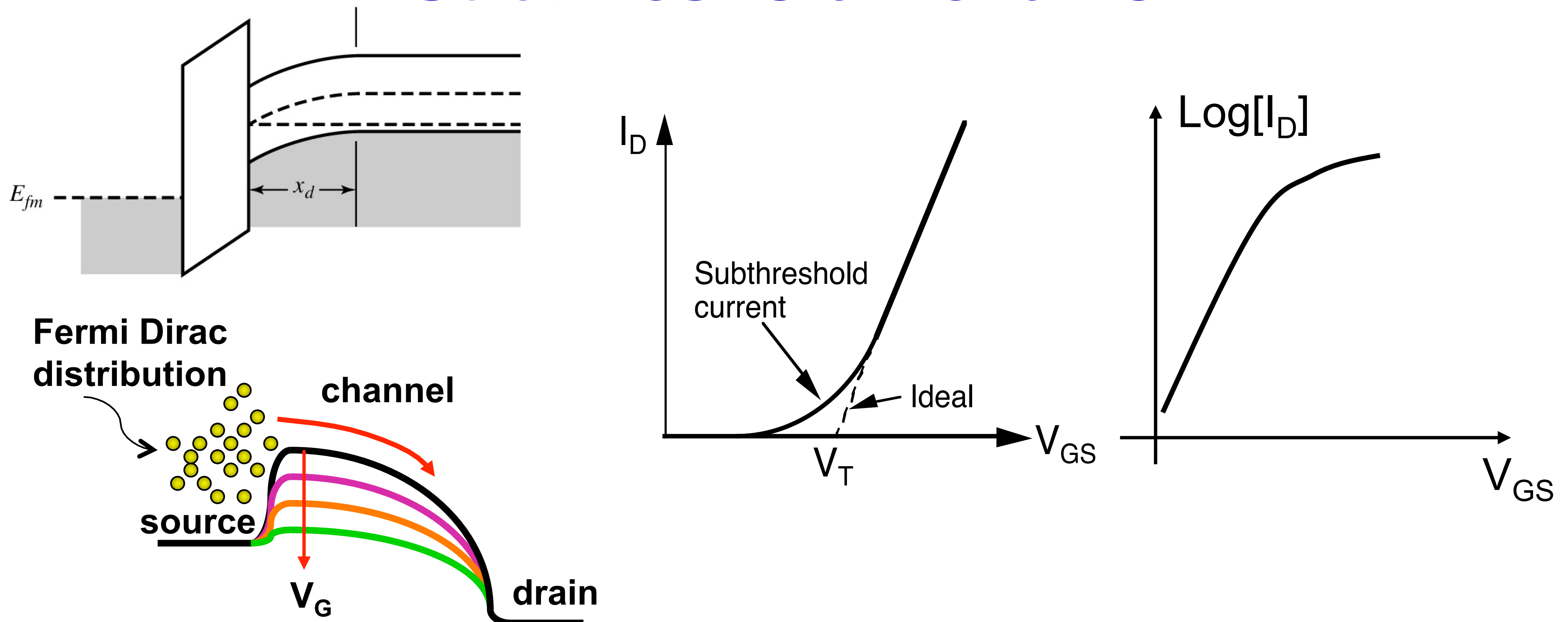
Diffusion | Drift



Sub-threshold regime

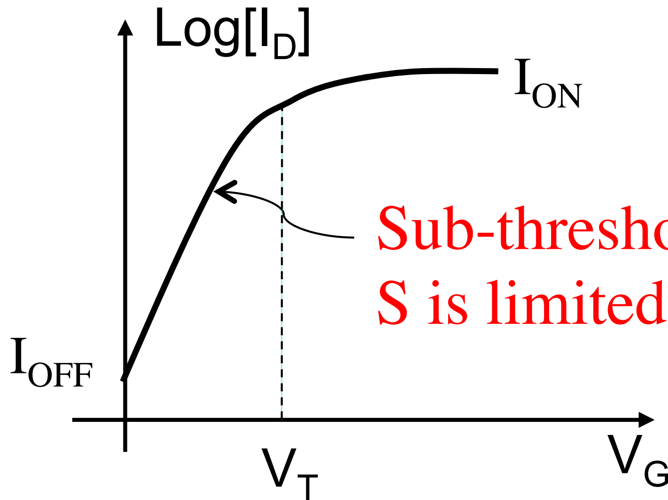
- Most of the V_{DS} drops across the reverse-biased S-D Junction. The Channel bands are still ~ flat. Therefore J_{drift} is negligible
- Gradient of free carriers along channel is large. Diffusion component $J_{diffusion}$ dominates.

Subthreshold Behavior



- When the surface is in weak inversion (i.e., $0 < \phi_s < -\phi_p$, $V_G < V_T$), a conducting channel starts to form and a low level of current flows between source and drain.
- Diffusion current due to carriers from source spilling over source barrier into channel due to application of V_G to lower ϕ_s
- Weak dependence on V_{DS} in long-channel FET

Sub-threshold Conduction



Inversion charge $Q_e(y) \propto \exp\left[\frac{\phi_s}{kT/q}\right]$

$$\phi_S = \frac{C_{ox}}{C_{ox}+C_S}(V_G - V_T) = \frac{(V_G - V_T)}{m}$$

where $m = \frac{C_{ox}+C_S}{C_{ox}}$

Drain current $I_D \propto Q_e \propto \exp\left[\frac{V_G - V_T}{mkT/q}\right]$

V_G

C_{ox}

ϕ_s

C_s = Channel Capacitance

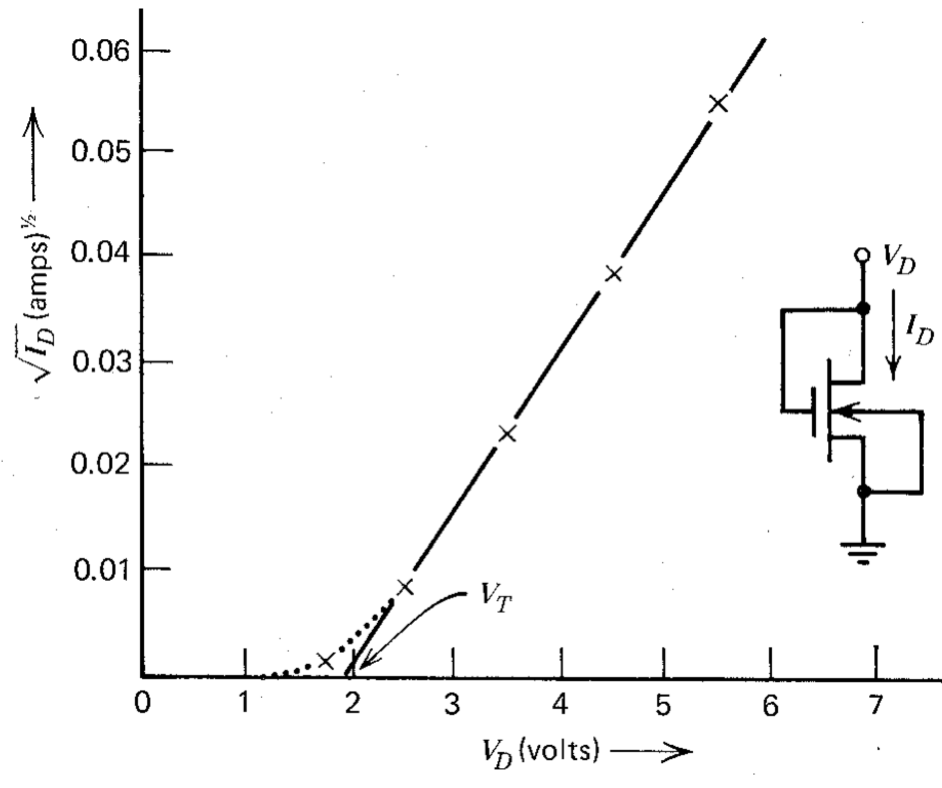
Subthreshold swing

$$S = \frac{\partial V_G}{\partial(\log I_D)} = \frac{\partial V_G}{\partial \phi_S}\frac{\partial \phi_S}{\partial(\log I_D)} = \left(1 + \frac{C_S}{C_{ox}}\right)\frac{kT}{q}$$

Gate to channel potential coupling m > 1 in MOSFET

60 mV/dec due to Fermi-Dirac distribution

V_T Extraction



$V_D = V_G$

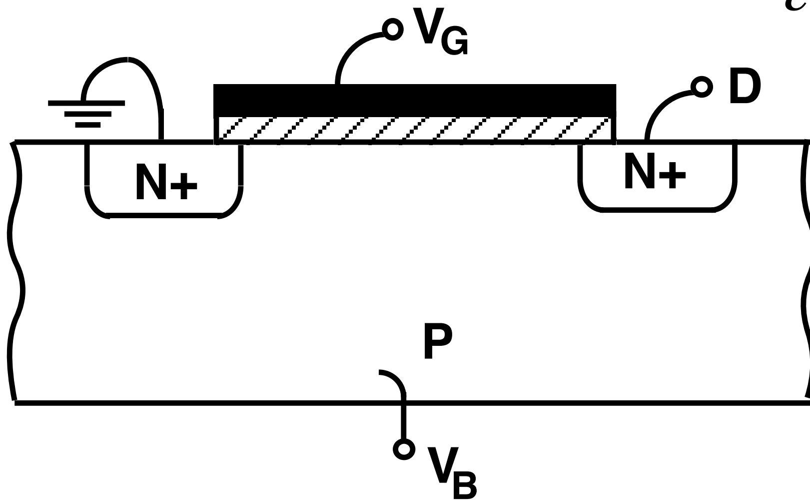
$$I_{D_{SAT}} \approx \frac{W}{2\ L} \mu_n \frac{\varepsilon_{ox}}{t_{ox}} (V_G - V_T)^2$$

$$= \frac{W}{2\ L} \mu_n \frac{\varepsilon_{ox}}{t_{ox}} (V_D - V_T)^2$$

- The intercept of $\sqrt{I_D}$ vs V_D gives the value of threshold voltage V_T. This technique is widely used to extract the value of V_T.
- The region depicted by the dotted curve below V_T is the WEAK INVERSION REGION.

Effect of Substrate (Back Gate) Bias

For small V_D $$V_T = V_{FB} + \frac{t_{ox}}{\varepsilon_{ox}} \sqrt{2\, K_s \varepsilon_o q N_a \left(-2\phi_p - V_B\right)} - 2\phi_p$$



The body voltage (or backside bias) makes it easier or harder to reach inversion: --> Change in threshold voltage (V_T).

The change in V_T due to V_B is described as

$$\Delta V_T = \frac{1}{C'_{ox}} \sqrt{2\varepsilon_s q N_a} \left[\sqrt{-2\phi_p - V_B} - \sqrt{-2\phi_p}\right] \qquad (8)$$

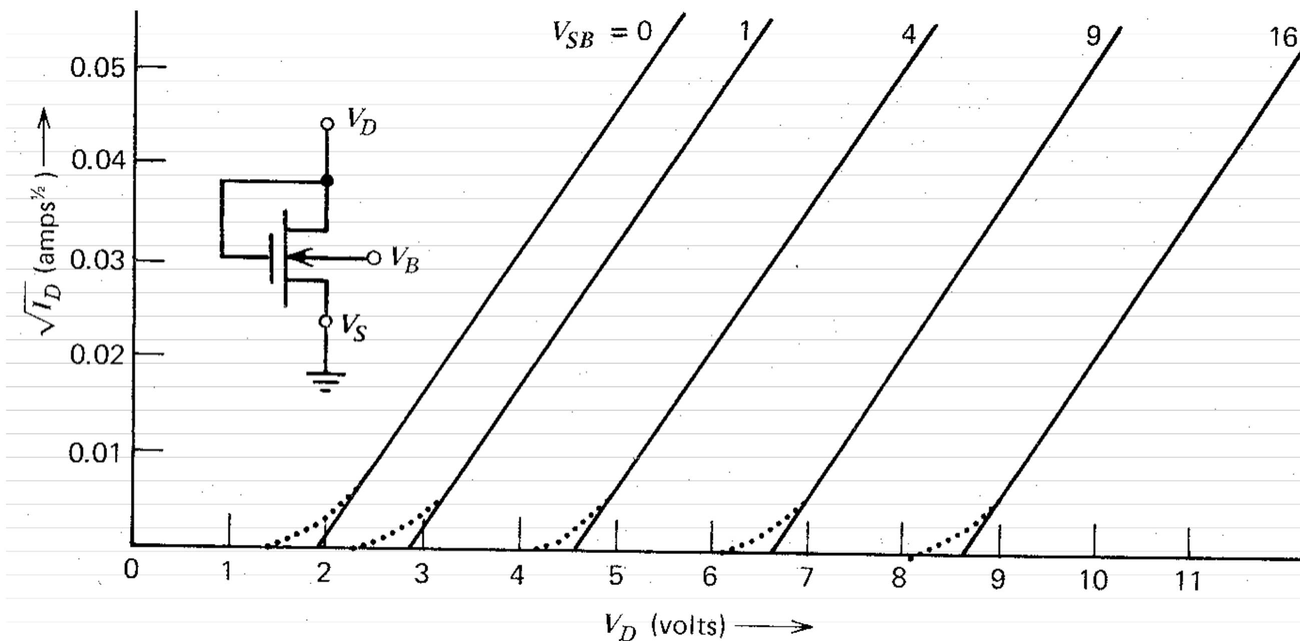
$$\approx \frac{1}{C'_{ox}} \sqrt{2\varepsilon_s q N_a} \sqrt{-V_B}$$

$$= \gamma \sqrt{-V_B} \qquad (9)$$

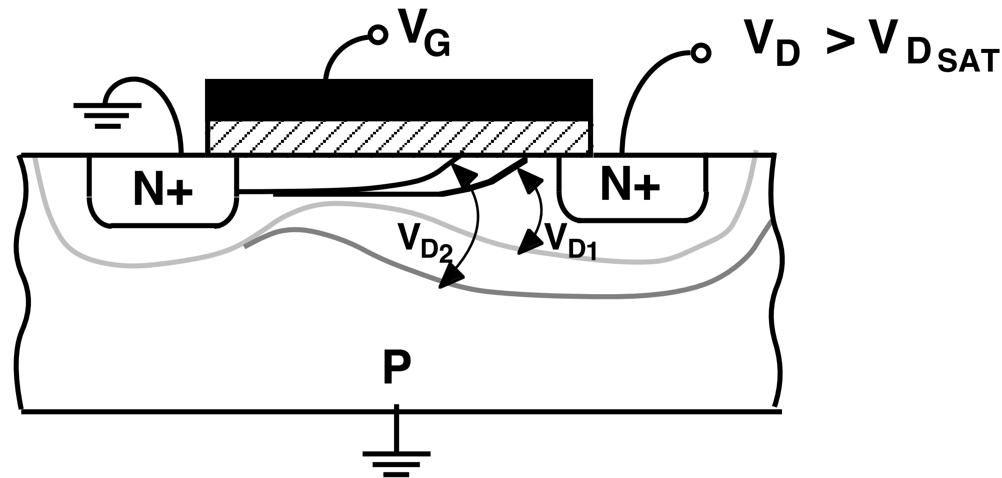
Where

$$\gamma = \frac{1}{C'_{ox}}\sqrt{2\varepsilon_s q N_a} \; = \; \text{body factor} \qquad (10)$$

Equations (5) and (7) can be used to approximate the I-V characteristic if $\boldsymbol{V_T}$ is replaced with $\boldsymbol{V_T + \Delta V_T}$.



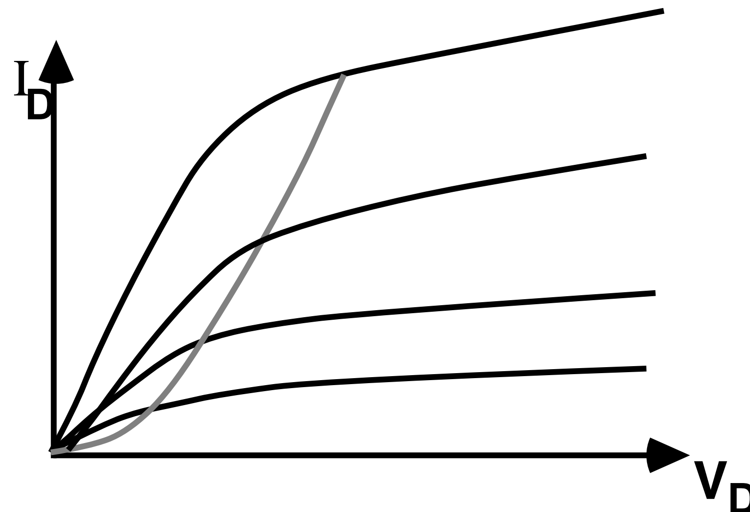
Channel Length Modulation



In the saturation region, as V_D ↑, the depletion region near drain expands, the pinch-off point of the channel moves back towards source. The effective channel becomes shorter, I_D ↑ because it is proportional to μ/L_{eff}. The depletion region expands as $\sqrt{V_D}$ assuming a step junction. Provided the device has a channel length >> ΔX_D then the change in channel length is approximated by difference in the depletion width of a step junction.

$$\Delta L \approx -\sqrt{\frac{2}{q}\frac{\varepsilon_s}{N_a}}\left(\sqrt{V_D} - \sqrt{V_{D_{SAT}}}\right) \qquad (11)$$

The decrease in ***L*** is responsible for an increase in I_D in the saturation region.



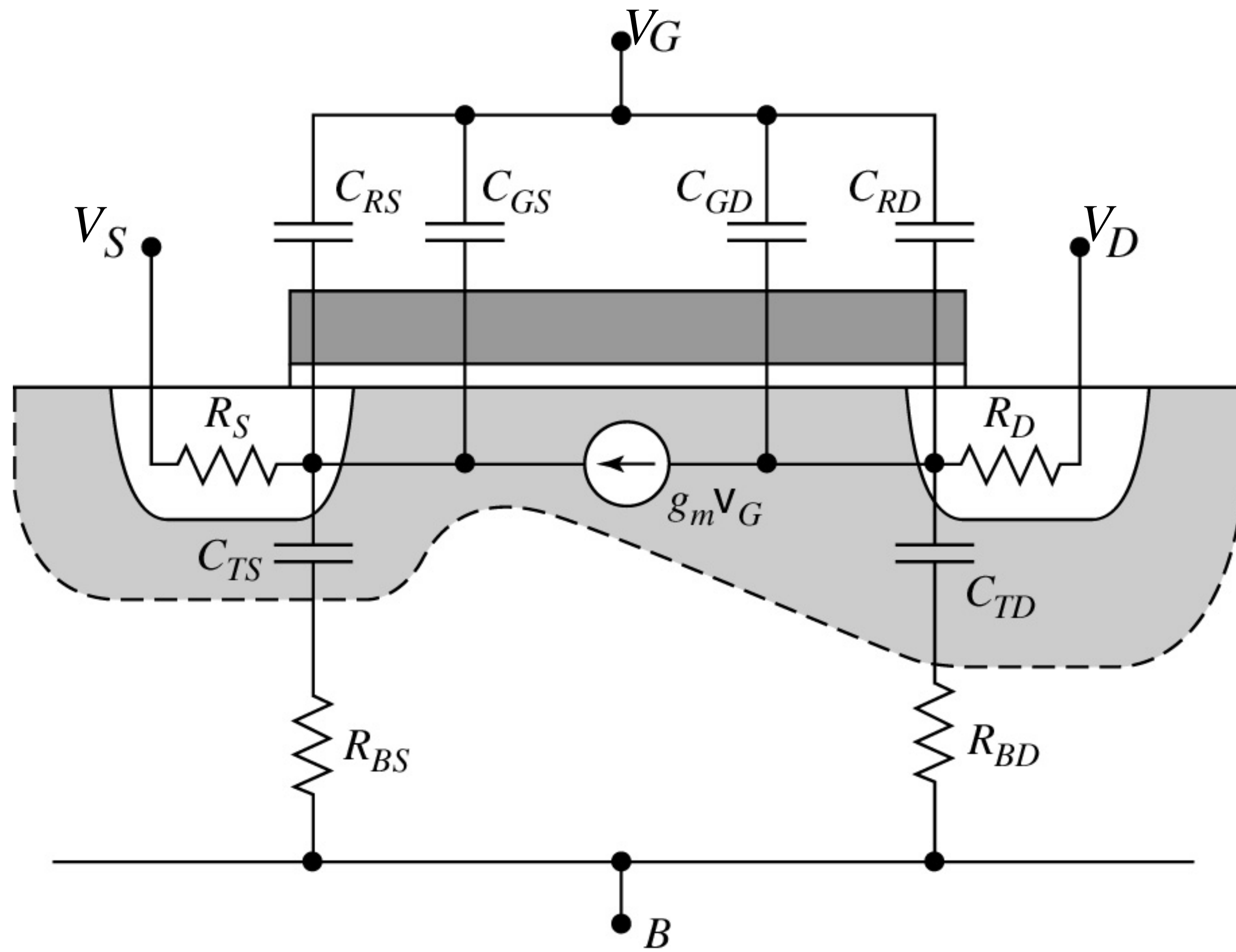
Therefore, a finite output impedance results. For most applications, this is modeled as

$$I_{D_{SAT}} = \frac{W}{2L} \mu_n C'_{ox} (V_G - V_T)^2 (1 + \lambda V_D) \qquad (12)$$

where λ = channel length modulation parameter

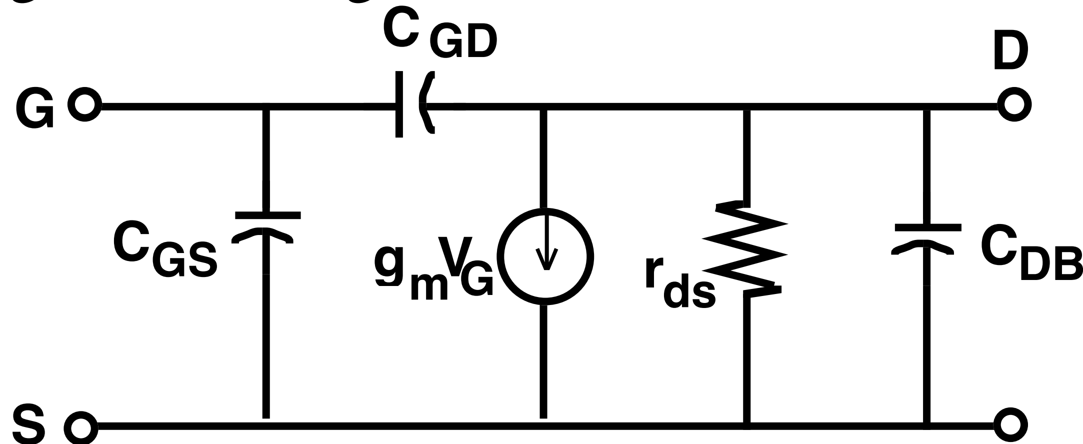
Circuit Models

The MOS transistor may be modeled in the following manner, by inspection of its physical structure.



Of the elements in the model, only the gate to channel capacitance is essential; the rest are parasitic elements which degrade performance. Technology improvements are generally designed to reduce these parasitics.

In many cases, the equivalent circuit can be reduced to the following for small signals:



Note that many of the parameters in the model are voltage sensitive. Accurate large signal model usually requires computer techniques.

Transconductance, g_m

Defined from the I_D-V_D characteristics in both the linear and saturation regions

$$I_D = \frac{W}{L}\mu_n \frac{\varepsilon_{ox}}{t_{ox}}[V_G - V_T]V_D$$

$$I_{D_{SAT}} \approx \frac{W}{2\ L}\mu_n \frac{\varepsilon_{ox}}{t_{ox}}(V_G - V_T)^2$$

The transconductance or gain of the device is defined as:

$$g_m = \left.\frac{\partial I_D}{\partial V_G}\right| V_D = const$$

$$g_m = \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D = const}$$

$$\approx \frac{W}{L}\mu_n \frac{\varepsilon_{ox}}{t_{ox}} V_D \qquad \text{for } V_D < V_{D_{SAT}},\ \textit{linear region}$$

$$\approx \frac{W}{L}\mu_n \frac{\varepsilon_{ox}}{t_{ox}} (V_G - V_T) \quad \text{for } V_D > V_{D_{SAT}},\ \textit{saturation region} \qquad (13)$$

B. Gate Capacitances, C_{GS} and C_{GD}

The gate capacitances vary as the device moves from the linear to saturation region,

C_{GS} = 1/2 C_{ox} to 2/3 C_{ox} from linear to saturation (14)

C_{GD} = 1/2 C_{ox} to 1/3 C_{ox} from linear to saturation (15)

Output Impedance, r_{ds}

The output impedance or resistance of the device is defined as:

$$r_{ds} = \left.\frac{\partial V_D}{\partial I_D}\right|_{V_G = const}$$

$$\approx \frac{1}{\left[\frac{W}{L}\mu_n \frac{\varepsilon_{ox}}{t_{ox}}(V_G - V_T)\right]} \quad \text{for } V_D < V_{D_{SAT}}, \ \textit{linear region}$$

$$\approx \frac{1}{\lambda I_D} \quad \text{for } V_D > V_{D_{SAT}}, \ \textit{saturation region} \quad (16)$$

Note: Small signal model applicable to analog applications is NOT APPLICABLE to digital applications.