

EE 332: Devices and Circuits II

Lecture 2: MOS Devices (Part 1)

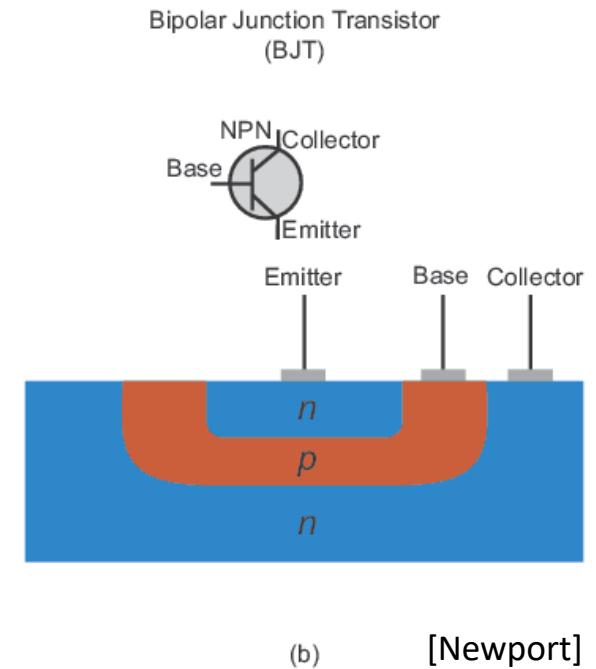
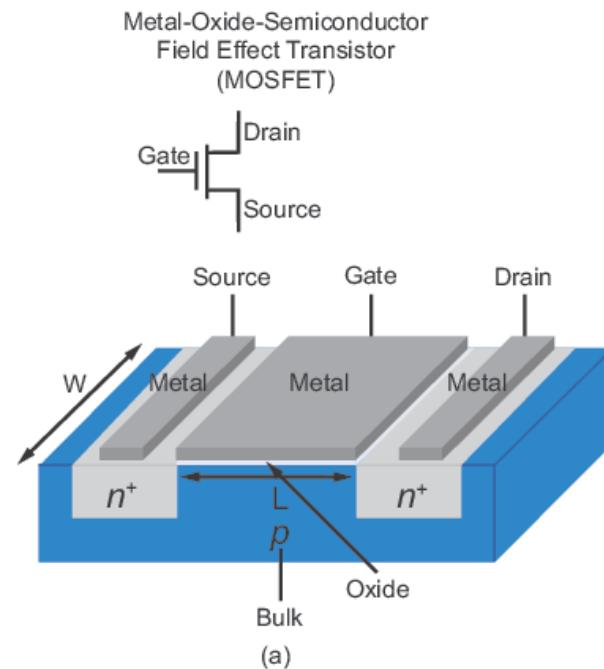
Prof. Sajjad Moazeni

smoazeni@uw.edu

Autumn 2022

MOSFET vs. BJT Devices

- Bipolar junction transistor (BJT)
 - Works based on the diffusion current
 - Can operate at higher frequencies
 - Lower noise
 - Larger intrinsic gain ($g_m r_o$)
- Metal-oxide-semiconductor field-effect transistor (MOSFET)
 - Works based on the drift current
 - Very scalable and compact
 - Lower power (excellent for mixed-signal ICs)
 - Widely used in today's ICs



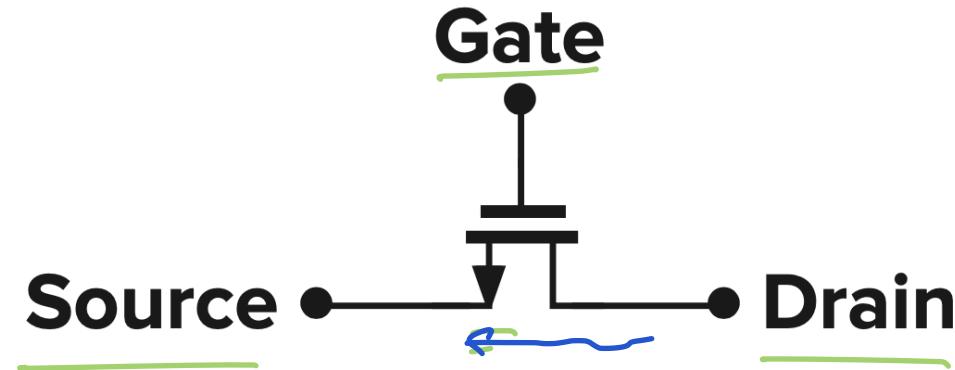
(b)

[Newport]

We only focus on MOSFETs in this course!

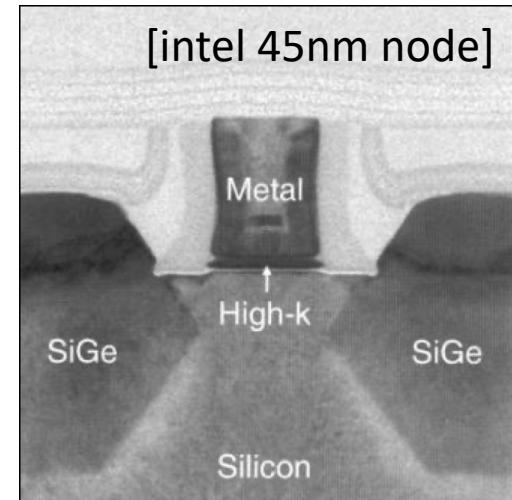
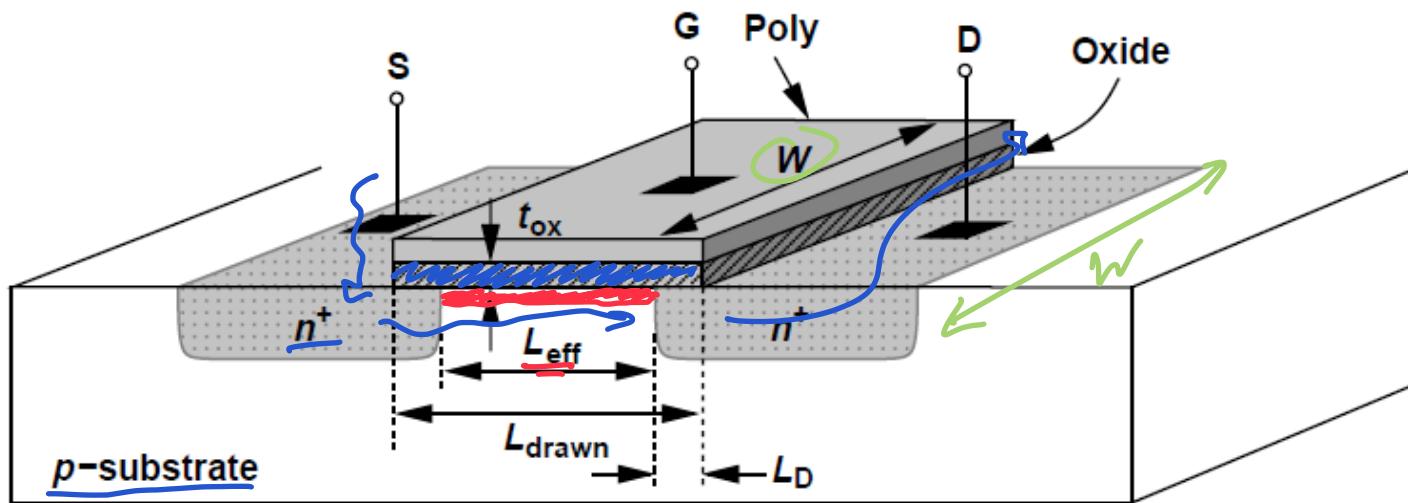
(However methods and circuit topologies are very similar between MOS and BJTs)

MOSFET as a Switch



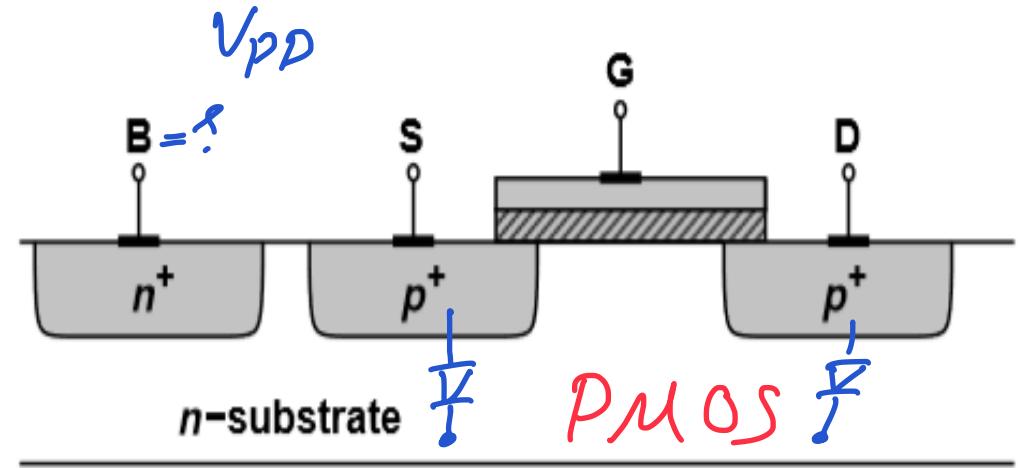
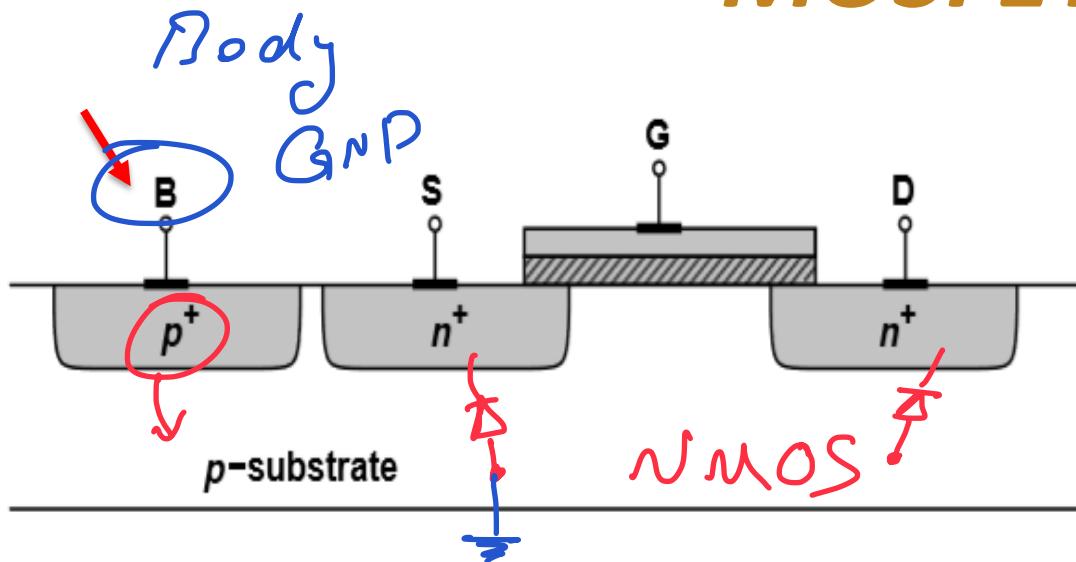
- When gate voltage is high, device is on (Recall this from CMOS logics in EE331).
- Source and drain are interchangeable.
- But,
 - At what gate voltage does the device turn on?
 - How much is the resistance between S and D?
 - What limits the speed of the device?

MOSFET Structure



- n-type MOS (NMOS) has n-doped source (S) & drain (D) on p-type substrate ("bulk" or "body")
- S/D junctions “side-diffuse” during fabrication so that effective length $L_{eff} = L_{drawn} - 2L_D$
- Typical values are $\underline{L_{eff} \approx 10 \text{ nm}}$ and $\underline{t_{ox} \approx 15 \text{ \AA}}$ $\underline{1 \text{ } 2 \text{ nm}}$
- The S terminal provides charge carriers and the D terminal collects them.
 - As voltages at the three terminals changes, the source and drain may exchange roles (Device structure is symmetric).

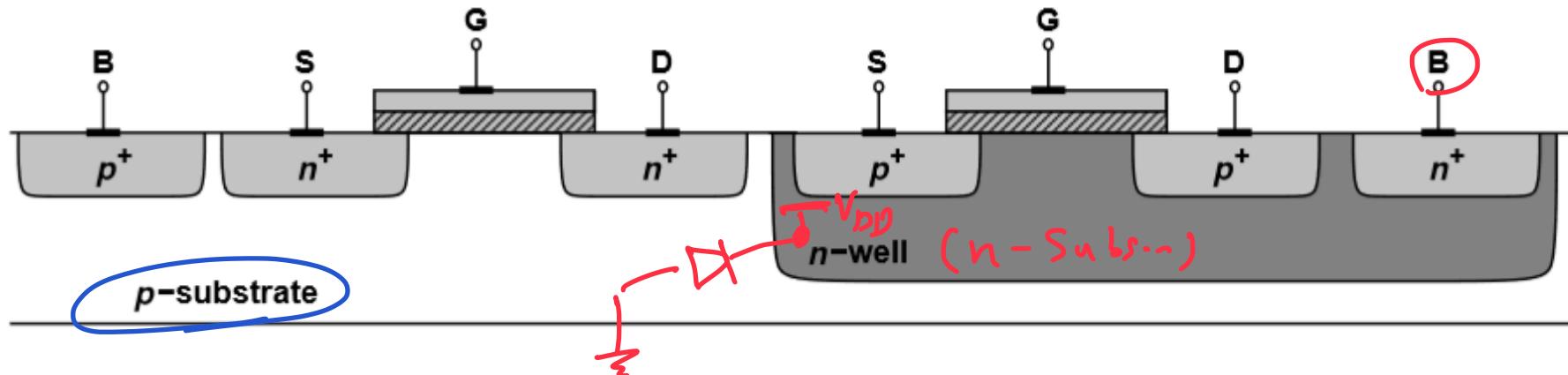
MOSFET Structure



- MOSFETs actually have four terminals.
- Substrate potential greatly influences device characteristics.
- Typically S/D junction diodes are reversed-biased and the NMOS substrate is connected to the most negative supply in the system.

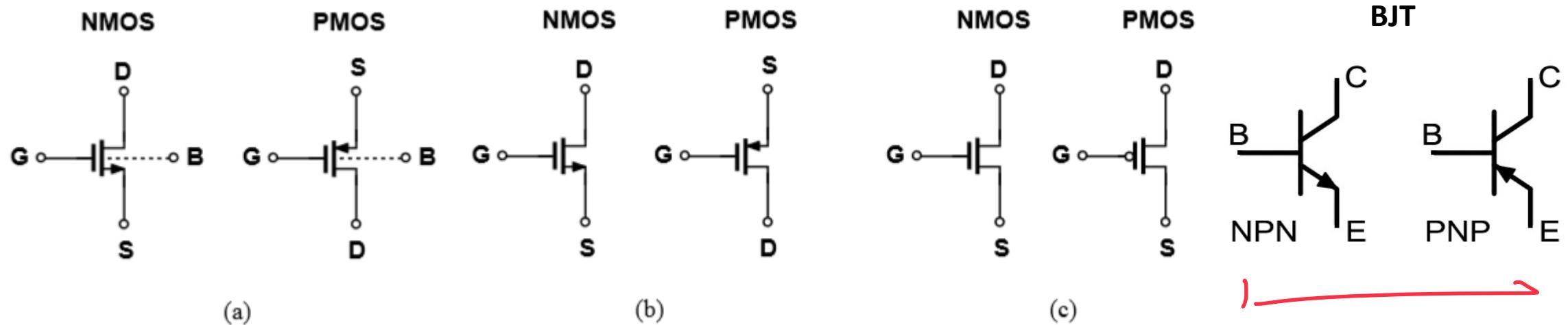
- PMOS is obtained by inverting all of the doping types (including the substrate).

MOSFET Structure



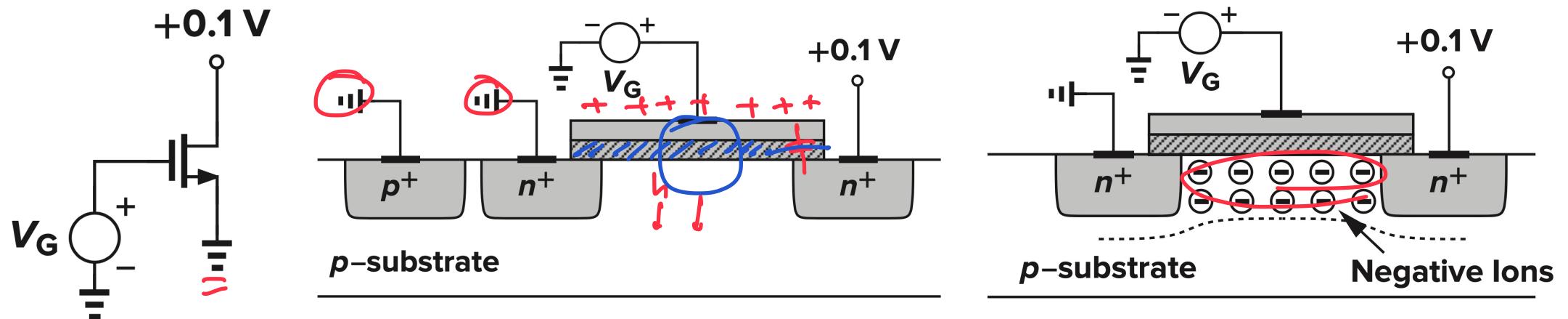
- In complementary MOS (CMOS) technologies both NMOS (NFET) and PMOS (PFET) are needed and fabricated on the same wafer.
- In today's CMOS, the PMOS is fabricated in an *n*-well, where the *n*-well is tied to the most positive supply voltage (normally called VDD).

MOS Symbols



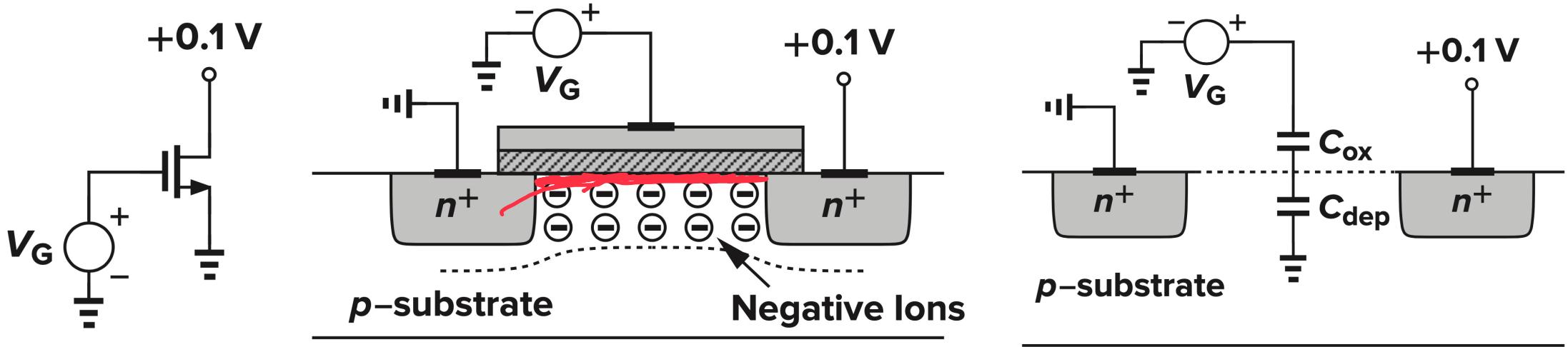
- Substrate is denoted by “B” (bulk).
- PMOS source is positioned on top since it has a higher potential than the gate.
- Most circuits have NMOS and PMOS bulk tied to ground and V_{DD} , respectively, so we tend to omit the connections (b,c).
- Digital circuits tend to incorporate “switch” symbols (c).

Threshold Voltage



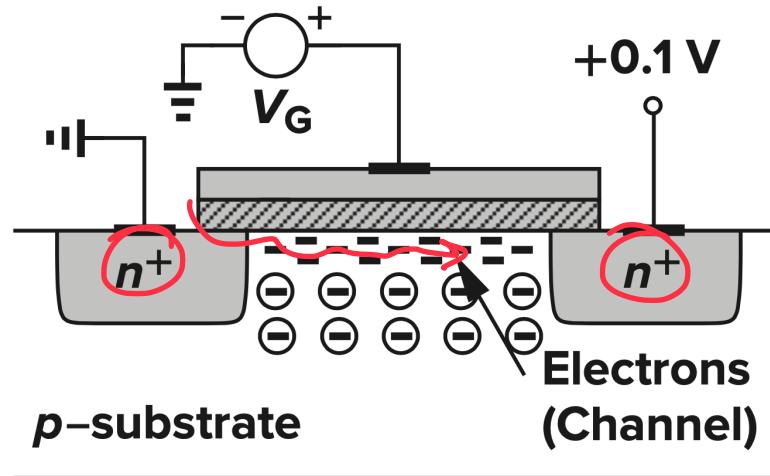
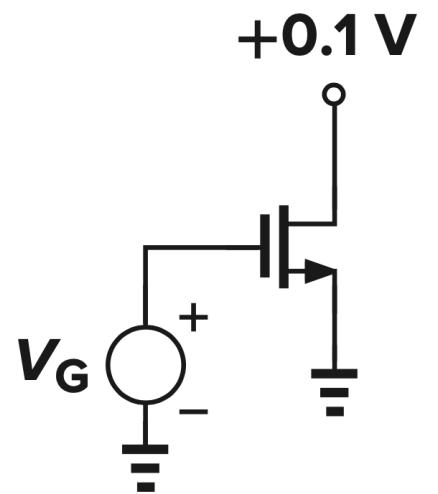
- As V_G increases from zero, holes in p -substrate are repelled leaving negative ions behind to form a depletion region.
- There are no charge carriers, so no current flow.

Threshold Voltage



- Increasing V_G further increases the width of the depletion region and the potential at the oxide-silicon interface.
- Structure resembles voltage divider consisting of gate-oxide capacitor and depletion region capacitor in series.

Threshold Voltage



- When interface potential reaches sufficiently positive value, electrons flow from the source to the interface and eventually to the drain.
- This creates a channel of charge carriers (inversion layer) beneath the gate oxide.
- The value of V_G at which the inversion layer occurs is the **threshold voltage (V_{TH})**.

Threshold Voltage

$$V_{TH} = \Phi_{MS} + 2\Phi_F + \frac{Q_{dep}}{C_{ox}}$$

$$\Phi_F = (kT/q) \ln(N_{sub}/n_i) \quad Q_{dep} = \sqrt{4q\epsilon_{si}|\Phi_F|N_{sub}}$$

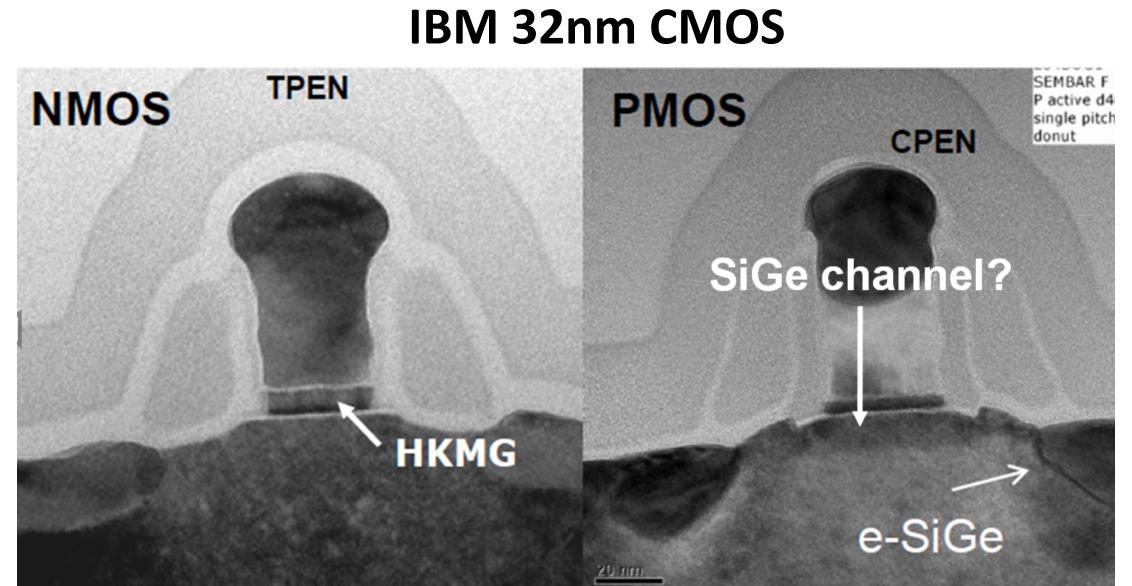
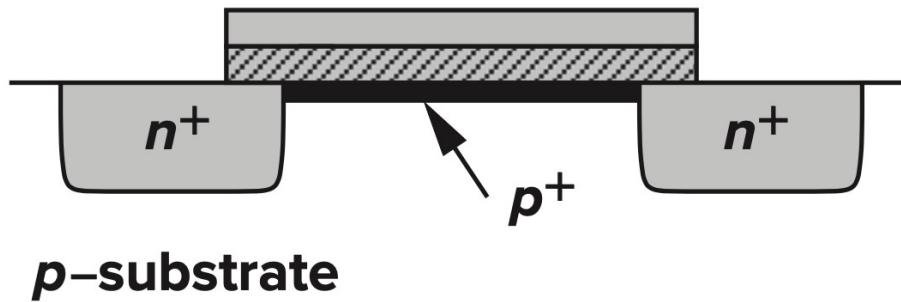
- Where

- Φ_{MS} is the difference between the work functions of the polysilicon gate and the silicon substrate.
- k is Boltzmann's constant, q is the electron charge ($kT/q \sim 25\text{mV}$ at room temp).
- N_{sub} is the doping density of the substrate.
- n_i is the density of electrons in undoped silicon.
- Q_{dep} is the charge in the depletion region.
- C_{ox} is the gate oxide capacitance per unit area.
- ϵ_{si} is the dielectric constant of silicon.

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \rightarrow C_{ox} \sim \frac{1}{20} \frac{\text{FF}}{\mu\text{m}^2}$$

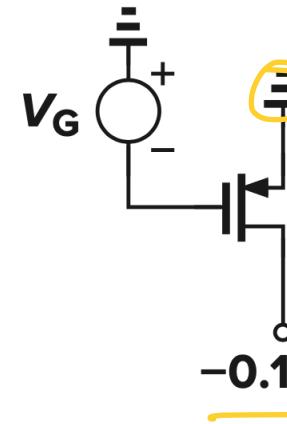
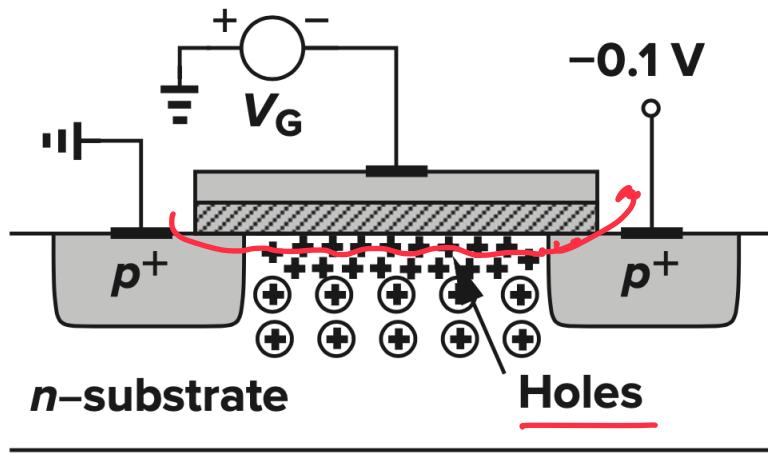
$t_{ox} \approx 2\text{nm}$

Threshold Voltage



- In practice, threshold voltage is adjusted by implanting dopants into the channel area during device fabrication.
- For NMOS, adding a thin sheet of p^+ increases the gate voltage necessary to deplete the region.

Threshold Voltage

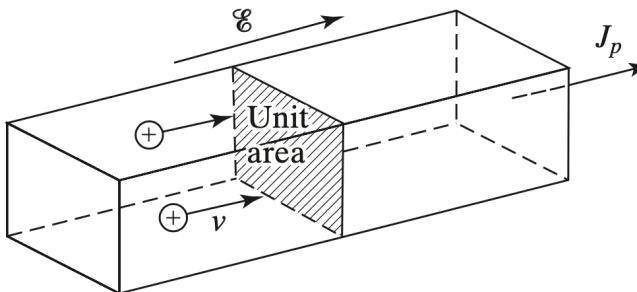


$$I_{SD} = f(V_S, V_G, V_D)$$

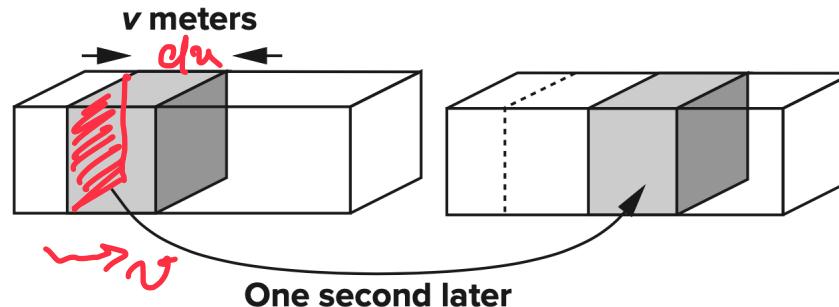
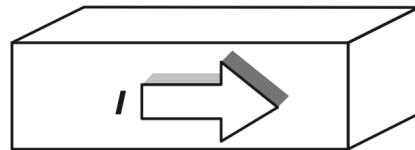
V_{th}

- Turn-on phenomena in PMOS is similar to that of NMOS but with all polarities reversed.
- If the gate-source voltage becomes sufficiently *negative*, an inversion layer consisting of holes is formed at the oxide-silicon interface, providing a conduction path between source and drain.
- PMOS threshold voltage is negative.

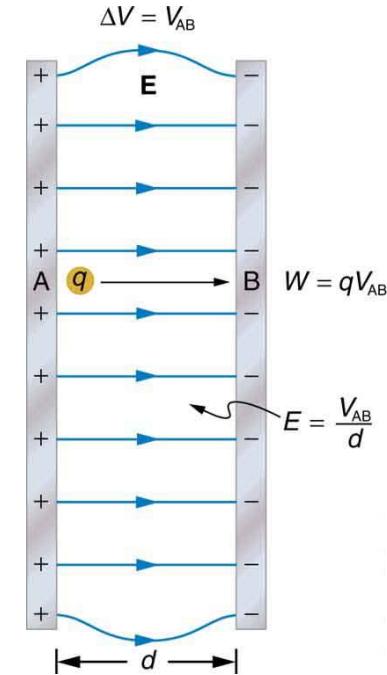
Derivation of I/V Characteristics (Basic Eqs.)



$$v = \mu E$$



$$I = \frac{dQ}{dt} \rightarrow \frac{Q_d v dt}{dt} \Rightarrow I = Q_d \cdot v$$



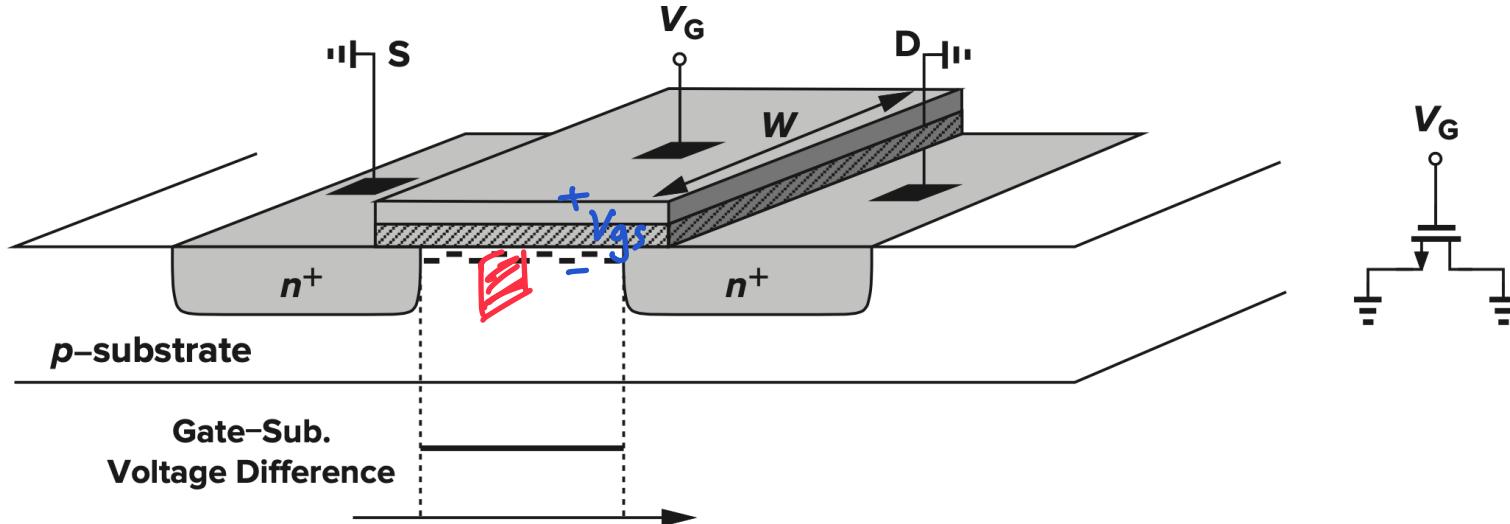
$$E(x) = -dV/dx$$

- Where:
 - Q_d is the mobile charge density along the direction of current I .
 - v is the charge velocity & μ is the carrier mobility.

$$C_{ox} \sim 17 \text{ fF}/\mu\text{m}^2$$

$$t_{ox} = 2 \text{ nm}$$

Derivation of I/V Characteristics

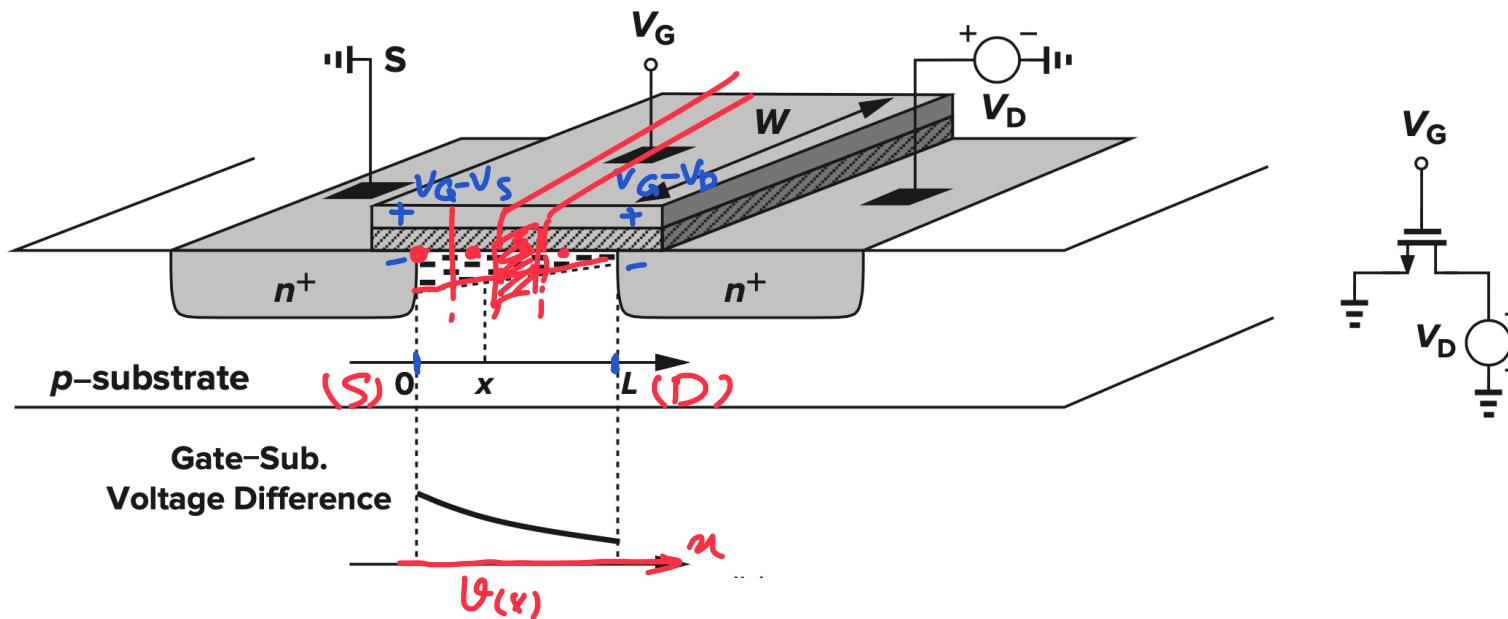
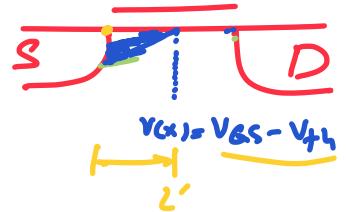


- Onset of inversion occurs at $V_{GS} = V_{TH}$.
- Inversion charge density produced by gate oxide capacitance is proportional to $V_{GS} - V_{TH}$ since for $V_{GS} \geq V_{TH}$, charge placed on the gate must be mirrored by charge in the channel, yielding a uniform channel charge density:

$$Q_d = \underbrace{WC_{ox}}_{\text{Total capacitance per unit length.}} \underbrace{(V_{GS} - V_{TH})}_{\text{Gate-Sub. Voltage Difference}}$$

- Where WC_{ox} is the total capacitance per unit length.

Derivation of I/V Characteristics



- Channel potential varies from zero at the source to V_D at the drain.
- Local voltage *difference* between the gate and the channel varies from V_G to $V_G - V_D$.
- Charge density now varies with respect to x :

$$Q_d(x) = WC_{ox}[V_{GS} - V(x) - V_{TH}]$$

where $V(x)$ is the channel potential at x .

$$\left. \begin{array}{l} V(0) = V_S = 0 \\ V(L) = V_D = V_{DS} \end{array} \right\}$$

Derivation of I/V Characteristics

- Since

$$\left\{ \begin{array}{l} I = Q_d \cdot v \\ v = \mu E \\ E(x) = -\frac{dV}{dx} \\ Q_d(x) = WC_{ox}[V_{GS} - V(x) - V_{TH}] \\ I_D = WC_{ox}[V_{GS} - V(x) - V_{TH}] \mu_n \left(\frac{dV(x)}{dx} \right) \end{array} \right.$$

$$I_D \times L = \int_{x=0}^L I_D dx = \int_{V=0}^{V_{DS}} WC_{ox} \mu_n [V_{GS} - V(x) - V_{TH}] dV$$

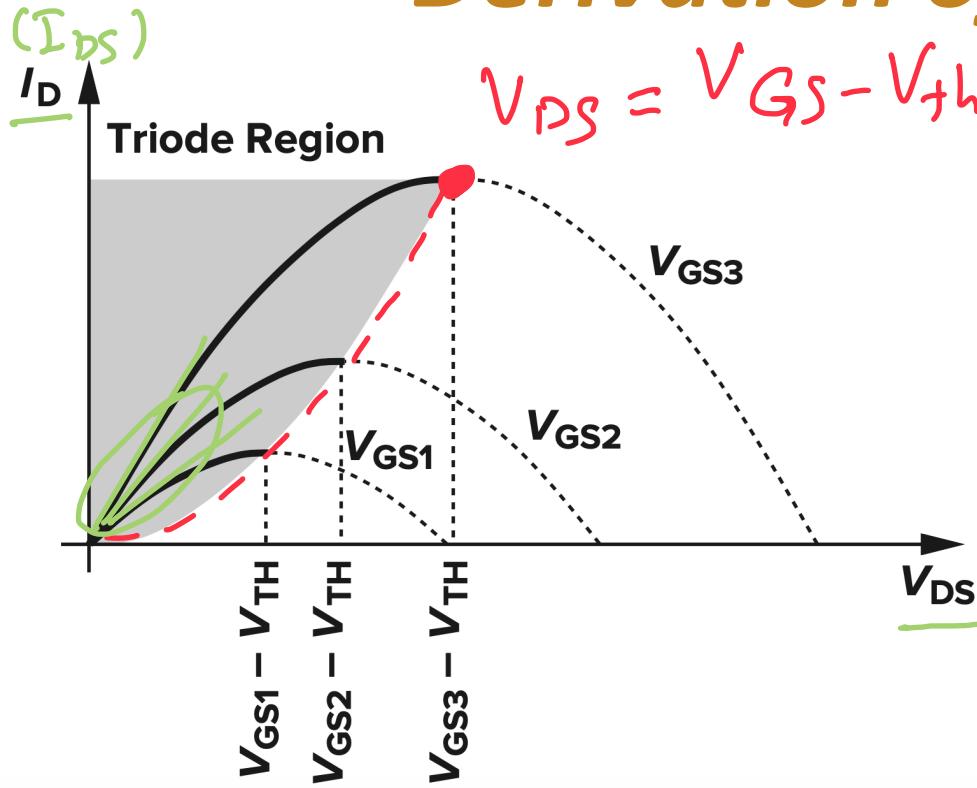
- A negative sign is added because the charge carriers are negative for NMOS.

$$I_D = \mu_n C_{ox} \frac{W}{L} [(V_{GS} - V_{TH}) \cdot V_{DS} - \frac{1}{2} V_{DS}^2]$$

aspect ratio

$$V_{DS} \leq V_{GS} - V_{TH}$$

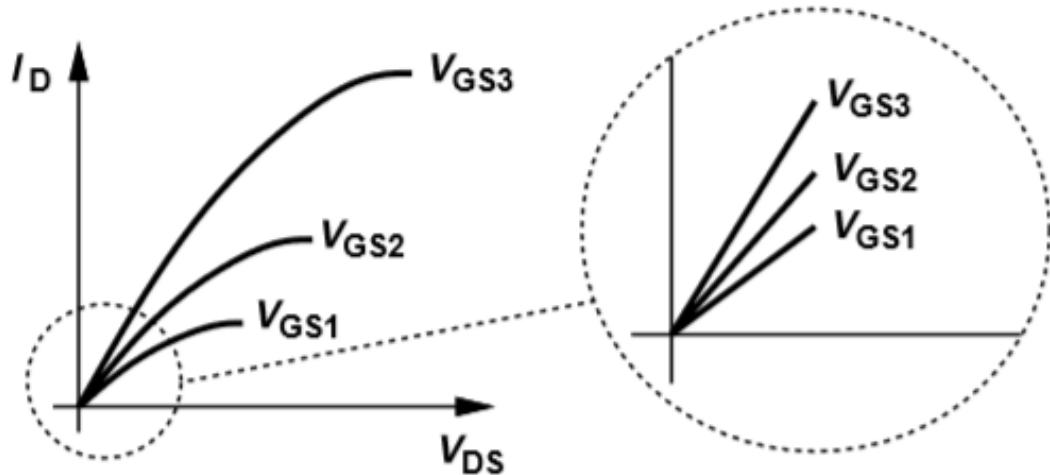
Derivation of I/V Characteristics



$$I_D = WC_{ox}[V_{GS} - V(x) - V_{TH}] \mu_n \left(\frac{dV(x)}{x} \right)$$
$$I_{D,max} = \frac{1}{2} \mu_n C_{ox} \left(\frac{W}{L} \right) (V_{GS} - V_{TH})^2$$

- $V_{GS} - V_{TH}$ is known as the “overdrive voltage.” (typically denoted by V_{OD})
- W/L is known as the “aspect ratio.”
- If $\underline{V_{DS}} \leq \underline{V_{GS} - V_{TH}}$, we say the device is operating in the “triode region” (or “linear region”)

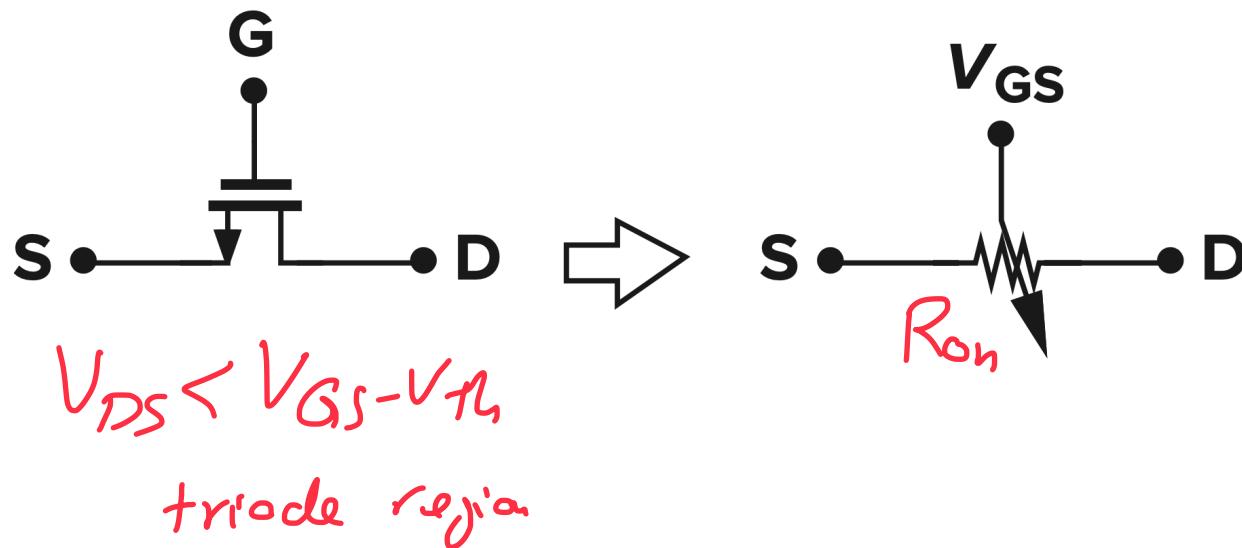
Derivation of I/V Characteristics



- If $V_{DS} \ll 2(V_{GS} - V_{TH})$, then $\rightarrow I_D \approx \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH}) V_{DS}$
- In this case, the drain current is a linear function of V_{DS} so the path from source to drain can be represented by a linear resistor:

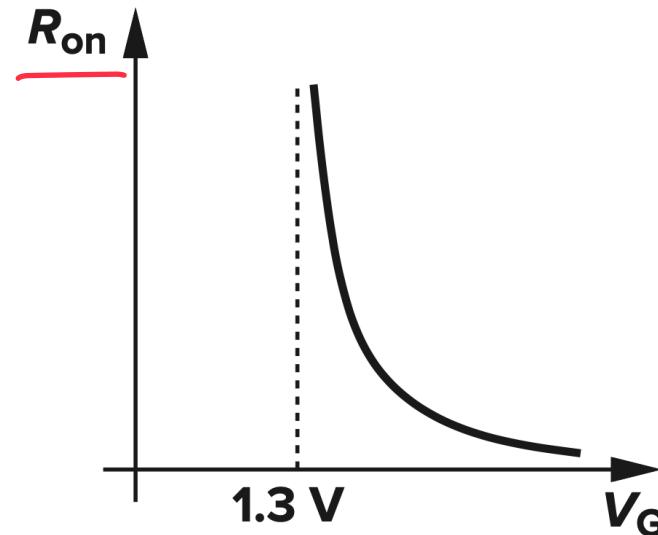
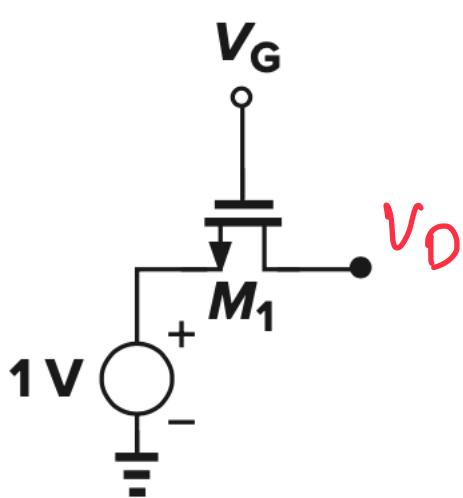
$$\frac{V_{DS}}{I_{DS}} = R_{on} = \frac{1}{\mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})}$$

Derivation of I/V Characteristics



- If $V_{DS} \ll 2(V_{GS} - V_{TH})$, the device is operating in “deep triode region.”
- In this region, a MOSFET can operate as a resistor whose value is controlled by the overdrive voltage.
- Unlike bipolar transistors, a MOS device may be on even if it carries no current.

Derivation of I/V Characteristics



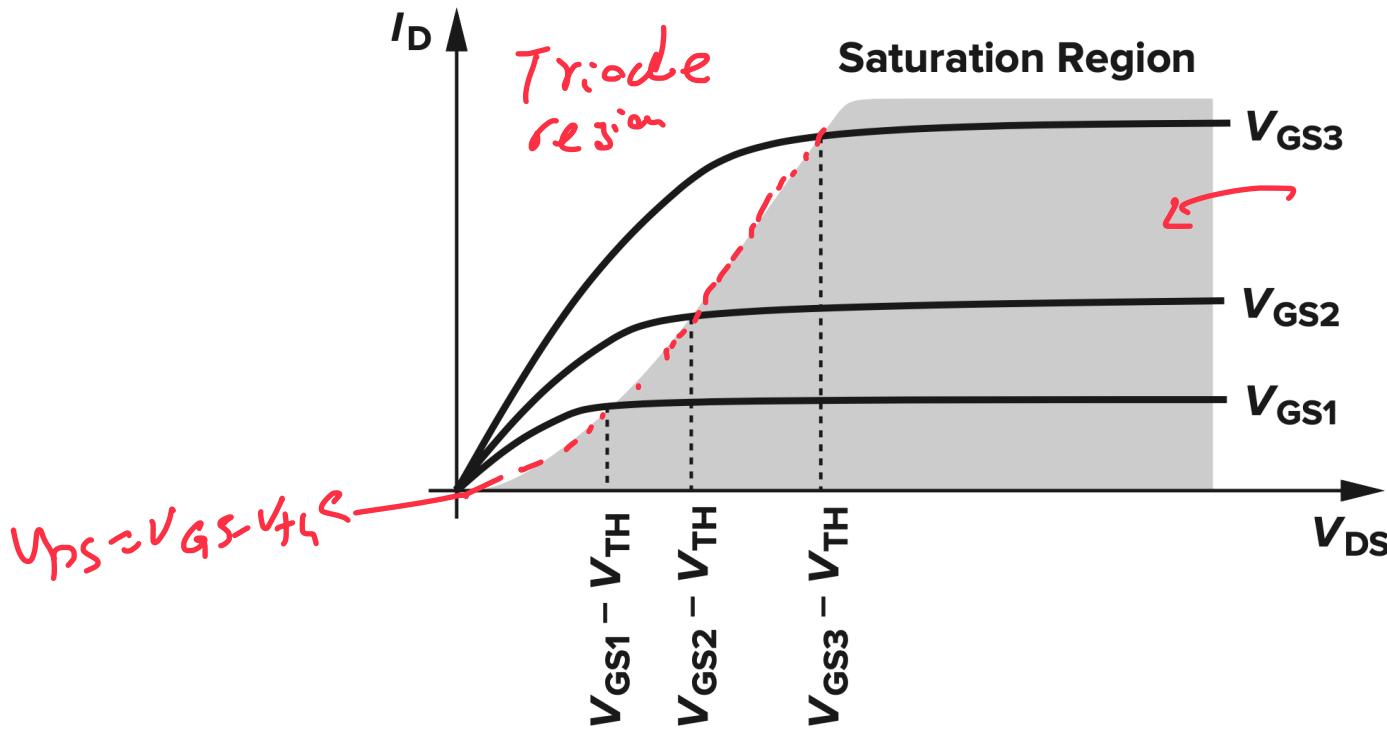
- For example, given the topology on the left and that

$$\mu_n C_{ox} = 50 \mu\text{A}/\text{V}^2$$

$\overbrace{W/L = 10}^{0.9 \mu\text{m}}$
 $\overbrace{V_{TH} = 0.3 \text{ V}}^{0.9 \mu\text{m}}$

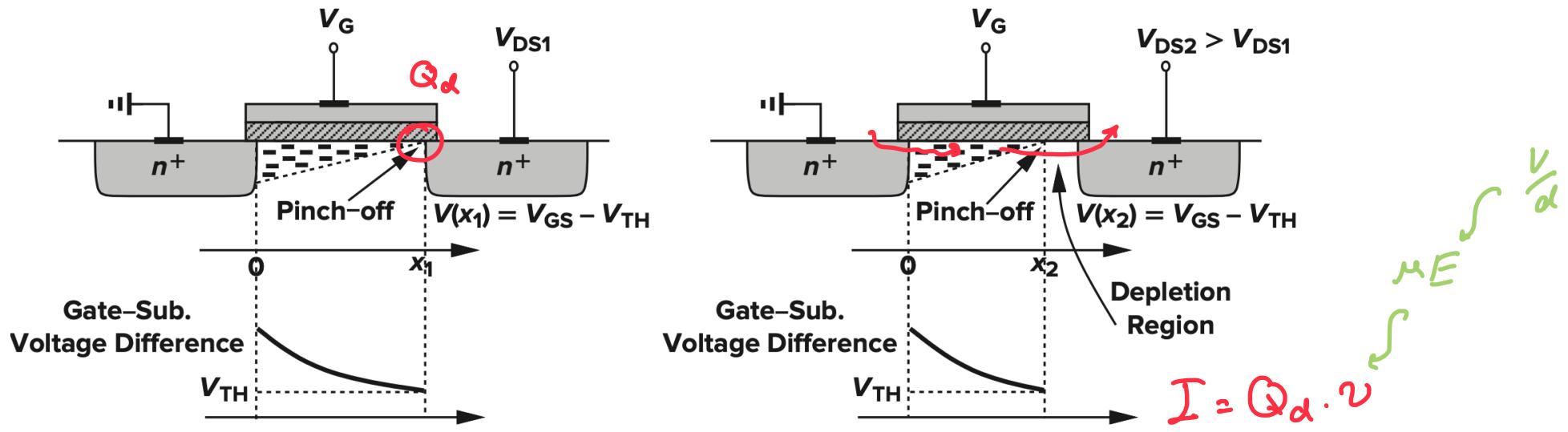
$$R_{on} = \frac{1}{50 \mu\text{A}/\text{V}^2 \times 10(V_G - 1 \text{ V} - 0.3 \text{ V})}.$$

Derivation of I/V Characteristics



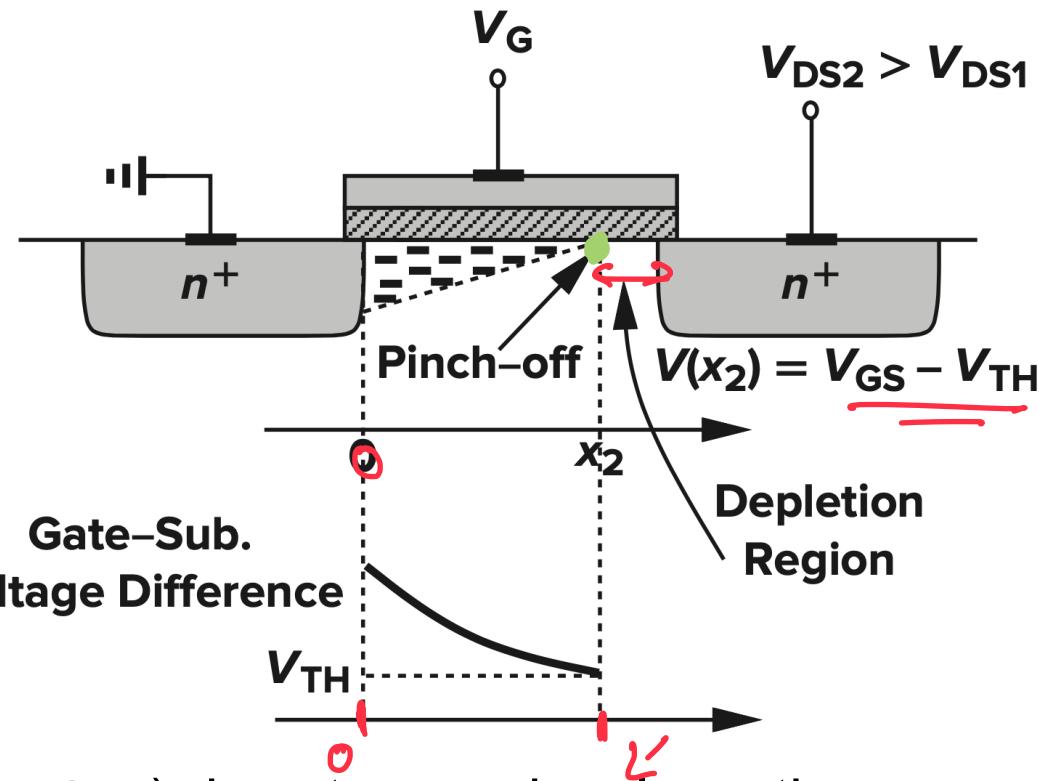
- In reality, if $V_{DS} > V_{GS} - V_{TH}$, I_D becomes relatively constant and we say that the device operates in “saturation region.”
- $V_{D,sat} = V_{GS} - V_{TH}$ denotes the minimum V_{DS} necessary for operation in saturation.

Derivation of I/V Characteristics



- If V_{DS} is slightly larger than $V_{GS} - V_{TH}$, the inversion layer stops at $x \leq L$, and the channel becomes “pinched off.”
- As V_{DS} increases, the point at which Q_D equals zero gradually moves towards the source.
- At some point along the channel, the local potential difference between the gate and the oxide-silicon interface is not sufficient to support an inversion layer.

Derivation of I/V Characteristics



$$E = -\frac{dV}{dx}$$

- Electron velocity ($v = I/Q_d$) rises tremendously as they approach the pinch-off point (where $Q_d \rightarrow 0$) and shoot through the depletion region near the drain junction and arrive at the drain terminal.
- Similar to the flow of water in a river reaching a waterfall!

Derivation of I/V Characteristics

$$\rightsquigarrow V_{GS} = \sqrt{\frac{2I_D}{\mu_n C_{ox} \frac{W}{L'}}} + V_{TH}$$

- Since the integral becomes

$$\int_{x=0}^{x=x_2=L'} I_D dx = \int_{V=0}^{V=V_{GS}-V_{TH}} W C_{ox} \mu_n [V_{GS} - V(x) - V_{TH}] dV$$

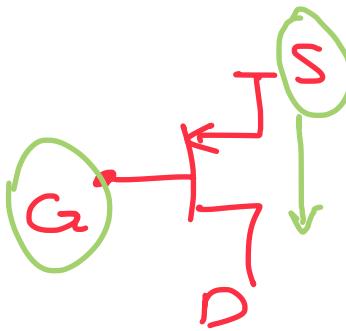
Sat. region

$$I_D = \frac{1}{2} \mu_n C_{ox} \left(\frac{W}{L'} \right) (V_{GS} - V_{TH})^2$$
$$V_{DS} > V_{GS} - V_{TH}$$

- I_D is relatively independent of V_{DS} if L' remains close to L .
- The device exhibits a “square-law” behavior.

$L' \approx L$

Derivation of I/V Characteristics



- For PMOS devices, the equations become

PMOS

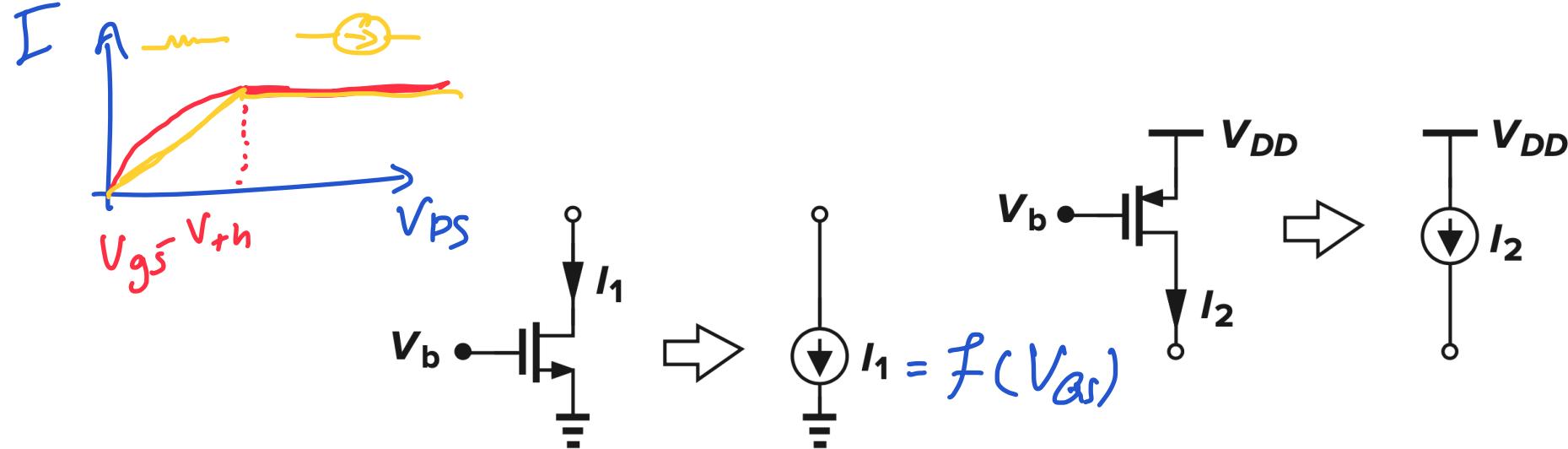
$$\left\{ \begin{array}{l} I_D = -\mu_p C_{ox} \frac{W}{L} \left[(V_{GS} - V_{TH}) V_{DS} - \frac{1}{2} V_{DS}^2 \right] \\ I_D = -\frac{1}{2} \mu_p C_{ox} \frac{W}{L'} (V_{GS} - V_{TH})^2 \end{array} \right.$$

Triode

Sat.

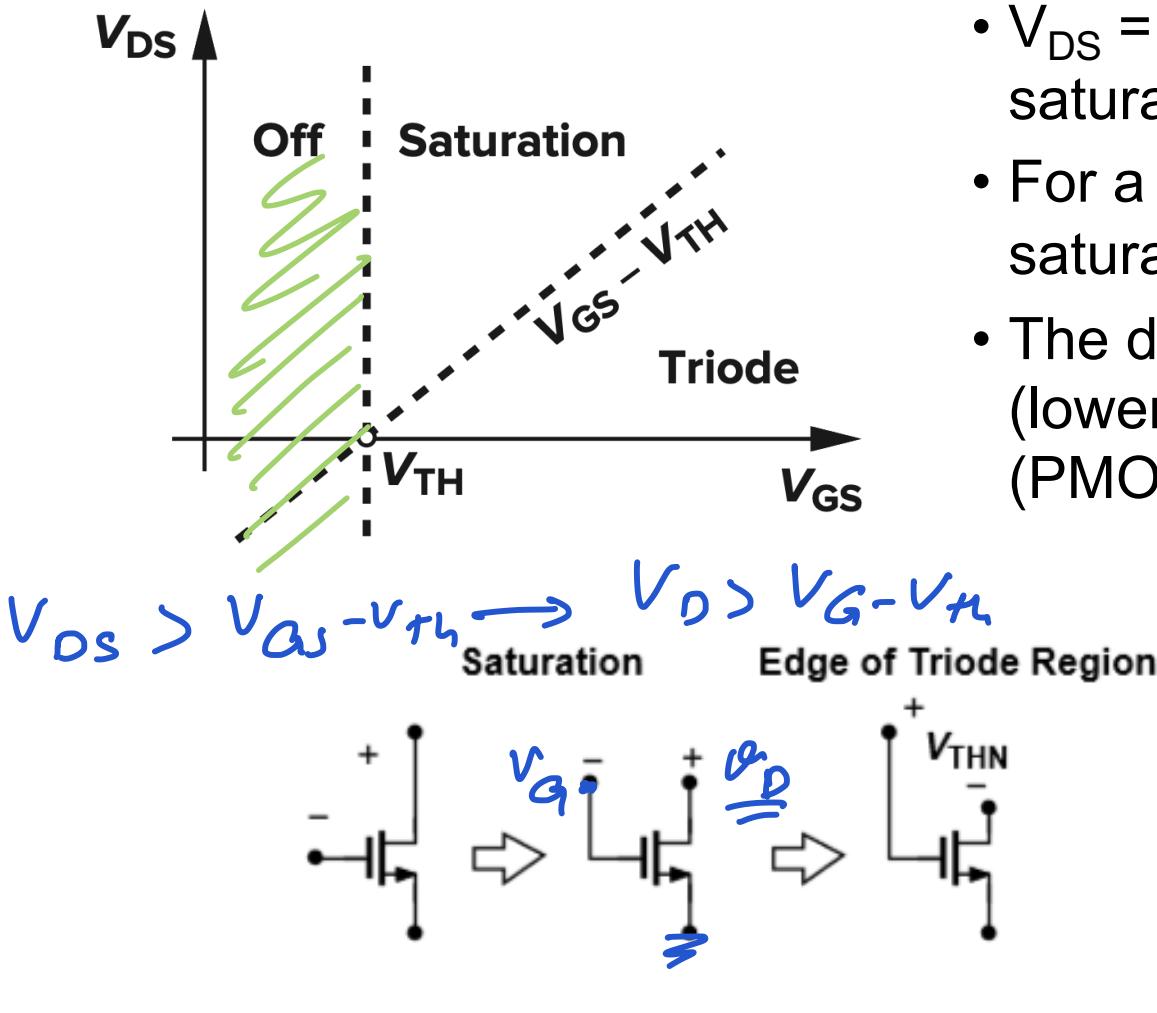
- The negative sign shows up due to the assumption that drain current flows from drain to source, whereas holes in a PMOS flow in the reverse direction.
- V_{GS} , V_{DS} , V_{TH} , and $V_{GS} - V_{TH}$ are negative for a PMOS transistor that is turned on.
- Since the mobility of holes is about $\frac{1}{2}$ the mobility of electrons, PMOS devices suffer from lower “current drive” capability.

Derivation of I/V Characteristics



- A saturated MOSFET can be used as a current source connected between the drain and the source.
- NMOS current sources inject current into ground while PMOS current sources draws current from V_{DD} .

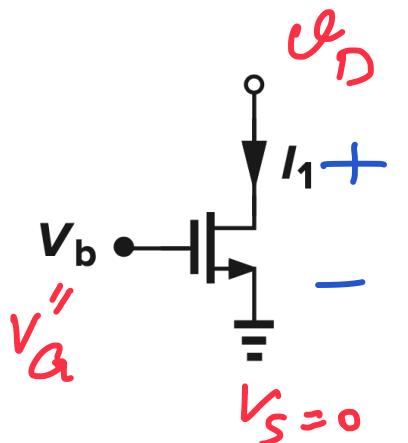
Derivation of I/V Characteristics



- $V_{DS} = V_{GS} - V_{TH} = V_{D,sat}$ is the line between saturation and triode region.
- For a given V_{DS} , the device eventually leaves saturation as V_{GS} increases.
- The drain is defined as the terminal with a higher (lower) voltage than the source for an NMOS (PMOS).

NMOS Example

- Calculate I_1 assuming transistor is in the Saturation Region, and ...
 - $L = 65\text{nm}$, $W = 1\mu\text{m}$, $\mu_n C_{ox} = 0.5\text{mA/V}^2$
 - $V_{th} = 0.6\text{V}$, $V_b = 1\text{V}$



$$I_D = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{th})^2$$

$\xrightarrow{0.9V}$

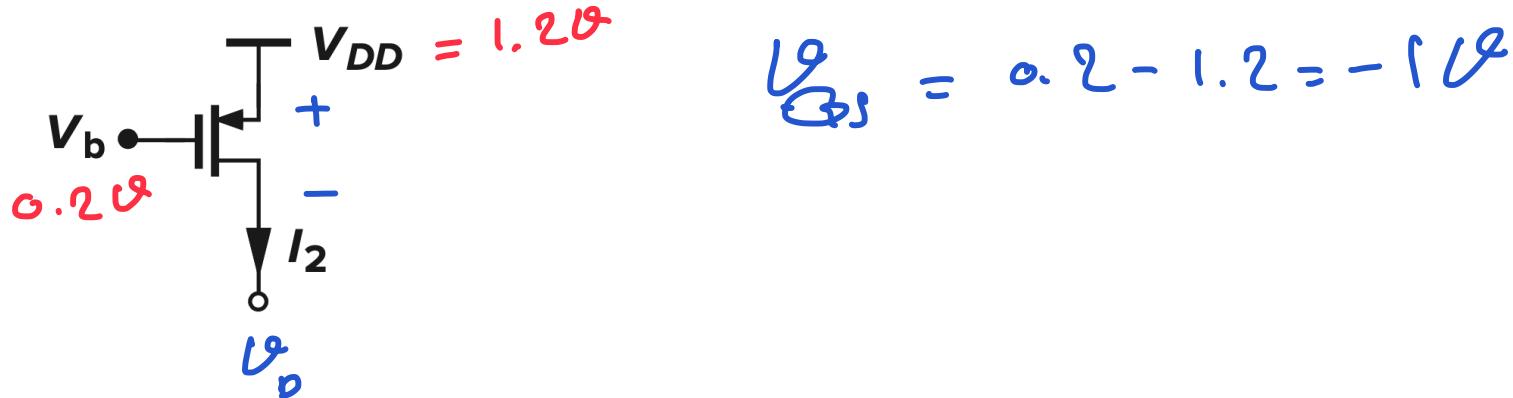
$$\Rightarrow V_{GS} - V_{th} = \sqrt{\frac{2I}{\mu_n C_{ox} \frac{W}{L}}}$$

- What are the conditions for the device to operate in saturation?

$$V_D > 1V - 0.6V = 0.4V$$

PMOS Example

- Calculate I_1 assuming transistor is in the Saturation Region, and ...
 - $L = 65\text{nm}$, $W = 1\mu\text{m}$, $\mu_p C_{ox} = 0.25\text{mA/V}^2$
 - $V_{th} = -0.6\text{V}$, $V_{DD} = 1.2\text{V}$, $V_b = 0.2\text{V}$



- What are the conditions for the device to operate in saturation?

$$V_D < |V_{GS}| - |V_{th}| = 0.4\text{V}$$

MOS Transconductance

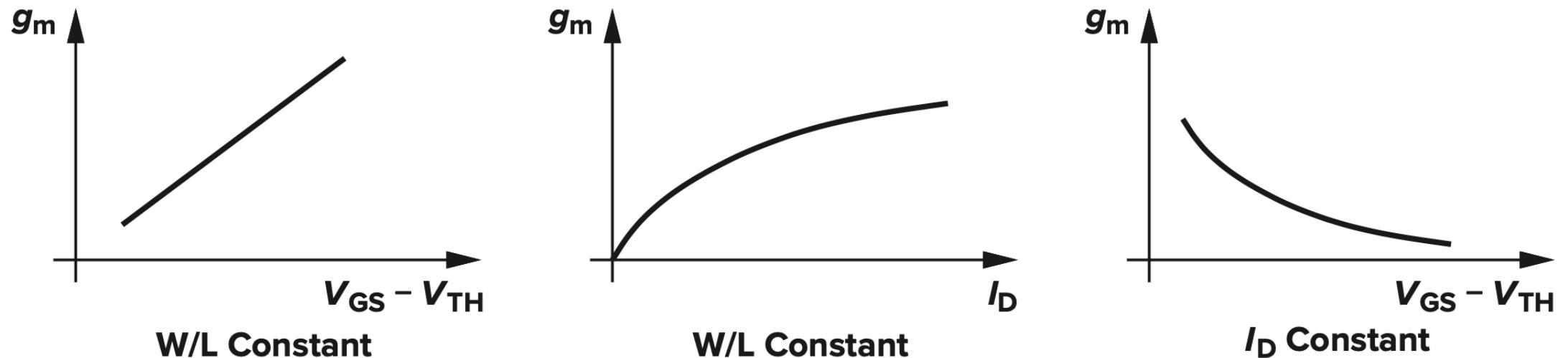
$$g_{ds} = \frac{\partial I_D}{\partial v_{DS}} \Big|_{v_{DS, cur}} = \frac{1}{r_o}$$

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \Big|_{V_{DS} \text{ const.}} = \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})$$
$$= \sqrt{2 \mu_n C_{ox} \frac{W}{L} I_D} = \frac{2 I_D}{V_{GS} - V_{TH}} = \frac{e \sum_D}{V_{OD}}$$

- Transconductance (usually defined in the saturation region) is defined as the change in drain current divided by the change in the gate-source voltage.
 - g_m represents the sensitivity of the device since a high value implies a small change in V_{GS} will result in a large change in I_D .
 - Transconductance in saturation region is equal to the inverse of R_{on} in the deep triode region.

$$g_m \text{ (triode)} = \frac{1}{R_{on}}$$

MOS Transconductance



- Each expression for transconductance is useful in studying its behavior.
- Drain current and overdrive voltage are *bias* values.
- If a small signal is applied to a device with defined bias values, we assume the signal amplitude is small enough that the variation in transconductance is negligible.

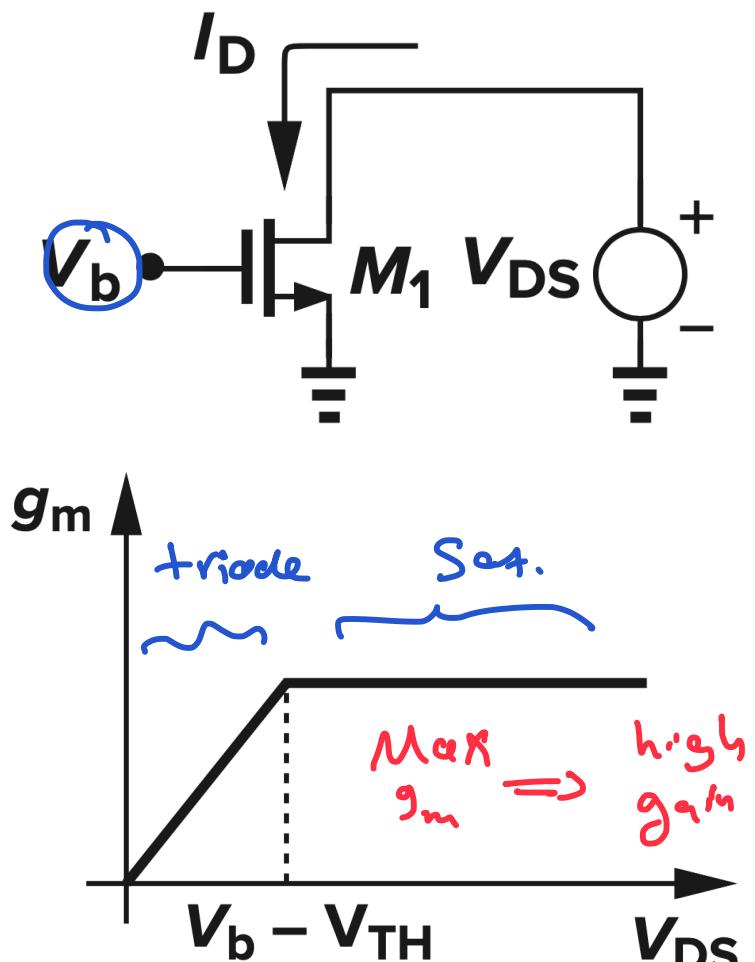
MOS Transconductance

$$i_d = g_m v_b$$

- To find the transconductance for the topology on the left with respect to V_{DS} ,
 - So long as $V_{DS} \geq V_b - V_{TH}$, M_1 is in saturation, so I_D is relatively constant, and therefore so is g_m .
 - When M_1 enters triode region ($V_{DS} < V_b - V_{TH}$),

$$\begin{aligned} g_m &= \frac{\partial}{\partial V_{GS}} \left\{ \frac{1}{2} \mu_n C_{ox} \frac{W}{L} [2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2] \right\} \\ &= \mu C_{ox} \frac{W}{L} V_{DS} \end{aligned}$$

- We are mostly (in this course) interested in biasing the circuit such that all of MOS devices will be in the saturation regime (why?)



MOS Transconductance

- For PMOS,

$$g_m = \frac{\partial I_D}{\partial V_{GS}}$$

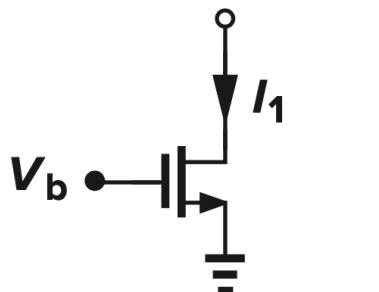
$$\begin{aligned} g_m &= -\mu_p C_{ox} \left(\frac{W}{L} \right) (V_{GS} - V_{TH}) \\ &= -\frac{2I_D}{V_{GS} - V_{TH}} \\ &= \sqrt{2\mu_p C_{ox} \left(\frac{W}{L} \right) I_D} \end{aligned}$$

- Why the g_m is similar to NMOS (i.e., not negative)?

g_m Example

- Calculate g_m assuming transistor is in the Saturation Region, and ...

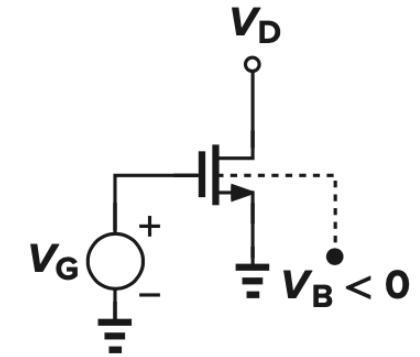
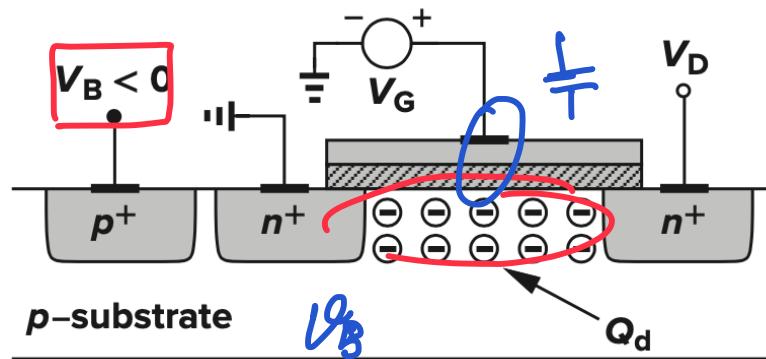
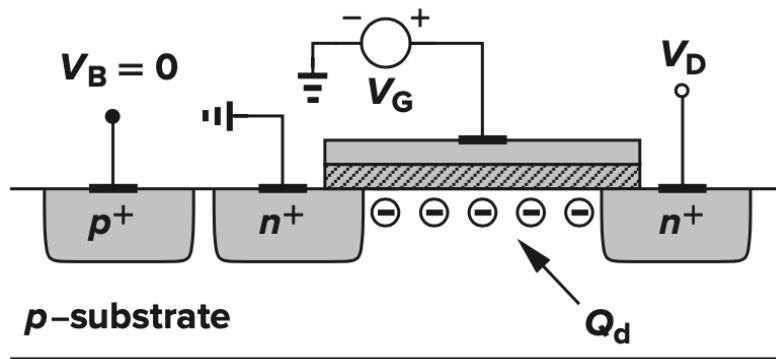
- $L = 65\text{nm}$, $W = 1\mu\text{m}$, $\mu_n C_{ox} = 0.5\text{mA/V}^2$
- $V_{th} = 0.6\text{V}$, $V_b = 1\text{V}$



$$g_m = ?$$

- What should be the unit for g_m ? $1/\text{ohm} = \text{S}$

Second-Order Effects (Body effect)



- Originally, with the bulk of an NMOS tied to ground, the threshold voltage was defined as

$$V_{TH} = \Phi_{MS} + 2\Phi_F + \frac{Q_{dep}}{C_{ox}}$$

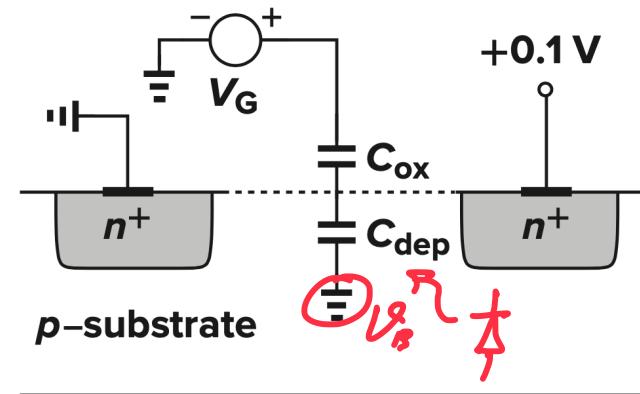
$\rightarrow Q_{dep} \uparrow \Rightarrow V_{TH} \uparrow$

- Decreasing the bulk voltage (V_B) increases the number of holes attracted to the substrate connection, which leaves a larger negative charge behind and makes the depletion region wider, increasing Q_{dep} and thus increasing V_{TH} .
- This is known as the “body effect” or “back-gate effect.” [This turns out to be confusing!!!!]

Second-Order Effects (Body effect)

$$Q_{dep} = \sqrt{4q\epsilon_{si}|\Phi_F|N_{sub}}$$

Sub dep. \rightarrow Fermi Voltage $\rightarrow \Phi_F + V_{SB}$



- With body effect, the expression which characterizes the dependence of threshold voltage on the bulk voltage is

$$V_{TH} = V_{TH0} + \gamma(\sqrt{2\Phi_F + V_{SB}} - \sqrt{|2\Phi_F|})$$

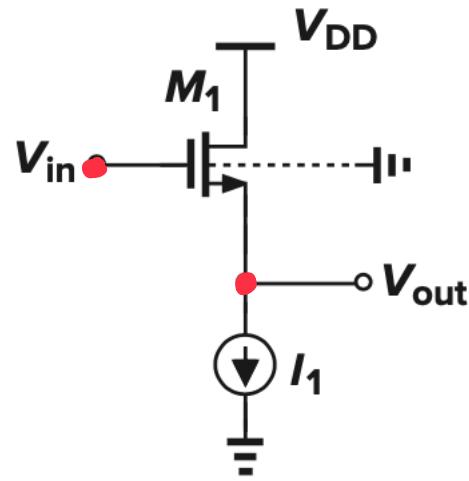
- Where,

$$\left\{ \begin{array}{l} - V_{TH0} = \Phi_{MS} + 2\Phi_F + \frac{Q_{dep}}{C_{ox}} \\ - \gamma = \sqrt{2q\epsilon_{si}N_{sub}}/C_{ox} \end{array} \right.$$

$V_{SB} = 0$

denotes the body effect coefficient.

Second-Order Effects (Body effect)



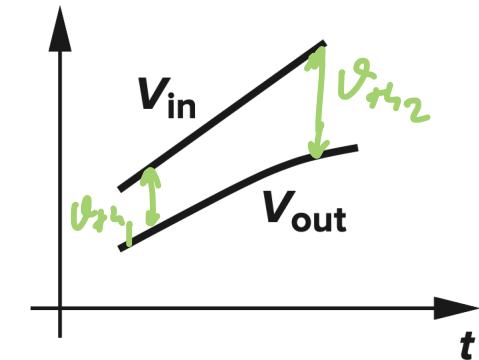
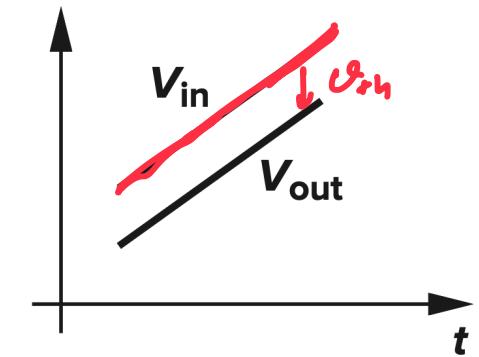
$$\vartheta_{SB} = V_{out+} - \sigma_2 V_{out}$$

- With body effect, as $V_{in,out}$ become more positive, V_{SB} increases, which increases V_{TH} and thus $V_{in} - V_{out}$ must increase to maintain a constant I_D .

$$I_1 = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} (V_{in} - V_{out} - V_{TH})^2$$

$\xrightarrow{\text{Simplifying}} V_{out+} = V_{in} - V_{TH}$

- Body effect manifests itself whenever the source voltage varies with respect to the bulk potential.
- Given the topology on the left and first ignoring body effect, as V_{in} varies, V_{out} follows the input because the drain current remains equal to I_1 , where



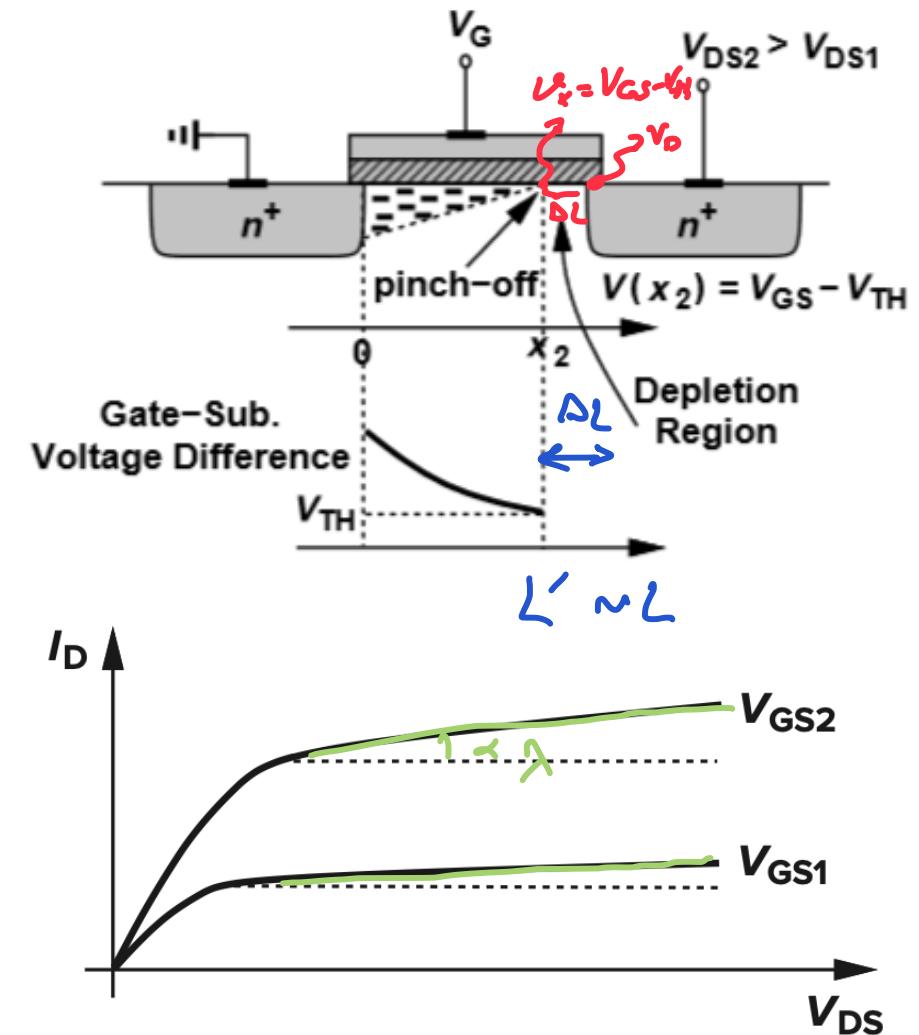
Second-Order Effects (Channel length modulation)

- Originally, when the device was in saturation region, drain current was characterized by

$$\frac{1}{L - \Delta L} \approx \frac{1 + \Delta V_{DS}}{L} \quad I_D = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})^2$$

- The actual length of the channel ($L' = L - \Delta L$) is a function of V_{DS} , which is an effect called “channel length modulation.”
- $1/L' \approx (1 + \Delta L/L)/L$, and $\Delta L/L = \lambda V_{DS}$, where λ is the channel-length modulation coefficient. Thus:

$$I_D \approx \frac{1}{2} \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})^2 (1 + \lambda V_{DS})$$



Second-Order Effects (Channel length modulation)

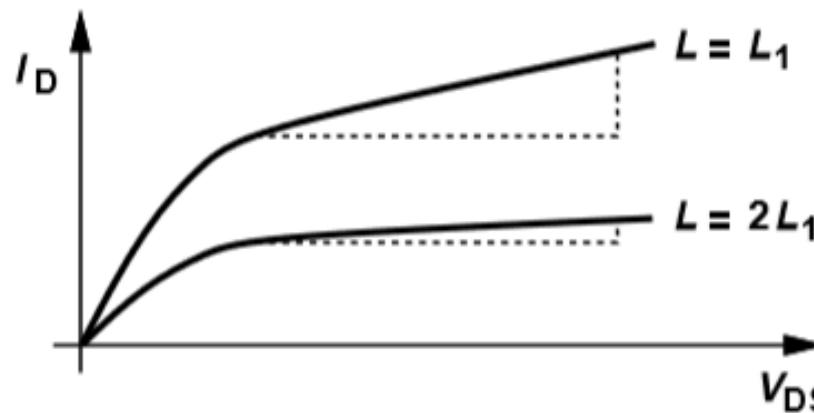
- With the effect of channel length modulation, the expressions derived for transconductance of the device that need modification are

$$\begin{aligned}\frac{\partial I_D}{\partial V_{GS}} &= g_m = \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH}) \underbrace{(1 + \lambda V_{DS})}_{\text{red}} \\ &= \sqrt{2\mu_n C_{ox} (W/L) I_D (1 + \lambda V_{DS})}\end{aligned}$$

- Notice g_m will be still as shown below, once you consider $(1 + \lambda V_{DS})$ factor in the new I_D :

$$= \sqrt{2\mu_n C_{ox} \frac{W}{L} I_D} = \boxed{\frac{2I_D}{V_{GS} - V_{TH}}}.$$

Second-Order Effects (Channel length modulation)



- Knowing that

$$I_D = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})^2 (1 + \lambda V_{DS}) , \rightarrow \text{slope } \propto \frac{1}{L} \cdot \lambda$$

$\lambda \propto 1/L$

and keeping all other parameters constant, we can see that if the length L is doubled, the slope of I_D vs. V_{DS} is divided by *four*.

- This is due to $\partial I_D / \partial V_{DS} \propto \lambda / L \propto 1 / L^2$.