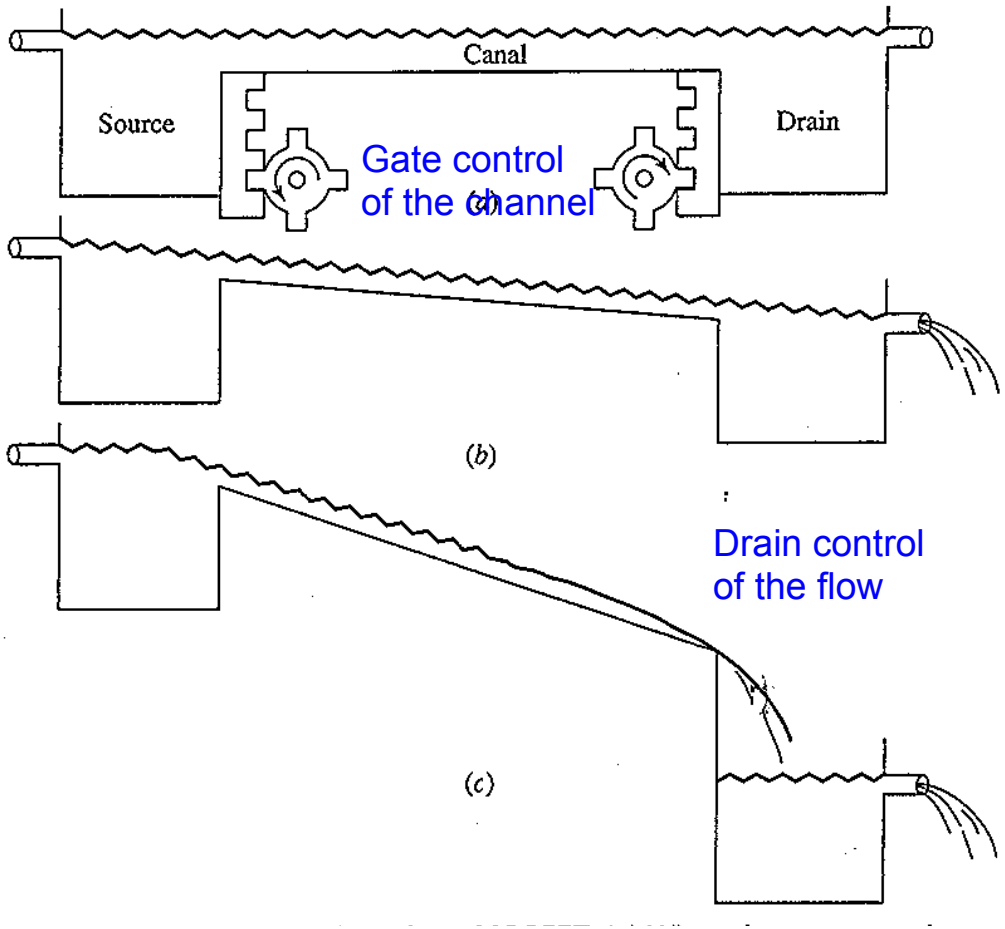
ECE 5205 Spring 2014

MOSFET OPERATION Part 1

Marius Orlowski
Virginia Polytechnic Institute and State University

Fluid Analogy to MOSFET Operation



- When the source and drain are level, there is no flow $V_{DS}=0$
- Whatever depth in the canal can be varied by the gear and track (V_{GS})
- When the drain is lower than the source, water flows along the canal
- The flow is limited by the channel capacity, lowering the drain further only increases the height of the waterfall at its edge

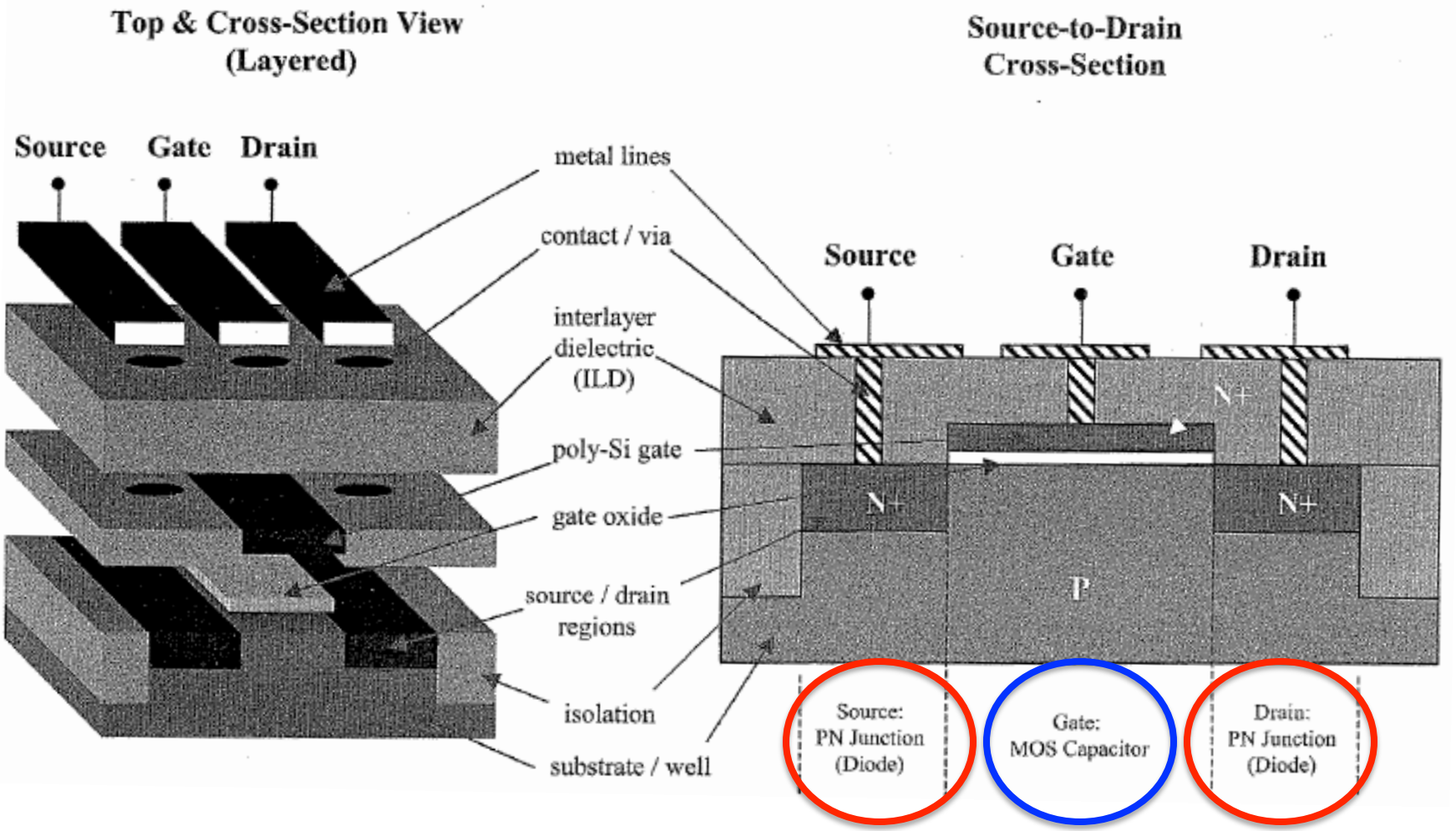
MOSFET Operation

Contents Part 1

- Overview (MOSFET Current Regimes)
- Inversion charge in a MOSFET
- Pao-Sah Double Integral Model
- Gradual channel approximation
- Charge-Sheet Approximation
- MOSFET I-V Characteristics
- Subthreshold Current
- Subthreshold Swing – a critical issue today

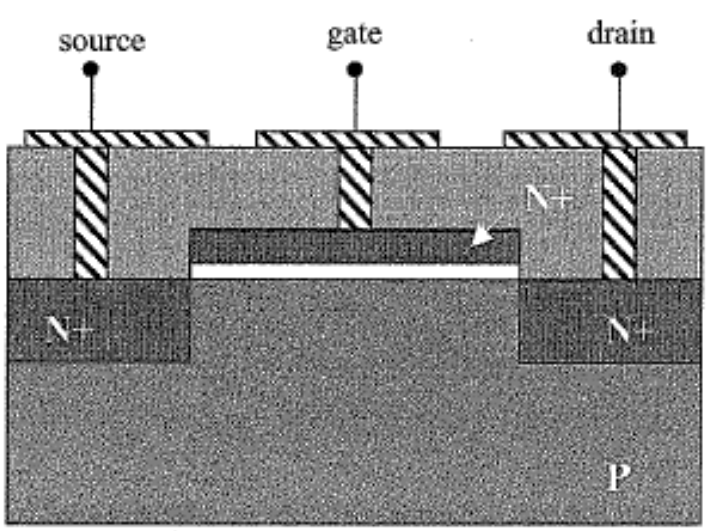
MOSFET - Layer/Feature ID

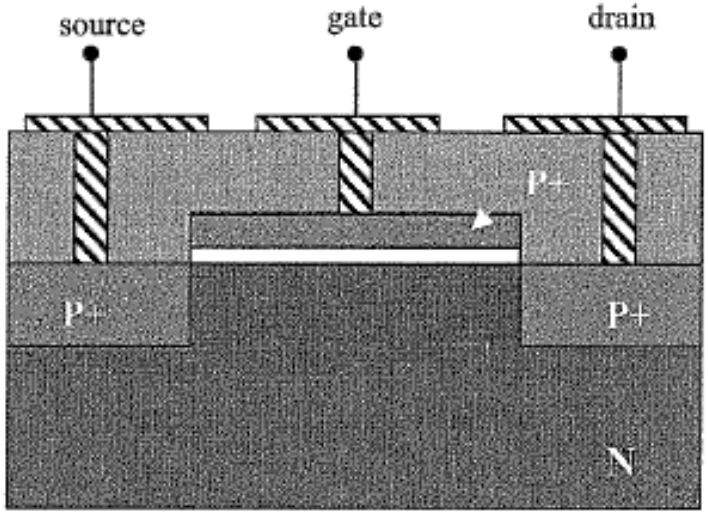
Metal-Oxide-Semiconductor Field-Effect Transistor



MOSFET - Two Types

device structure

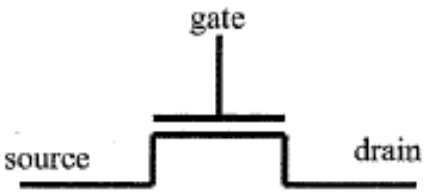


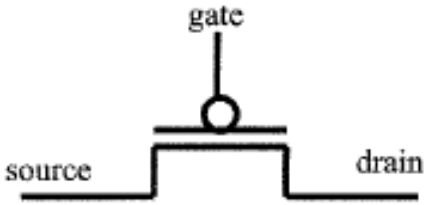


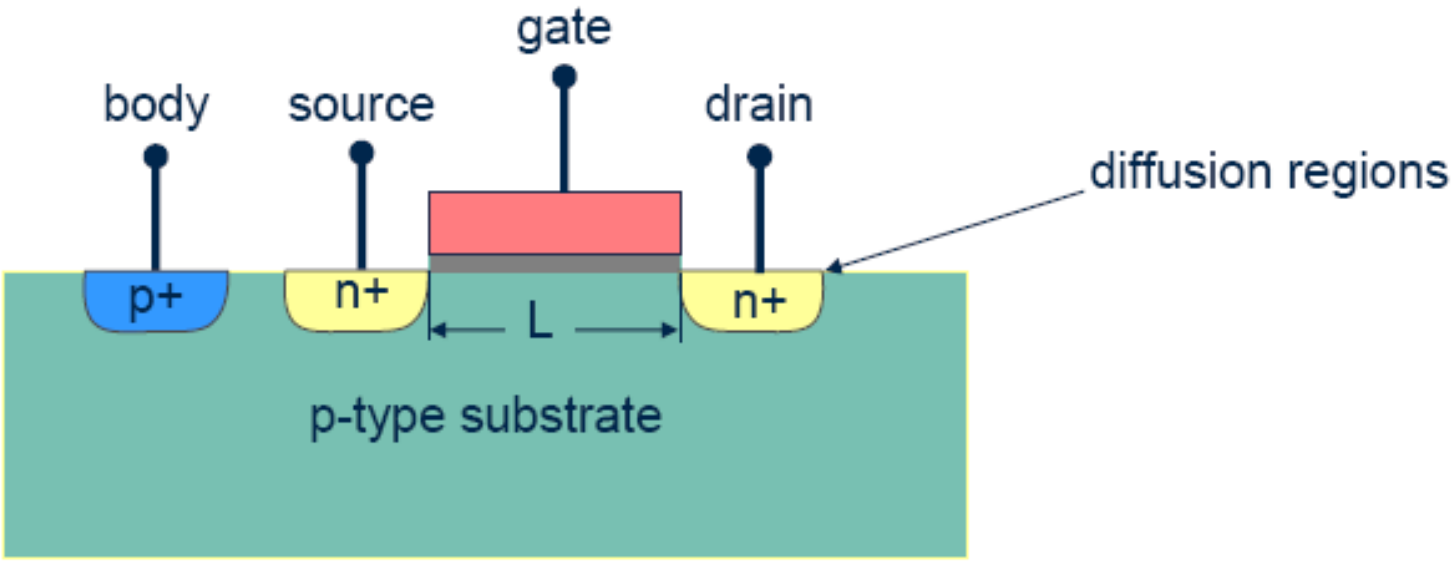
N-Channel Transistor

P-Channel Transistor

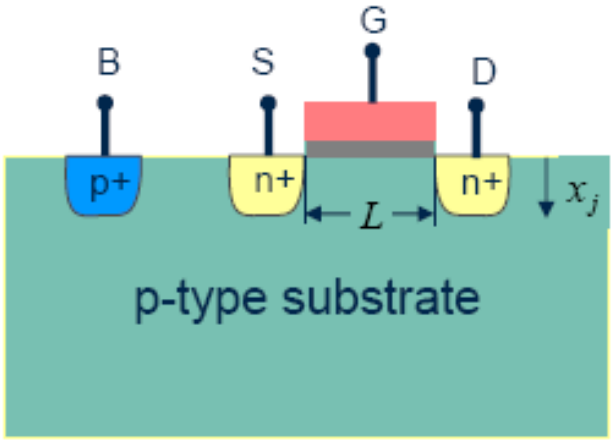
circuit symbol

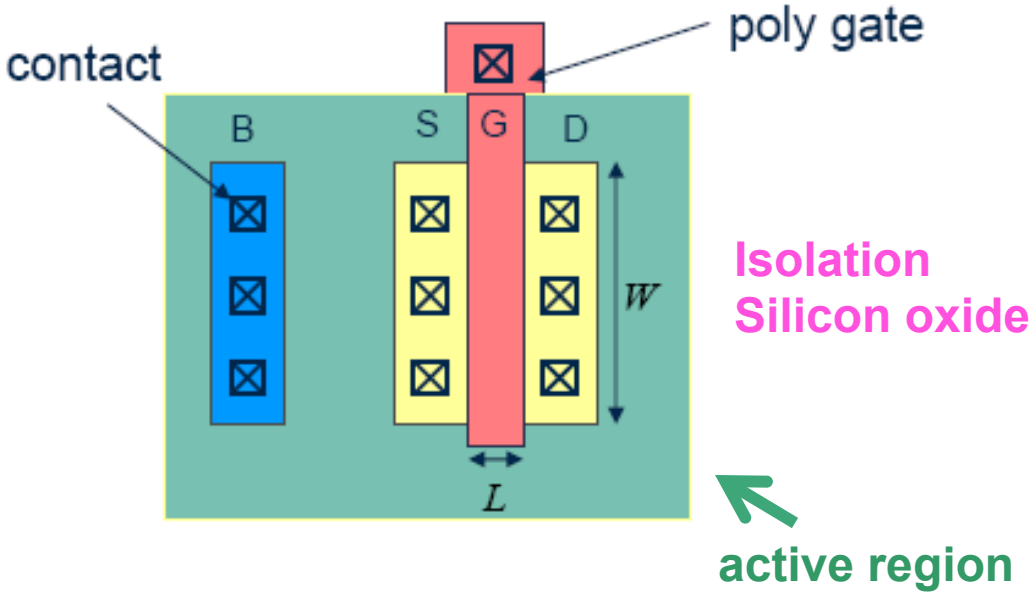






- Add two junctions around MOS capacitor
- The regions forms PN junctions with substrate
- MOSFET is a four terminal device
- The body is usually grounded (or at a DC potential)
- For ICs, the body contact is at surface

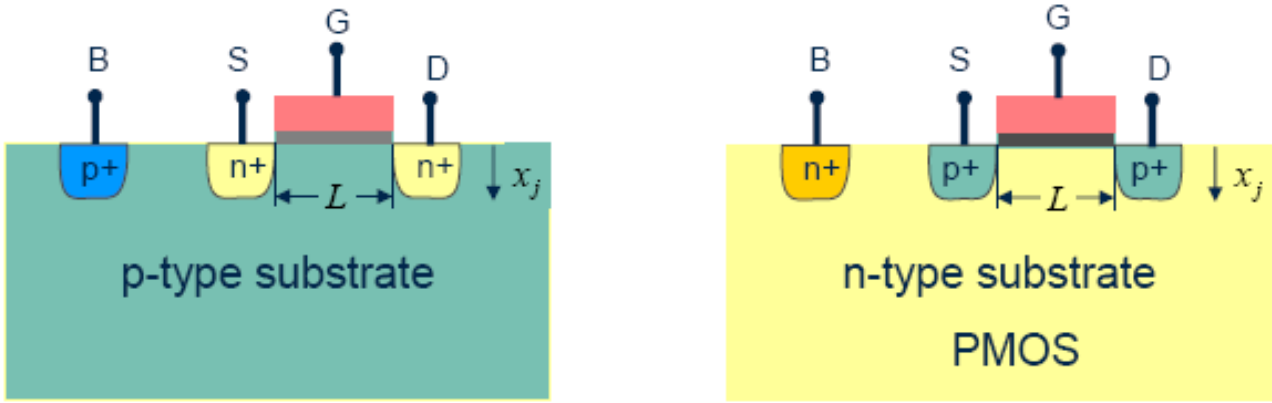




- **Planar process: complete structure can be specified by a 2D layout**
- **Design engineer can control the transistor width *W* and *L***
- **Process engineer controls t_{ox}, N_a, x_j, etc.**

Gate is always contacted outside active region (or on isolation)

PMOS & NMOS

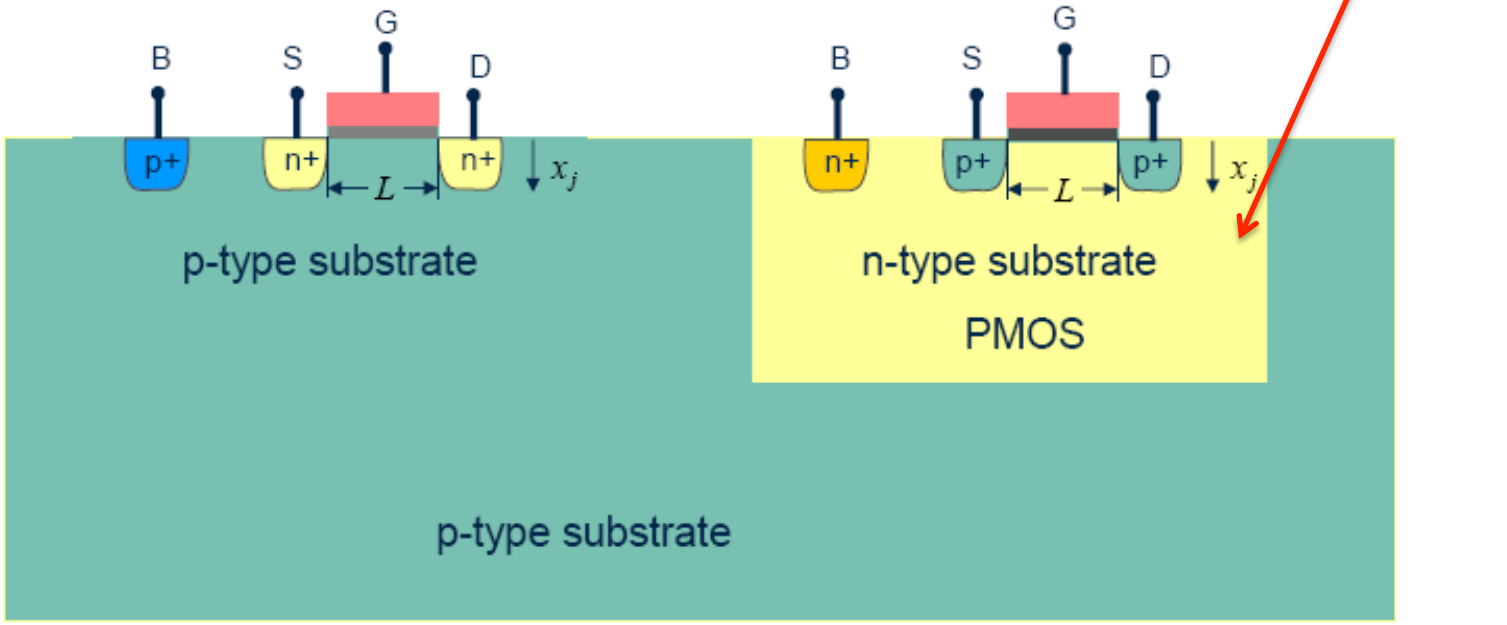


- **A MOSFET by any other name is still a MOSFET:**
  - NMOS, PMOS, nMOS, pMOS
  - NFET, PFET
  - IGFET Insulated gate field effect transistor
  - Other flavors: JFET, MESFET junction gate field-effect transistor
- **CMOS technology: The ability to fabricate NMOS and PMOS devices simultaneously**

MESFET metal semiconductor field effect transistor

in a p-substrate wafer PMOSFET is embedded in an n-well

CMOS



- Complementary MOS: Both P and N type devices
- Create a n-type body in a p-type substrate through compensation. This new region is called a "well".
- To isolate the PMOS from the NMOS, the well must be reverse biased (pn junction)

Four types of MOSFETs

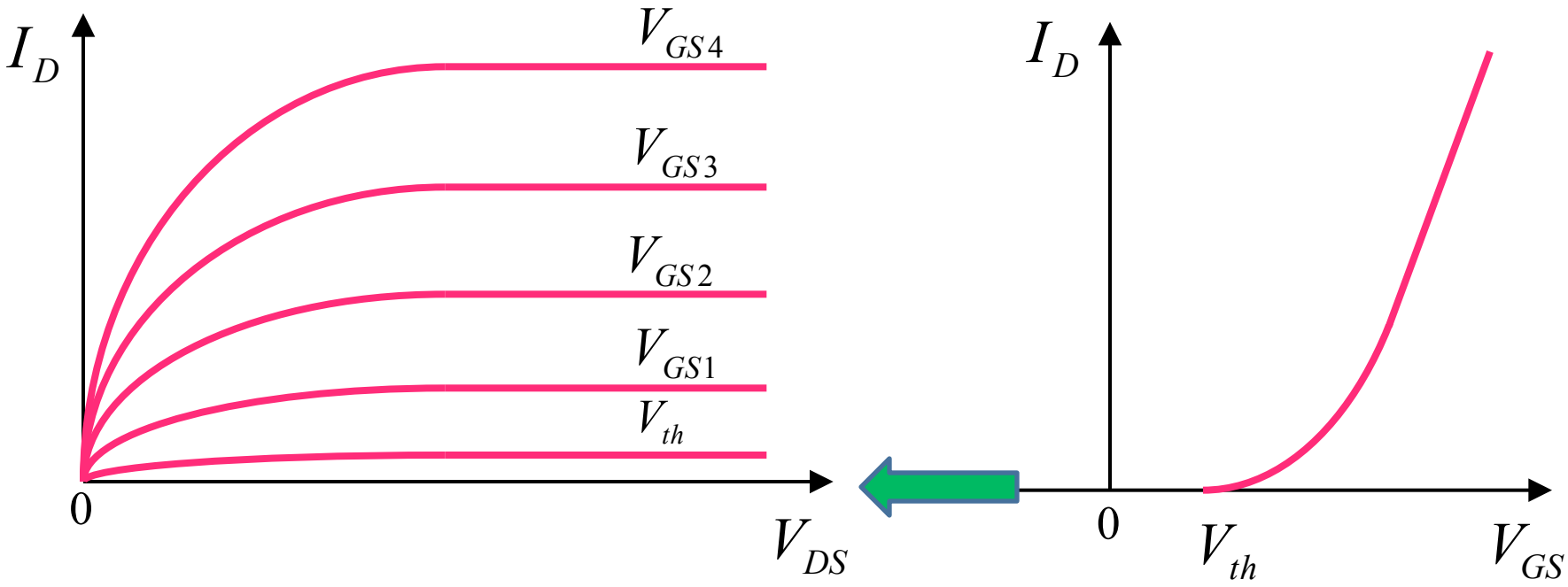
n-channel **enhancement** MOSFET

dc output characteristics

$$I_D = f\left(V_{DS}\right)\Big|_{V_{GS}}$$

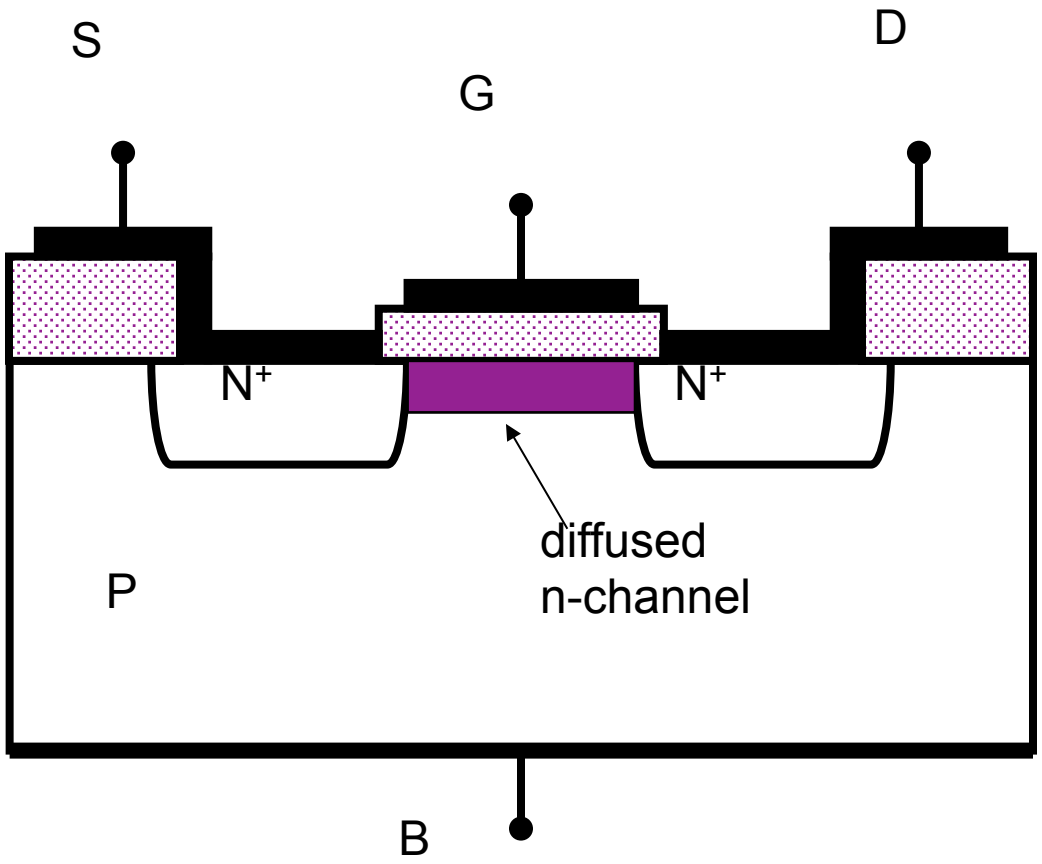
dc transfer characteristics

$$I_D = f\left(V_{GS}\right)\Big|_{V_{DS}}$$



NMOS: a positive Gate voltage has to be applied to turn the transistor on. The channel has to be “enhanced”. Transistor is “normally off”

n-channel **depletion** MOSFET



This transistor is “normally on”.
A negative gate voltage has to be applied to turn the transistor off.

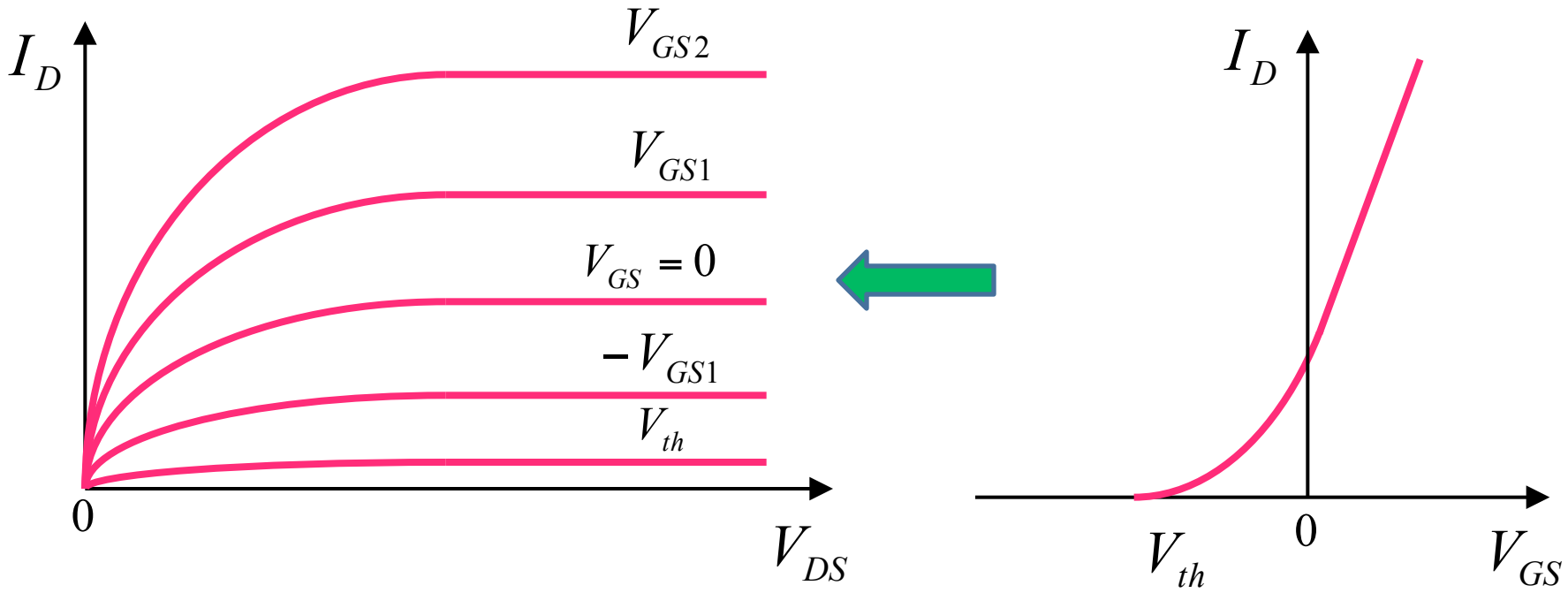
n-channel **depletion** MOSFET

dc output characteristics

$$I_D = f\left(V_{DS}\right)\Big|_{V_{GS}}$$

dc transfer characteristics

$$I_D = f\left(V_{GS}\right)\Big|_{V_{DS}}$$

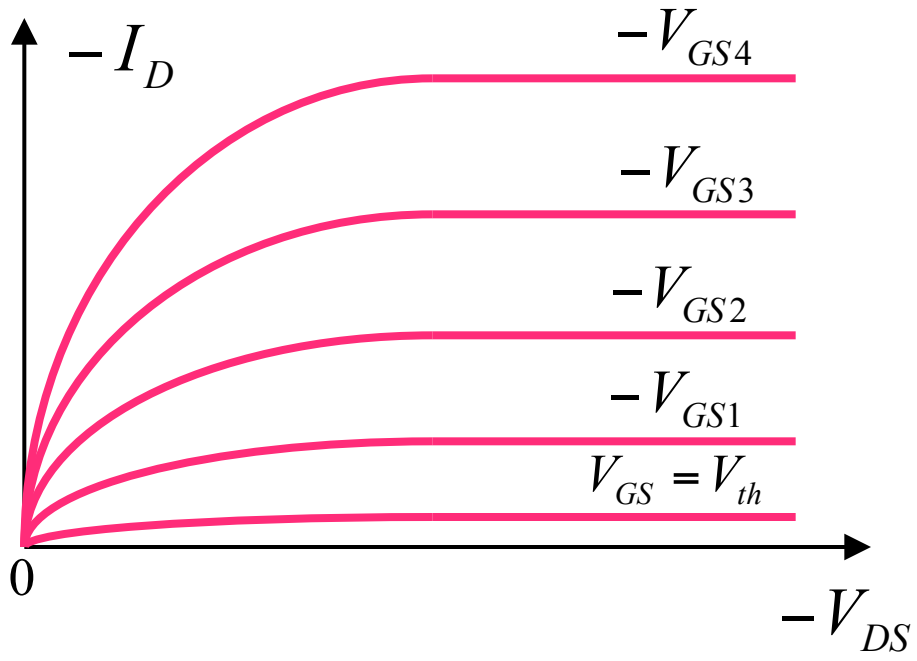


Transistor “normally on”. A negative gate voltage has to be applied to turn the transistor off.

p-channel **enhancement** MOSFET

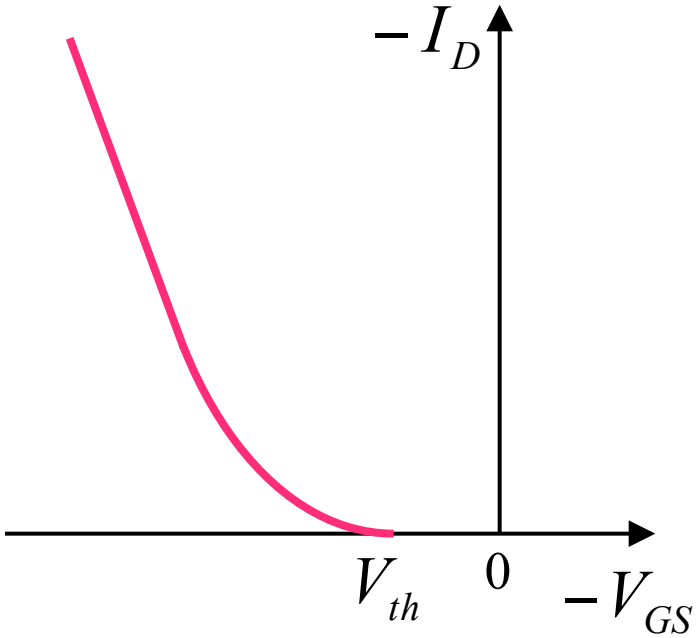
dc output characteristics

$$I_D = f\left(V_{DS}\right)\Big|_{V_{GS}}$$



dc transfer characteristics

$$I_D = f\left(V_{GS}\right)\Big|_{V_{DS}}$$

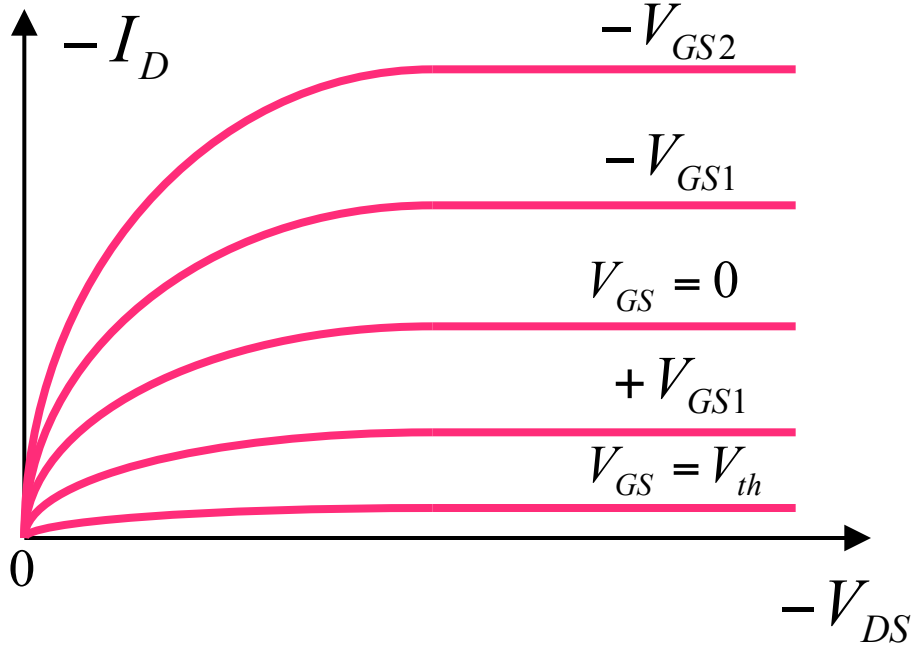


PMOS: a negative Gate voltage has to be applied to turn the transistor on. The channel has to be “enhanced”. Transistor is “normally off”

p-channel **depletion** MOSFET

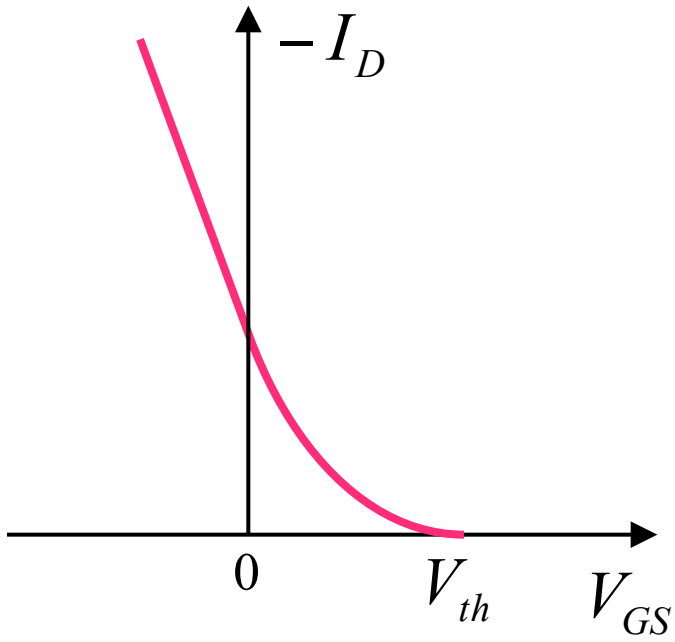
dc output characteristics

$$I_D = f(V_{DS})\Big|_{V_{GS}}$$



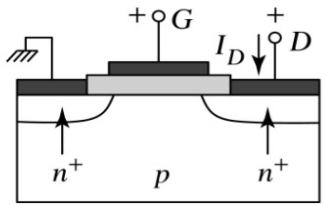
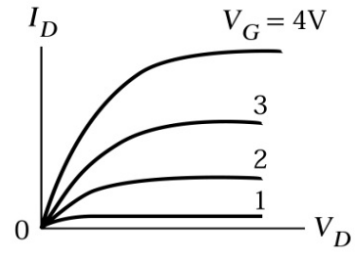
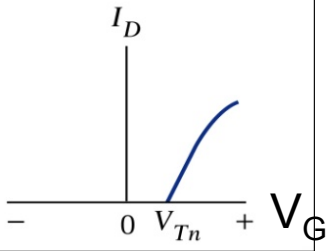
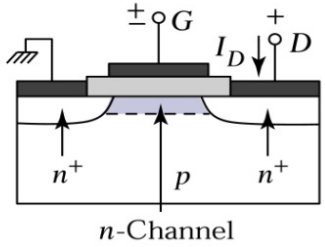
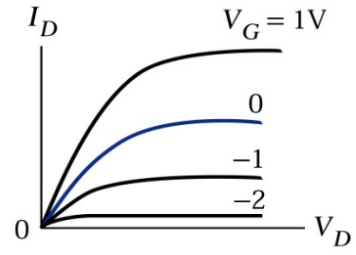
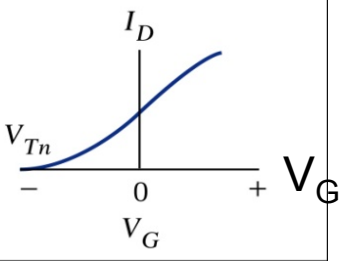
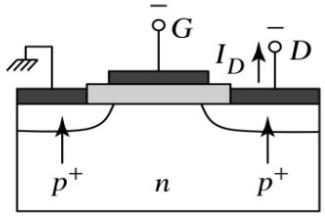
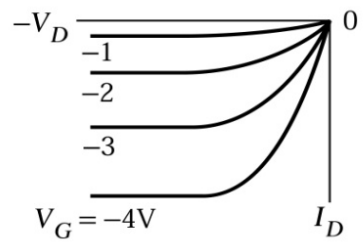
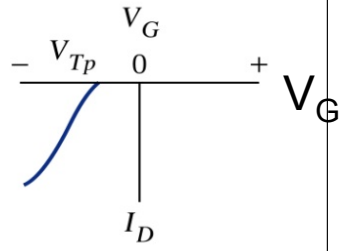
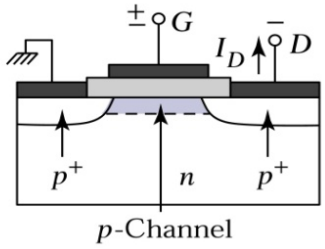
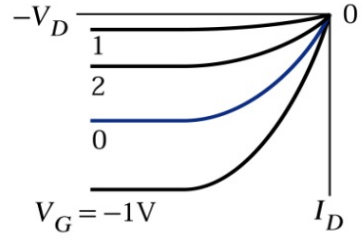
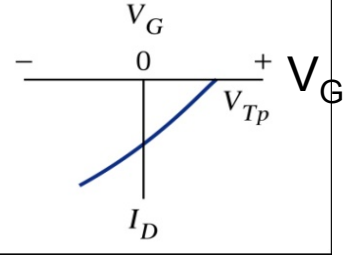
dc transfer characteristics

$$I_D = f(V_{GS})\Big|_{V_{DS}}$$

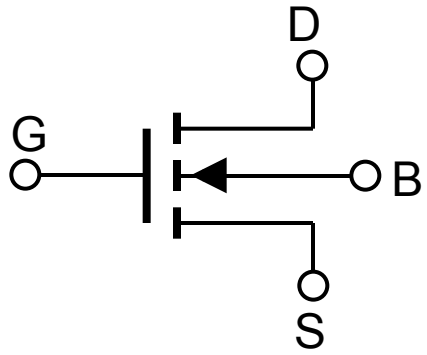
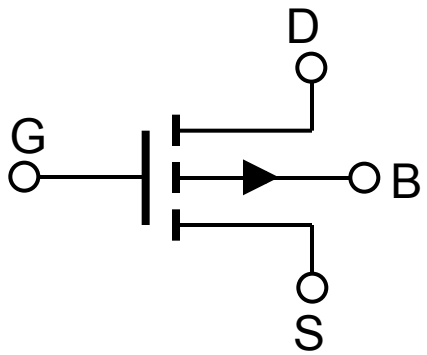
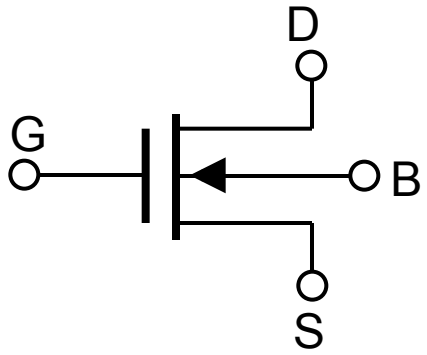
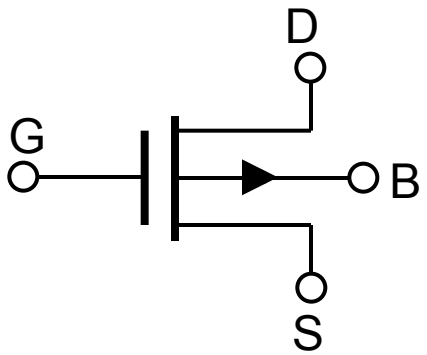


Transistor "normally on". A positive gate voltage has to be applied to turn the transistor off.

Summary: Four types of MOSFETs

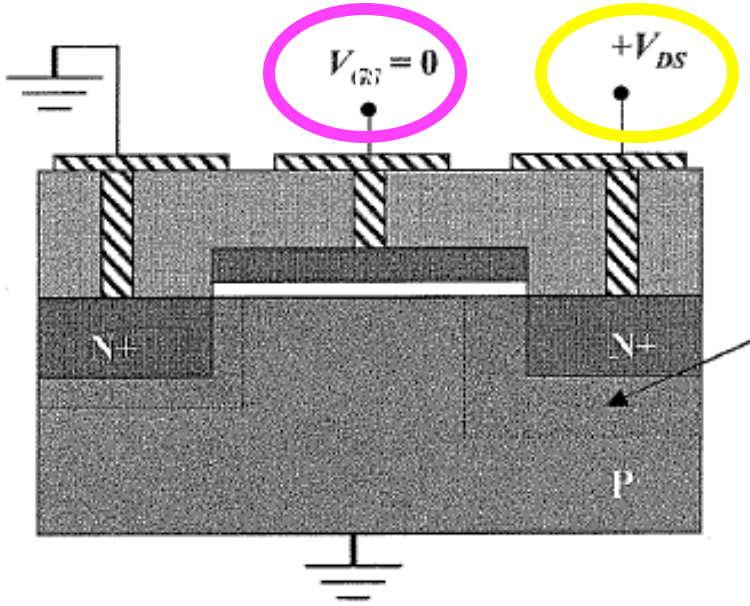
Type	Cross Section	Output Characteristics	Transfer Characteristics
n-Channel Enhancement (Normally Off)			
n-Channel Depletion (Normally On)			
p-Channel Enhancement (Normally Off)			
p-Channel Depletion (Normally On)			

MOSFET circuit symbols

	n-channel	p-channel
enhancement		
depletion		

The Basics - Turning the MOSFET On & Off (cont.)

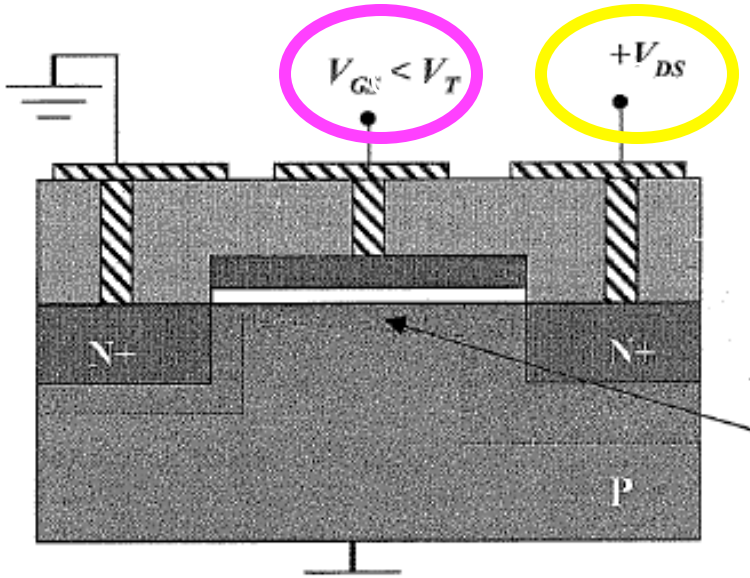
No Applied Gate Voltage
MOSFET is "OFF"



reverse-biased PN junction:
depletion region widens,
only leakage current

no current flow

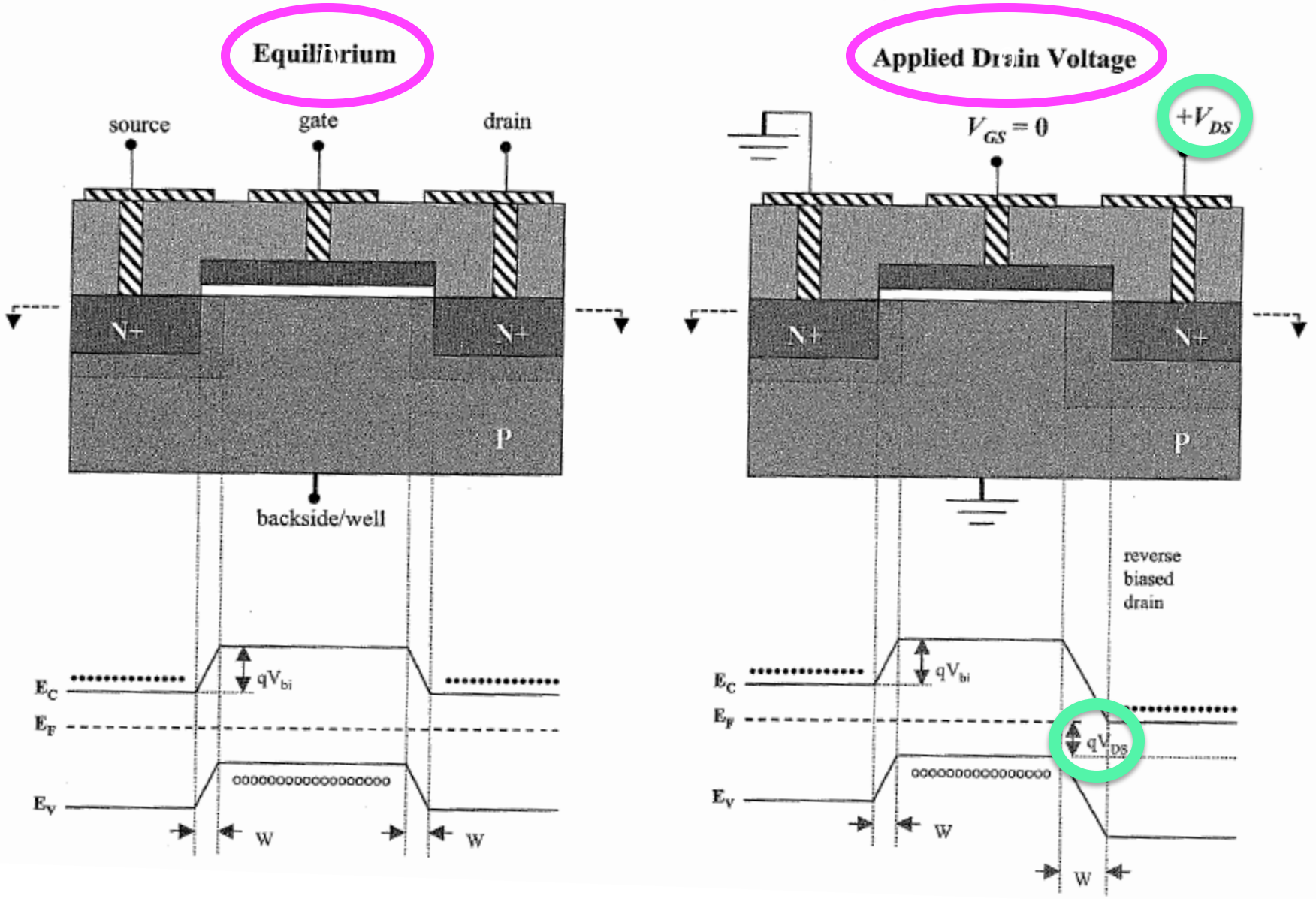
Subthreshold
MOSFET is "OFF"



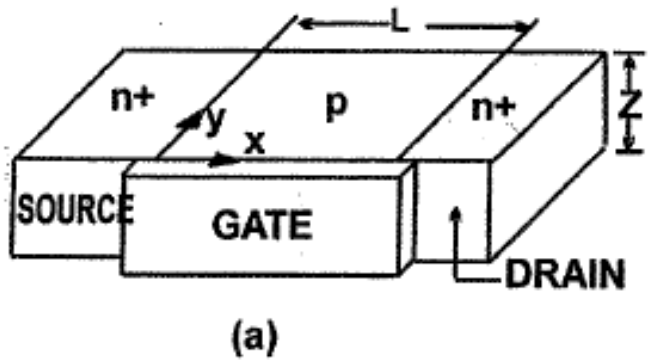
MOS capacitor is
biased into depletion,
only leakage current

no current flow

MOSFET - Band Diagram

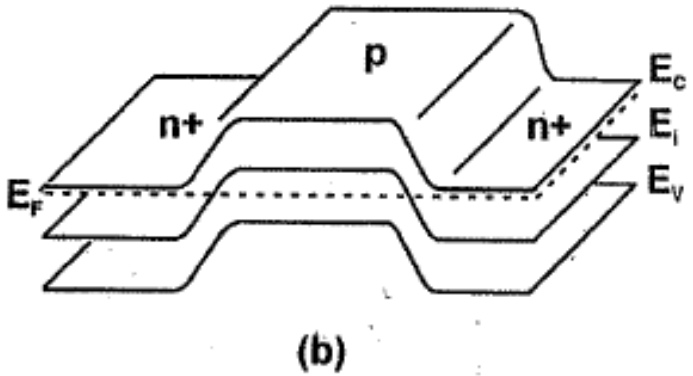


MOSFET - 3-D Band Diagram



(a)

Gate bias to turn MOS on



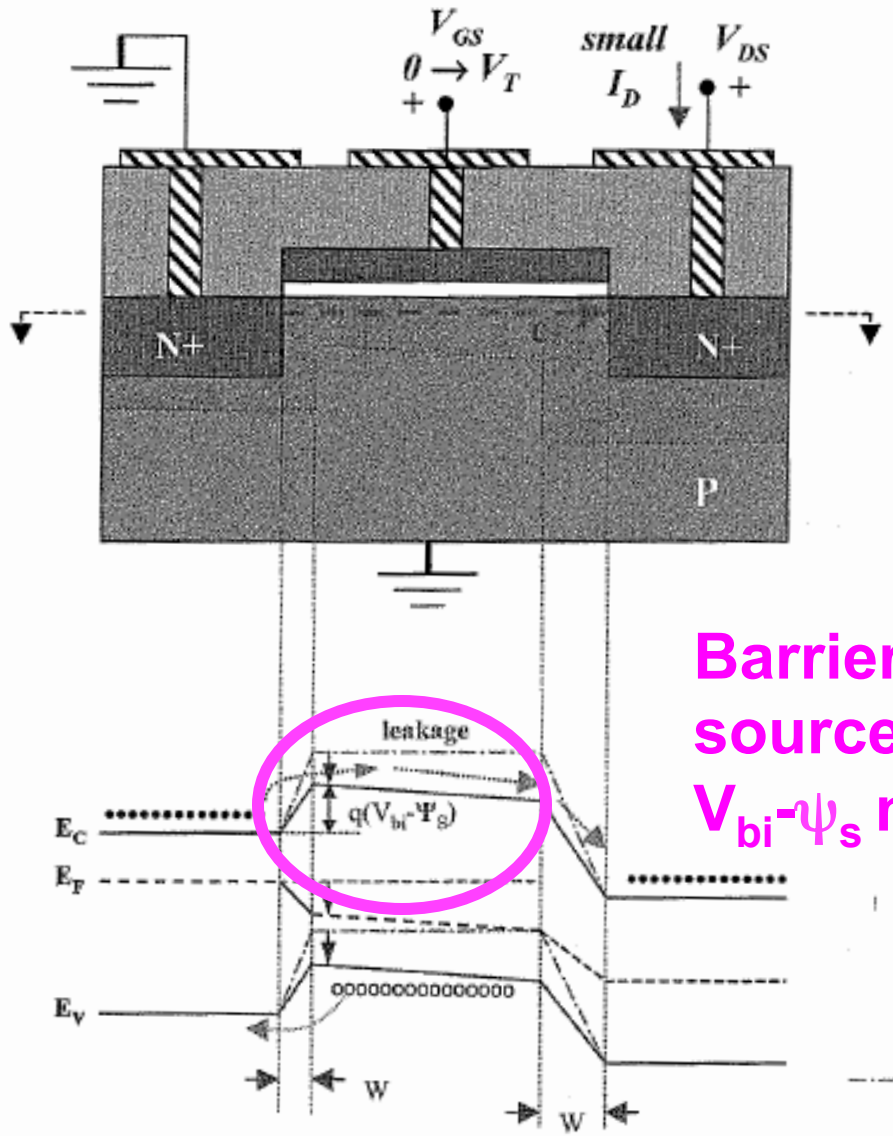
(b)

Gate bias ON and drain bias

MOSFET - Subthreshold

MOSFET OFF
"no current flow"
but…..
leakage current

Gate MOS capacitor is biased into depletion

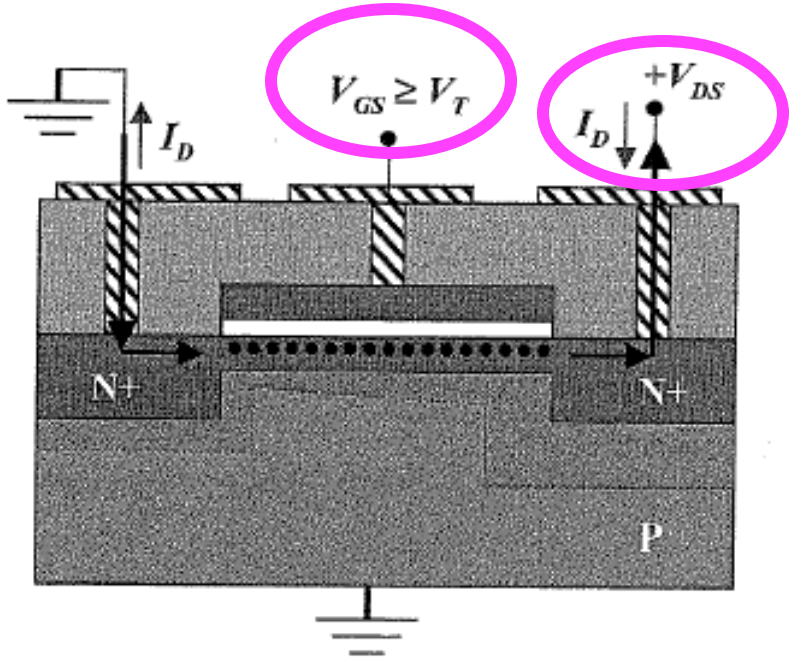


Barrier height at the source junction is V_{bi}-ψ_s near the source

from previous band diagram, for reference

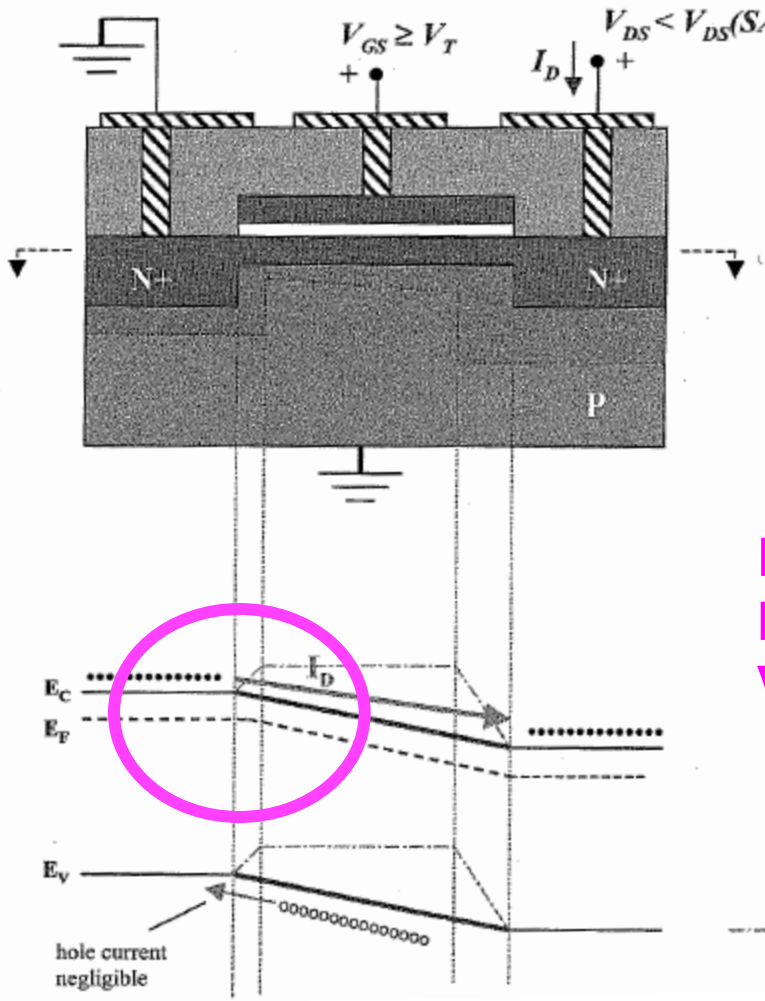
The Basics - Turning the MOSFET On & Off (cont.)

At or Beyond Threshold MOSFET is "ON"



Note: The depth of the channel is exaggerated for illustration purposes in this diagram. It is actually more like a "sheet" of charge right under the gate oxide.

MOSFET - Linear Regime



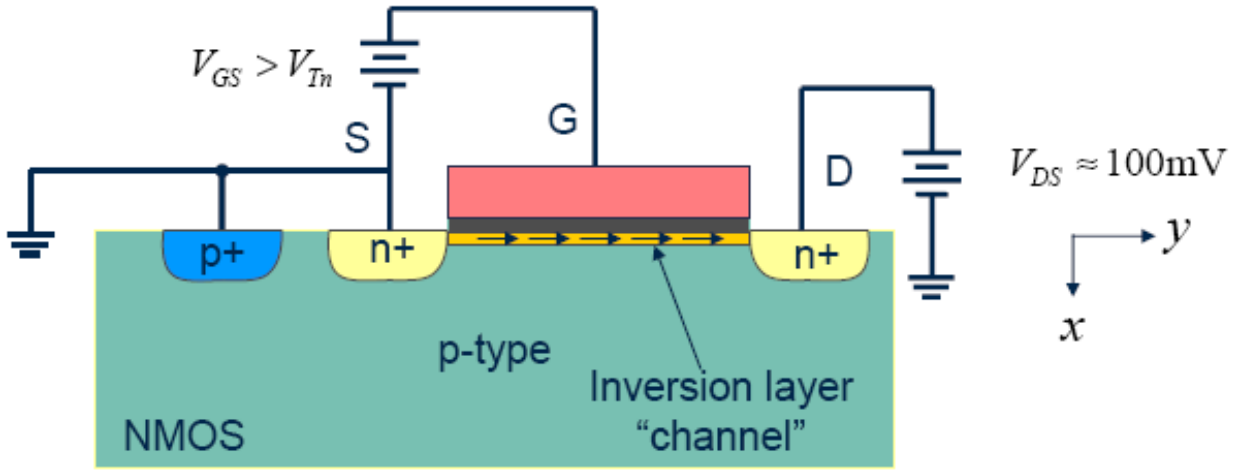
N-type channel, behaves like a resistor (ohmic)

Note that the barrier disappeared V_{bi}-ψ_s =0

from previous band diagram, for reference

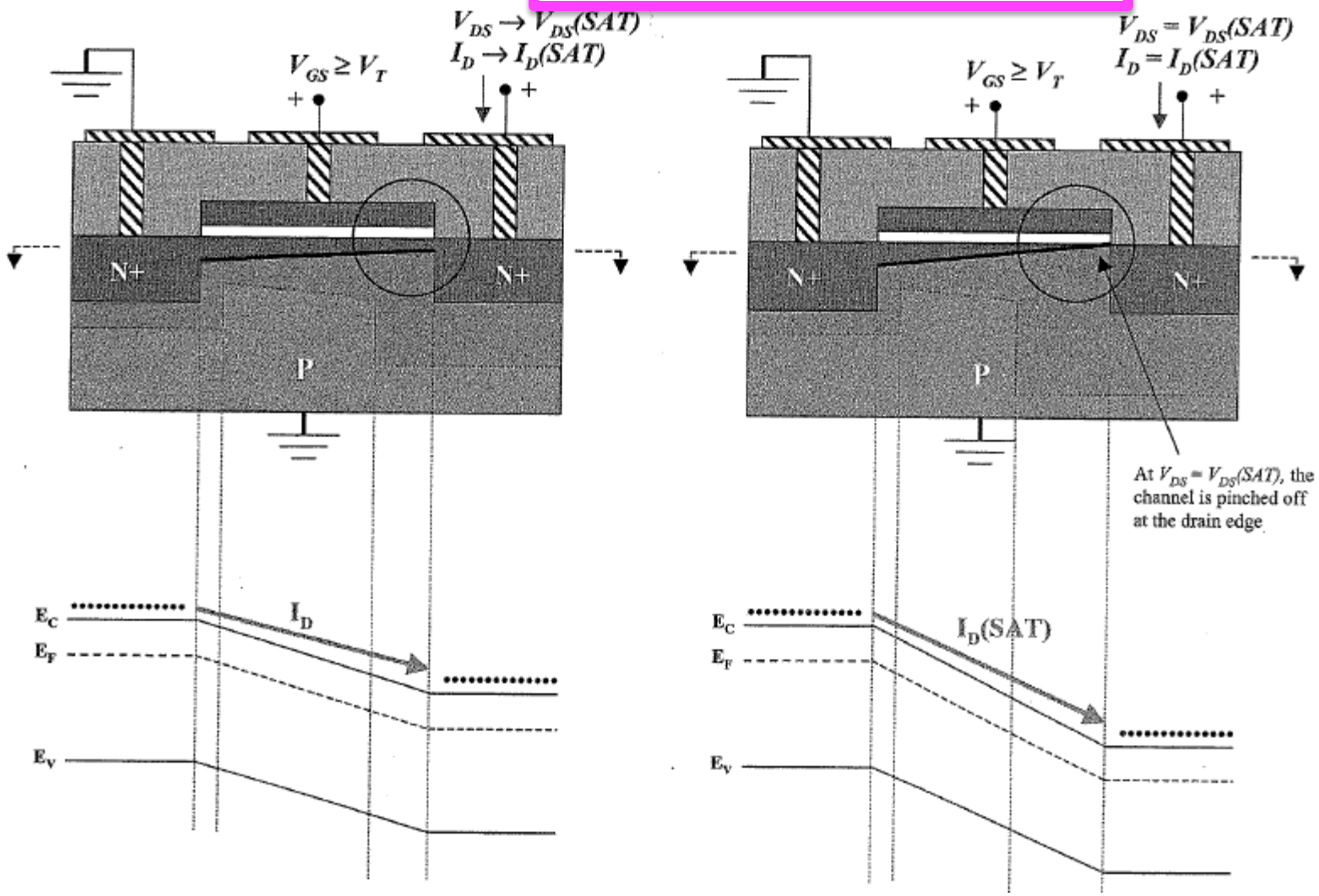
Common biasing of the MOSFET in linear regime

"Linear" Region Current



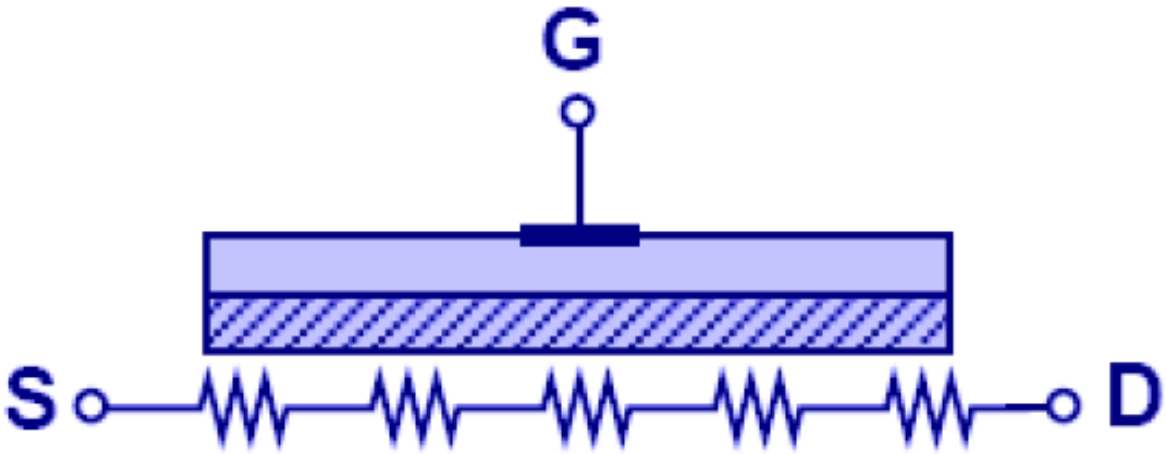
- If the gate is biased above threshold, the surface is inverted
- This inverted region forms a channel that connects the drain and gate
- If a drain voltage is applied positive, electrons will flow from source to drain

MOSFET - Linear Into Saturation



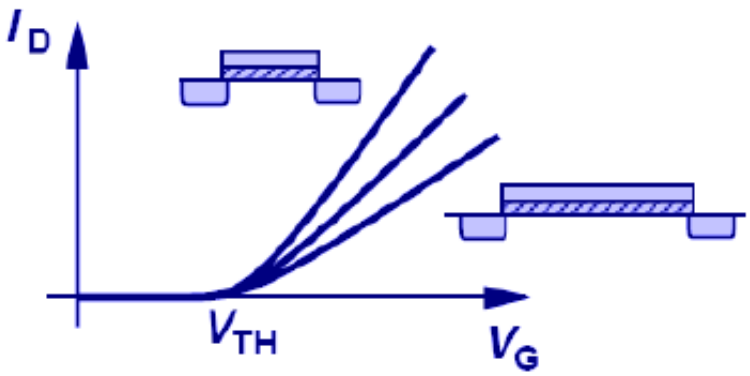
In the linear regime channel behaves like a gate-controlled resistor

Voltage-Dependent Resistor

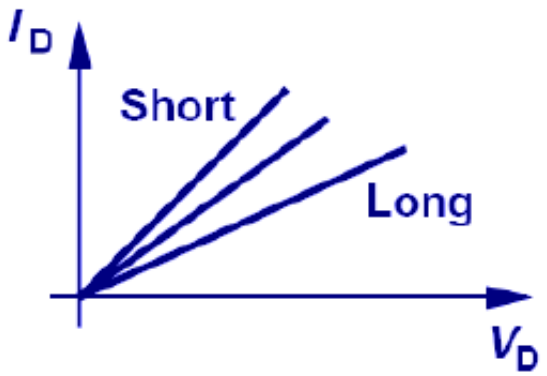


- The inversion channel of a MOSFET can be seen as a resistor.
- Since the charge density inside the channel depends on the gate voltage, this resistance is also voltage-dependent.

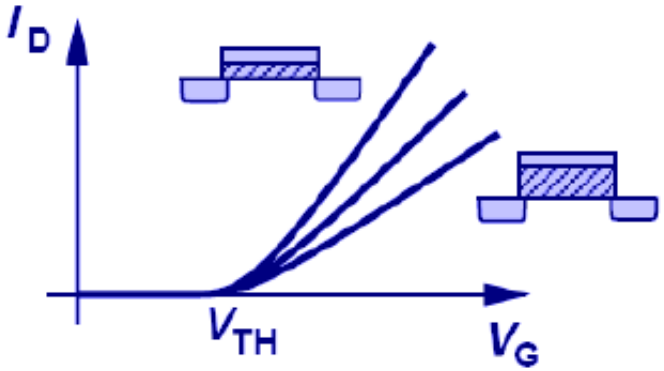
L and t_{ox} Dependence



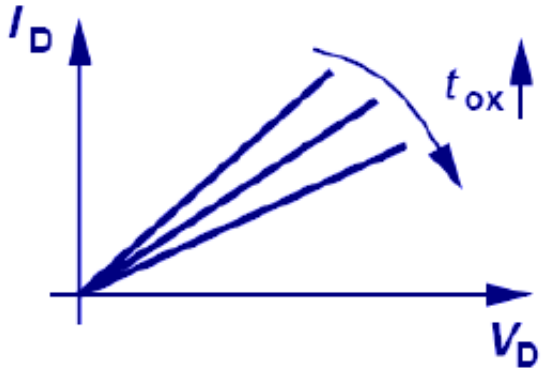
(a)



(b)



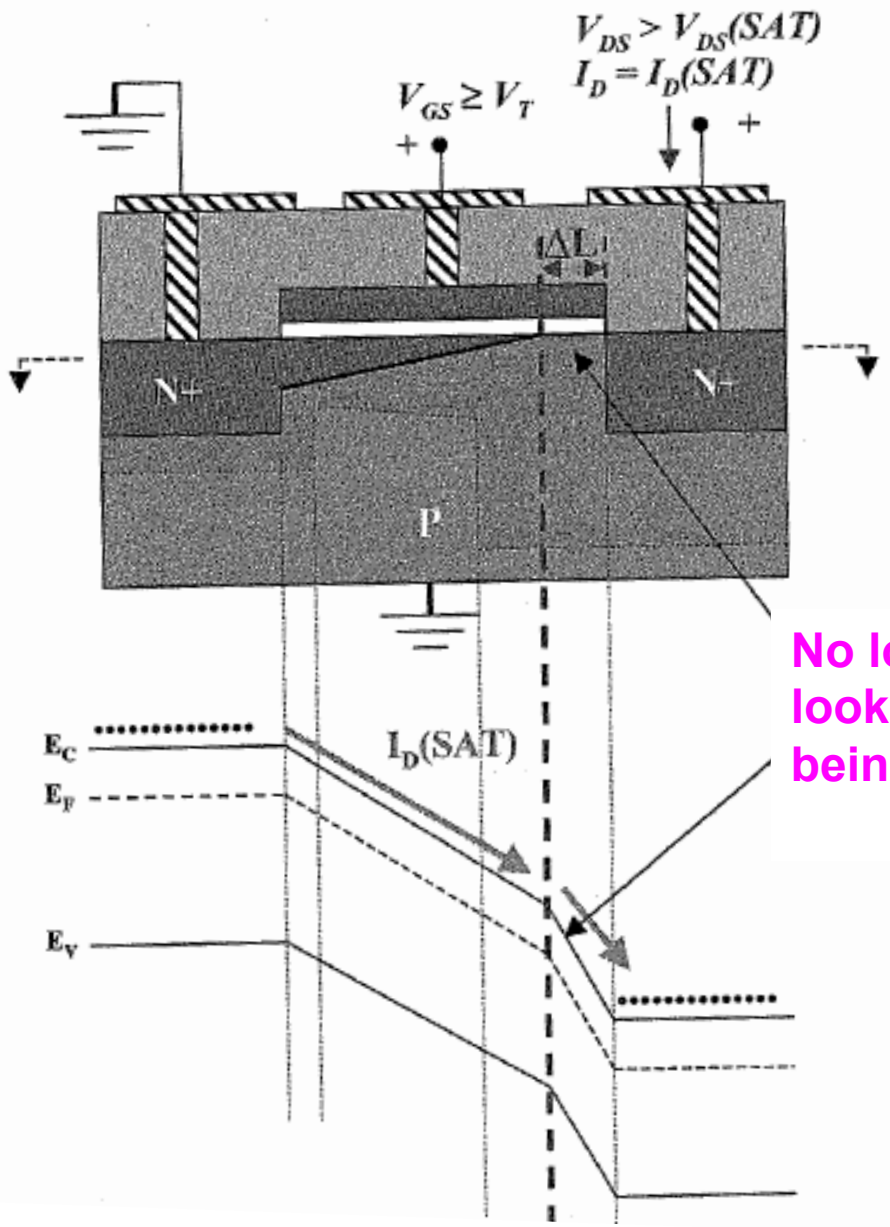
(c)



(d)

- **Small gate length and oxide thickness yield low channel resistance, which will increase the drain current.**

MOSFET - Saturation

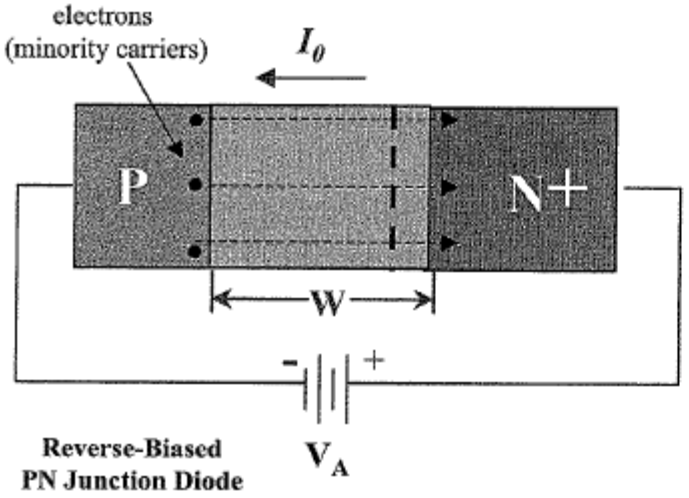


$$V_{DS}(SAT) \approx V_{GS} - V_T$$

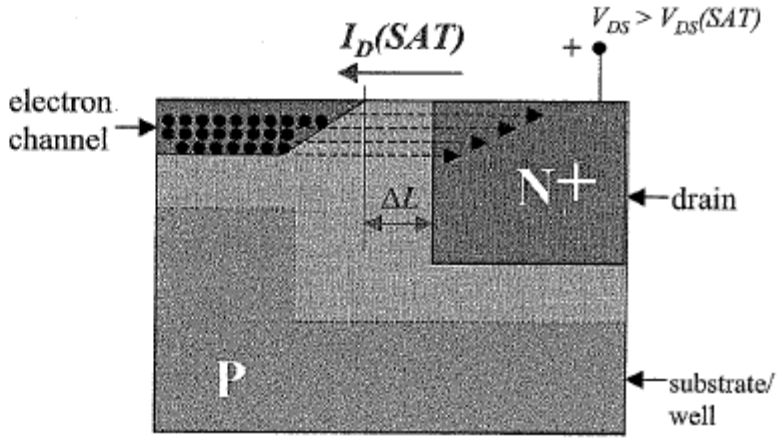
No longer behaving like a resistor; looks now like a depletion region being fed by the channel.

Analogy with pn diode in reverse bias condition

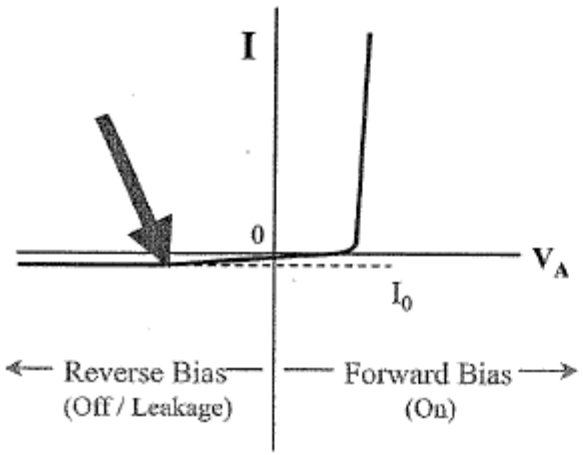
MOSFET - Saturation (cont.)

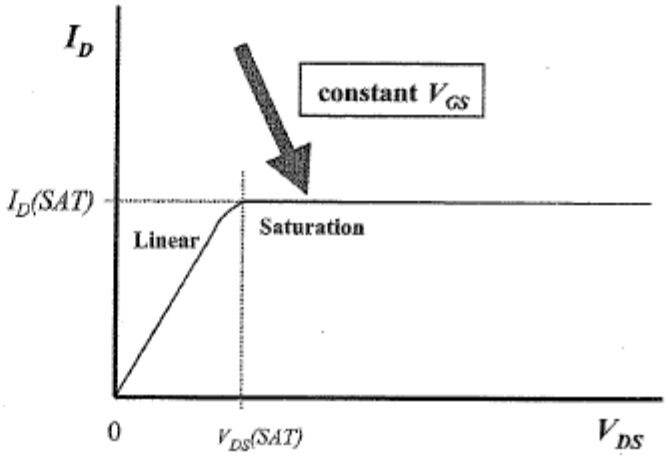


Reverse-Biased PN Junction Diode

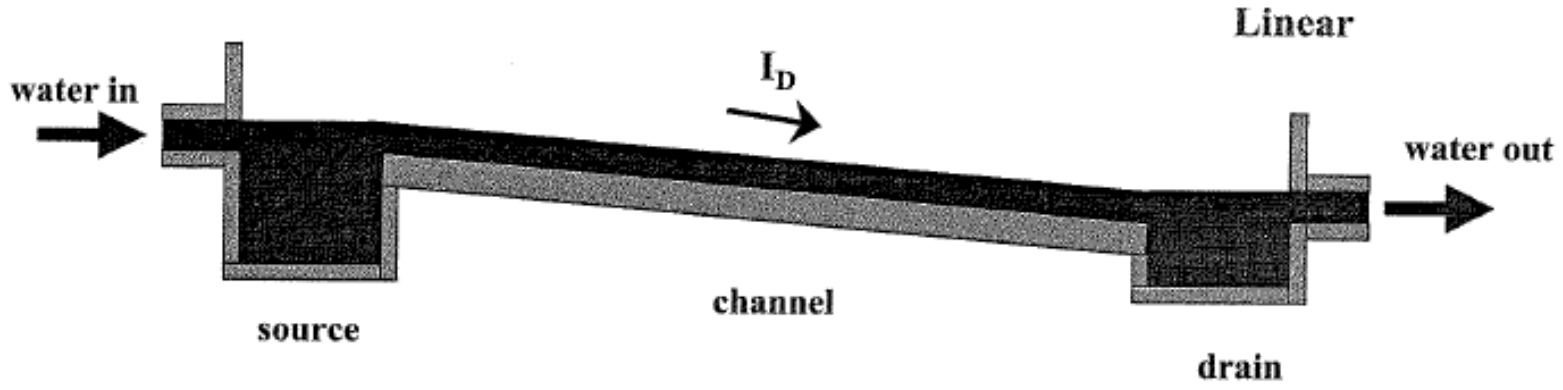


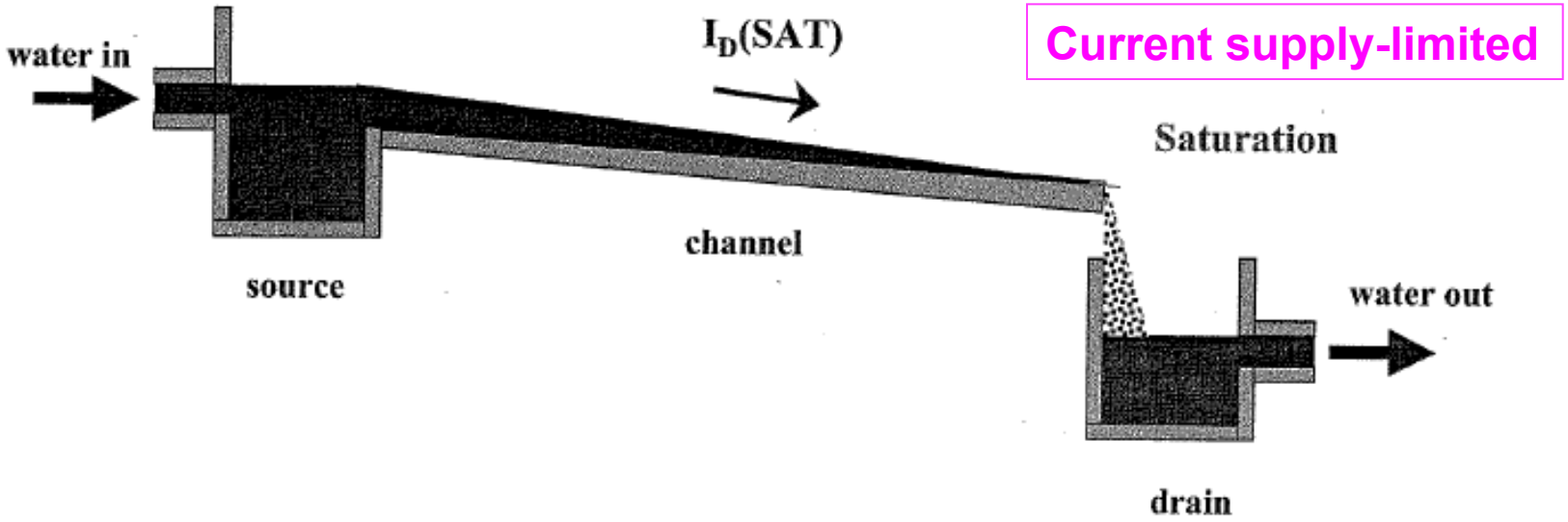
MOSFET in Saturation



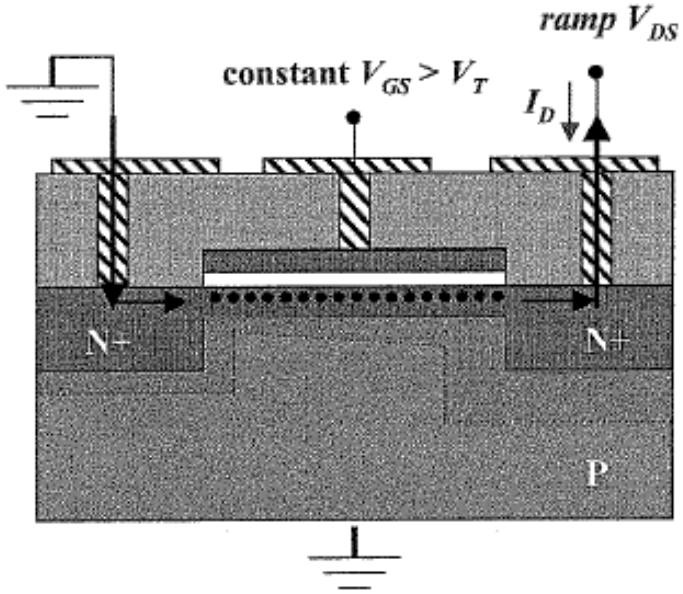


MOSFET - Saturation Analogy

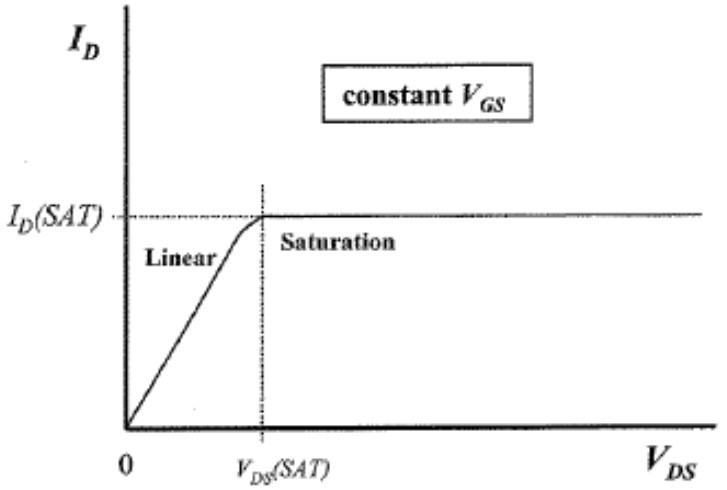




MOSFET - I-V Characteristic



For a constant gate voltage (above threshold), ramp V_{DS} and measure I_D.

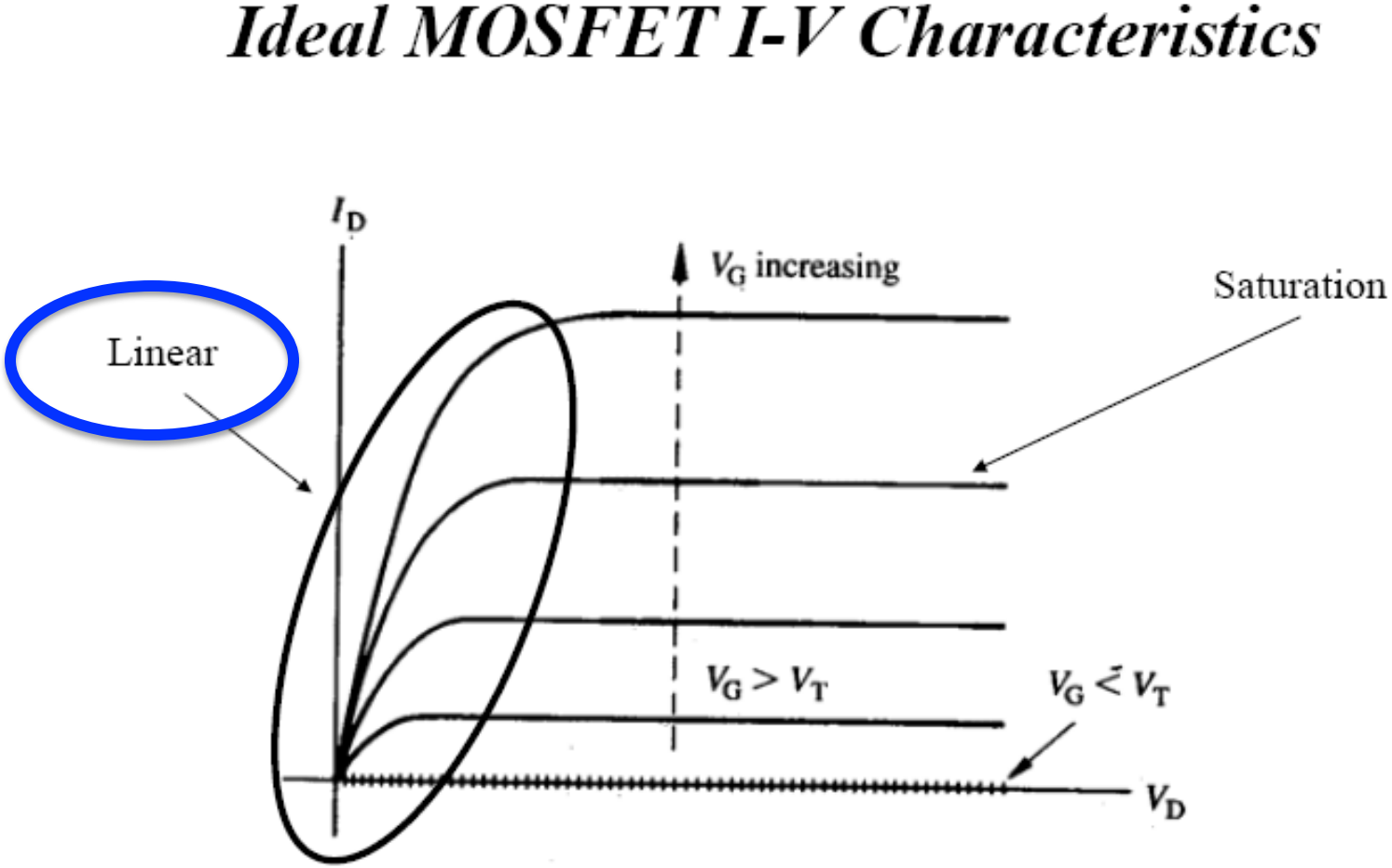


When

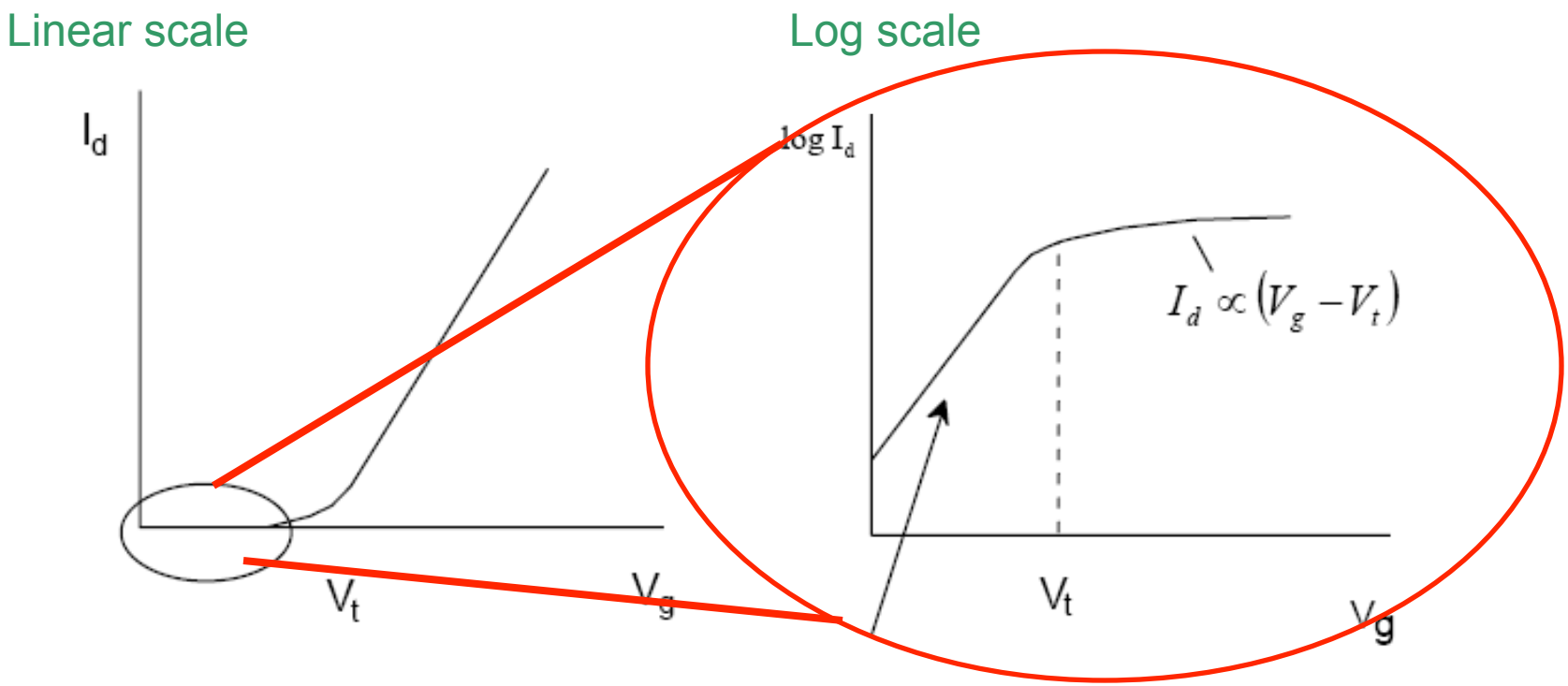
$V_{GS}-V_T = V_{dsat} < V_D$

→ saturation

Linear and Saturation Regime



Subthreshold Current Regime



subthreshold weak inversion

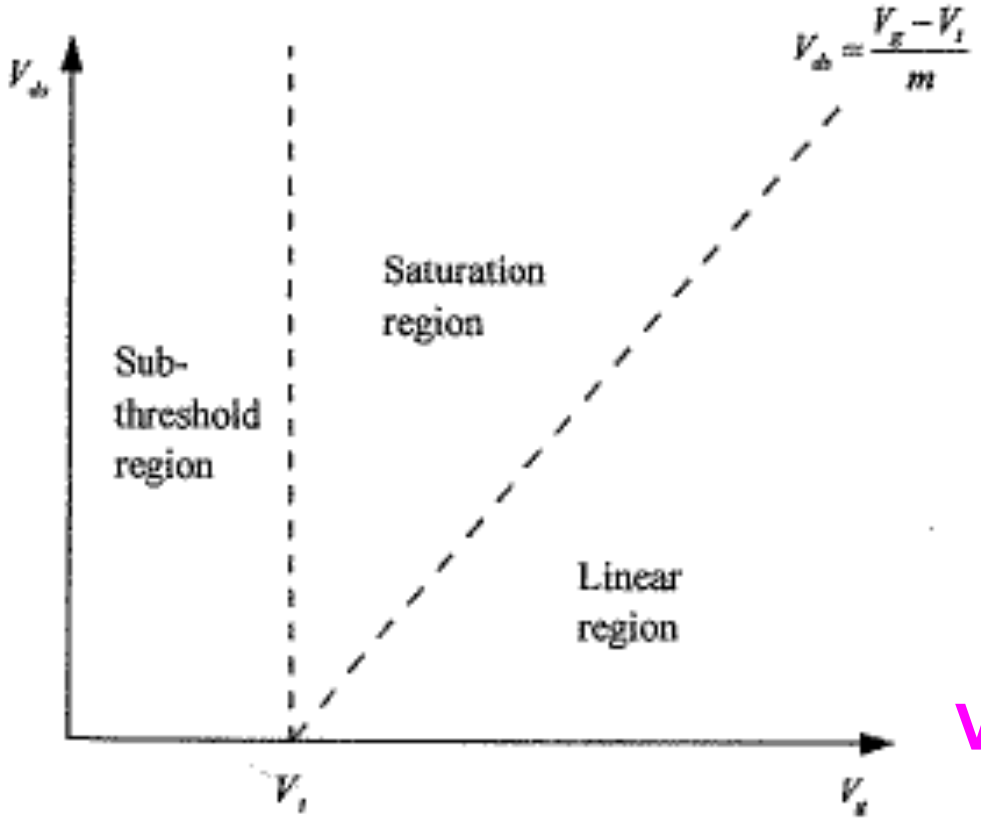
$$Q_{inv} \propto e^{qV_g / nkT}$$

$$I_d \propto e^{qV_g / nkT}$$

MOSFET Current Regimes

$$m = 1 + \frac{\sqrt{\varepsilon_s q N_A / 2\psi_B}}{C_{ox}} = 1 + \frac{C_d}{C_{ox}} = 1 + 3\frac{t_{ox}}{W_{\max}}$$

$$\varepsilon_s / \varepsilon_{ox} \approx 3$$



V drain

V gate

FIGURE 3.9. Three regions of MOSFET operation in the V_{ds}–V_g plane.

Analytical Formulae for MOSFET I-V Characteristics

Recall MOS Capacitor

$$p(x) = N_A \exp(-\beta\psi)$$

$$n(x) = \frac{n_i^2}{N_A} \exp(\beta\psi)$$

Surface potential ψ_s, surface electrical field E_s, and total charge in silicon Q_s

$$E_s = \{(\frac{2kT}{q})^2 (\frac{qN_A\beta}{2\varepsilon_s})[(\exp(-\beta\psi_s) + \beta\psi_s - 1) + \frac{n_i^2}{N_A}(\exp(\beta\psi_s) - \beta\psi_s - 1)]\}^{1/2}$$

$$Q_s = -\varepsilon_s E_s$$

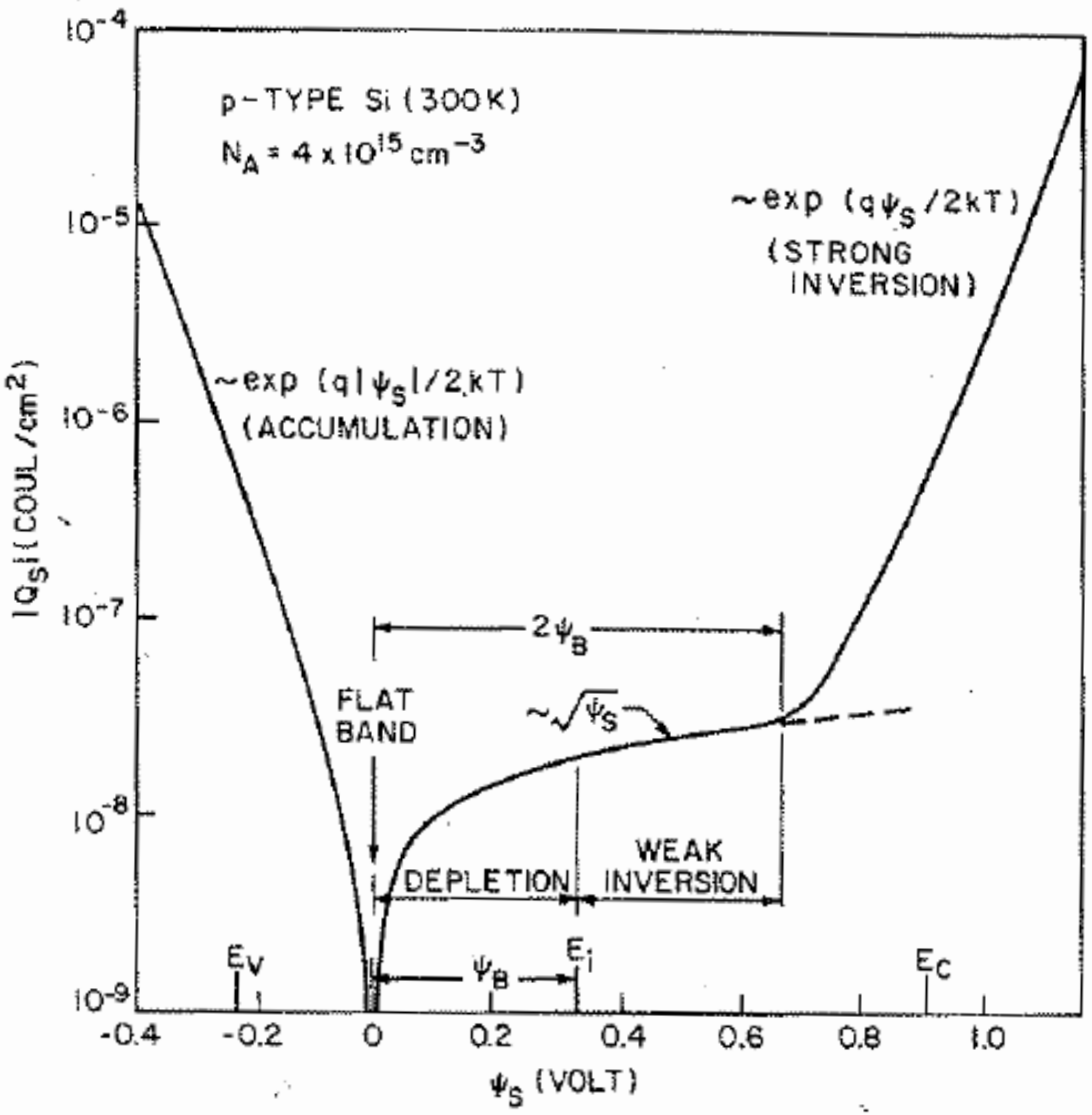
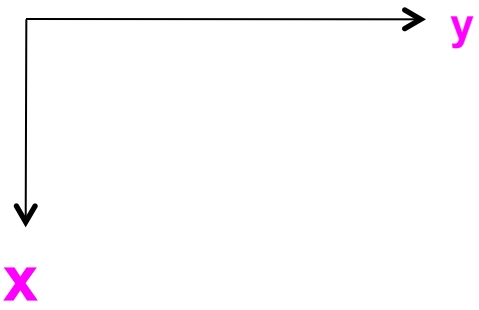


Fig. 5 Variation of space-charge density in the semiconductor as a function of the surface potential ψ_s for a *p*-type silicon with $N_A = 4 \times 10^{15}$ cm^{-3} at room temperature; ψ_B is the potential difference between the Fermi level and the intrinsic level of the bulk semiconductor. (After Garrett and Brattain, Ref. 13.)

Inversion Concentration in a MOSFET

Electron concentration in the inversion channel along (y) the gate



$$n(x,y) = \frac{n_i^2}{N_A}\exp(\beta[\psi - V(y)])$$

ψ inversion potential (band bending) with ψ_s at the surface

$$E_s^2 =$$

$$= (\frac{2kTN_A}{\varepsilon_s})^2[(\exp(-\beta\psi_s) + \beta\psi_s - 1) + \frac{n_i^2}{N_A}(\exp(-\beta V)\exp(\beta\psi_s) - \beta\psi_s - 1)]$$

$$\psi(0,y) = V(y) + 2\psi_B$$

$$W_{\max} = \sqrt{\frac{2\varepsilon_s[V(y) + 2\psi_B]}{qN_A}}$$

Gradual Channel Approximation

1-D Model of the MOSFET:

Key assumption in the *gradual channel approximation* (GCA) is to assume that the variation of the electric field in the y-direction (along the channel) is much less than the corresponding variation in the x-direction (perpendicular to the channel, i.e. gate field).

This allows to reduce Poisson's equation to the 1-D form (x component only).

GCA is valid for most of the channel regions except beyond the pinch-off point.

Gradual Channel Approximation

Current density at any point (x,y)

$$J_n(x,y) = -q\mu_n n(x,y)\frac{dV(y)}{dy}$$

To get the current we multiply the density by W (width of the transistor) and integrate over the inversion channel thickness in x direction.

$$I_{ds}(y) = qW\int_0^{x_i} \mu_n n(x,y)\frac{dV(y)}{dy}dx$$

The depth of the inversion channel x_i is defined as the bottom of the inversion layer where $\psi=\psi_B$

Gradual Channel Approximation

$$Q_n(y) = -q\int_0^{x_i} n(x,y)dx \qquad I_{ds}(y) = qW\int_0^{x_i} \mu_n n(x,y)\frac{dV(y)}{dy}dx$$

$$I_{ds}(y) = -\mu_n W\frac{dV(y)}{dy}Q_n(V)$$

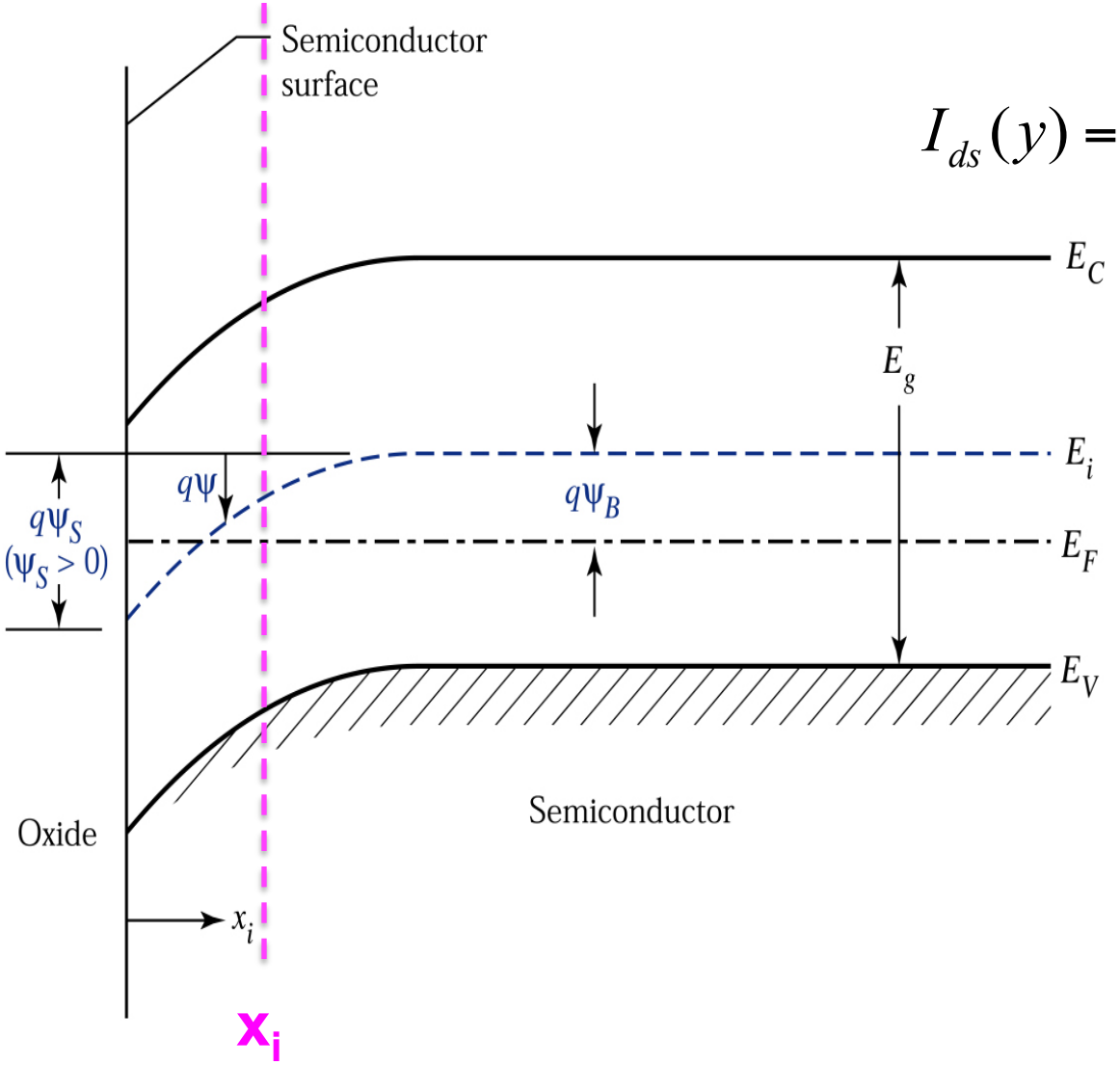
$$\int_o^L I_{ds}(y)dy = -\int_o^L \mu_n W\frac{dV(y)}{dy}Q_n(V)dy$$

Note that since V is function of y only, V is interchangeable with y.

We can multiply both sides by dy and integrate from 0 to L. Since current continuity requires I_{ds} to be constant, i.e. independent of y. Therefore we obtain:

$$I_{ds} = \mu_{eff}\frac{W}{L}\int_0^{V_{ds}} [-Q_n(V)]dV$$

Gradual Channel Approximation channel thickness



$$I_{ds}(y) = qW\int_0^{x_i} \mu_n n(x,y)\frac{dV(y)}{dy}dx$$

0 corresponds to ψ_s and x_i corresponds to ψ_b
recall $\psi_s=2\psi_b$ for the onset of inversion

$$\int_0^{x_i} \rightarrow \int_{2\psi_s}^{\psi_B}$$

PAO and SAH's Double Integral

$$n(x,y) = n(\psi, V) = \frac{n_i^2}{N_A} \exp(\beta[\psi - V\ ])$$

$$Q_n(V) = -q \int_{\psi_s}^{\psi_B} n(\psi, V) \frac{dx}{d\psi} d\psi$$

$$= -q \int_{\psi_s}^{\psi_B} \frac{n_i^2 / N_A \exp[\beta(\psi - V)]}{E(\psi, V)} d\psi$$

Pao and Sah's Double Integral

$$I_{ds} = q\mu_{eff}\frac{W}{L}\int_0^{V_{ds}}\left(\int_{\psi_s}^{\psi_B}\frac{n_i^2 / N_A \exp[\beta(\psi - V)]}{E(\psi, V)}d\psi\right)dV$$

Now we need to relate ψ_s to the gate voltage:

$$V_g = V_{FB} + \psi_s - \frac{Q_s}{C_{ox}} =$$

$$= V_{FB} + \psi_s + \frac{\sqrt{2\varepsilon_s kTN_A}}{C_{ox}}\left[\frac{q\psi_s}{kT} + \frac{n_i^2}{N_A}\exp[\beta(\psi_s - V)]\right]^{1/2}$$

Which is an implicit equation for $\psi_s(V)$.
Both equations have to be solved numerically.

Pao and Sah's Double Integral

$$V_g = V_{FB} + \psi_s + \frac{\sqrt{2\varepsilon_s kTN_A}}{C_{ox}}\left[\frac{q\psi_s}{kT} + \frac{n_i^2}{N_A}\exp[\beta(\psi_s - V)]\right]^{1/2}$$

The last equation has been obtained from $Q_s = -\varepsilon_s E_s(\psi_s)$

and

And

$$V_g = V_{FB} + \psi_s - \frac{Q_s}{C_{ox}}$$

$$E_s^2 =$$

$$= \left(\frac{2kTN_A}{\varepsilon_s}\right)[(\exp(-\beta\psi) + \beta\psi - 1) + \frac{n_i^2}{N_A}(\exp(-\beta V)\exp(\beta\psi) - \beta\psi - 1)]\}^{1/2}$$

where we kept only the two significant terms in strong inversion.

Charge Sheet Approximation

In order to derive an analytical formula we simplify the inversion charge using charge-sheet approximation, in which the inversion-layer thickness x_i is treated as zero. In other words, it assumes that all the inversion charges are located at the silicon interface like a sheet of charge and that there is no potential drop or band bending across the inversion layer. After the onset of inversion, the surface potential is pinned at $\psi_s=2\psi_B+V(y)$ as discussed earlier.

We proceed as follows:
1) we determine the depletion charge, 2) the total charge density in silicon, and 3) calculate the inversion charge density:

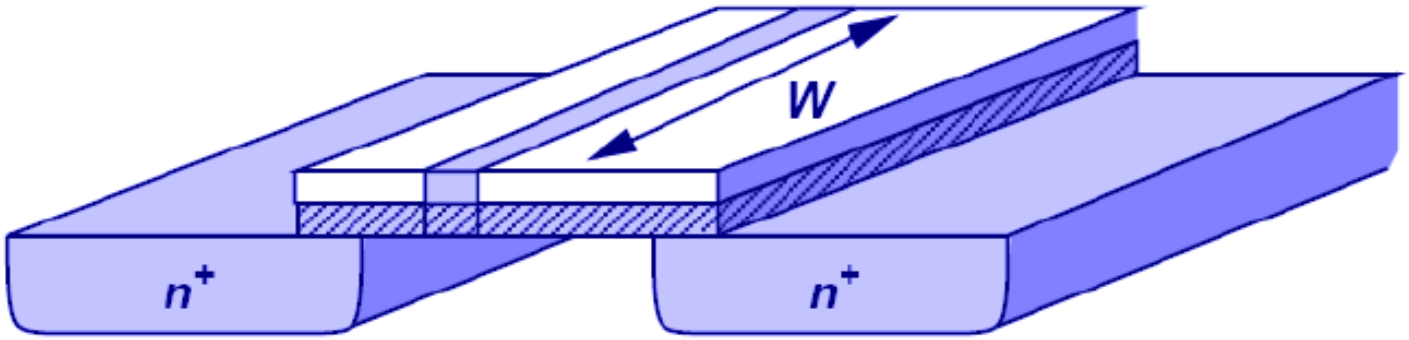
$$Q_d = -qN_A W_{\max} = -\sqrt{2\varepsilon_s qN_A(2\psi_B + V)}$$

$$Q_s = -C_{ox}(V_g - V_{FB} - \psi_s) =$$

$$= -C_{ox}(V_g - V_{FB} - 2\psi_B - V)$$

where V is the potential drop due to the drain bias

Channel Charge Density



$$Q = WC_{ox}(V_{GS} - V_{TH})$$

no reference to channel thickness

- The channel charge density is equal to the gate capacitance times the gate voltage in excess of the threshold voltage.

Charge Sheet Approximation

Therefore the inversion charge is

$$Q_n = Q_s - Q_d =$$

$$= -C_{ox}(V_g - V_{FB} - 2\psi_B - V) + \sqrt{2\varepsilon_s q N_A (2\psi_B + V)}$$

Substituting Q_n into current equation

$$I_{ds} = \mu_{eff} W \int_0^{V_{ds}} [-Q_n(V)] dV$$

and carrying out the integration in variable V, one obtains

$$I_{ds} =$$

$$\mu_{eff} C_{ox} \frac{W}{L} \left[(V_g - V_{FB} - 2\psi_B - \frac{V_{ds}}{2}) \cdot V_{ds} - \frac{2\sqrt{2\varepsilon_s q N_A}}{3C_{ox}} [(2\psi_B + V_{ds})^{3/2} - (2\psi_B)^{3/2}] \right]$$

Charge Sheet Approximation

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L} \left[(V_g - V_{FB} - 2\psi_B - \frac{V_{ds}}{2}) \cdot V_{ds} - \frac{2\sqrt{2\varepsilon_s q N_A}}{3C_{ox}} [(2\psi_B + V_{ds})^{3/2} - (2\psi_B)^{3/2}] \right]$$

The equation indicates that for a given V_g, the drain current first increases linearly with the drain voltage V_{ds} (called also the linear or triode region), then gradually levels off to a saturated value (saturation region).

Charge Sheet Approximation Linear Region

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left[(V_g - V_{FB} - 2\psi_B - \frac{V_{ds}}{2})\cdot V_{ds} - \frac{2\sqrt{2\varepsilon_s qN_A}}{3C_{ox}}[(2\psi_B + V_{ds})^{3/2} - (2\psi_B)^{3/2}]\right]$$

When V_{ds} is small we can expand the last equation into a power series in V_{ds} and keep only the lowest-order (first-order terms in V_{ds})

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left[V_g - V_{FB} - 2\psi_B - \frac{\sqrt{4\psi_B \varepsilon_s qN_A}}{C_{ox}}\right]\cdot V_{ds} =$$

$$= \mu_{eff} C_{ox} \frac{W}{L}(V_g - V_T)V_{ds}$$

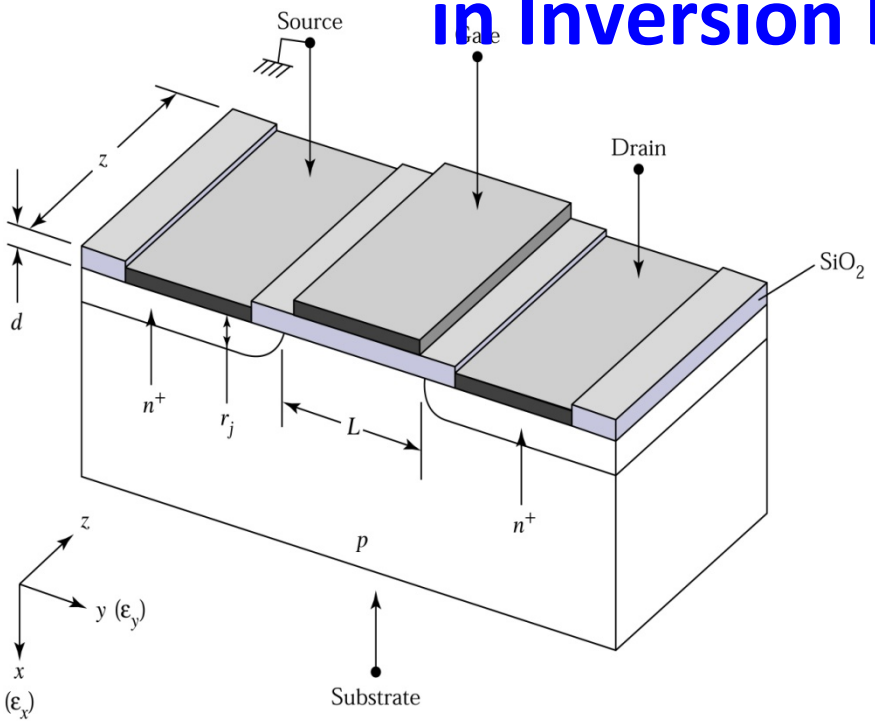
Where V_T is

$$V_T = V_{FB} - 2\psi_B - \frac{\sqrt{4\psi_B \varepsilon_s qN_A}}{C_{ox}}$$

(already familiar to us from our MOS capacitor derivation) and we have assumed That $V_{ds}/2 << V_{gs} - V_T$

Simple Estimate of MOSFET Current in Inversion Regime

$$I_D = \frac{Q}{t}$$



Mobile charge in the channel

$$Q = C_{ox}(V_G - V_T)$$

Estimate of the transit time

$$t = \frac{L}{v_d} = \frac{L}{E\mu} = \frac{L}{(V_{ds}/L)\mu} = \frac{L^2}{V_{ds}\mu}$$

$$Q_n = C_{ox}^{tot} V = \frac{\varepsilon_{ox} A}{t_{ox}}(V_{GS} - V_T) = \frac{\varepsilon_{ox} ZL}{t_{ox}}(V_{GS} - V_T) = ZLC_{ox}(V_{GS} - V_T)$$

$$I_d = \frac{Z\mu C_{ox}^{'}}{L}(V_{gs} - V_T)V_{ds}$$

Charge Sheet Approximation Linear Region

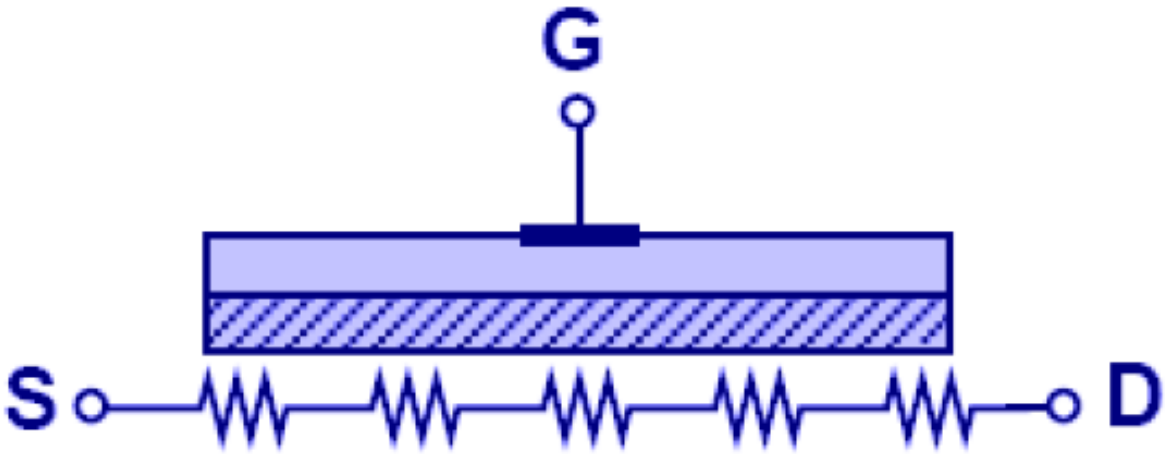
When V_g is below V_T there is very little current flowing and the MOSFET is in subthreshold regime (to be addressed next).

In the linear regime, the MOSFET simply acts like a resistor with a sheet resistivity

$$\rho_{sh} = \frac{1}{\mu_{eff} C_{ox} \left(V_g - V_T \right)}$$

modulated by the gate voltage.

Voltage-Dependent Resistor

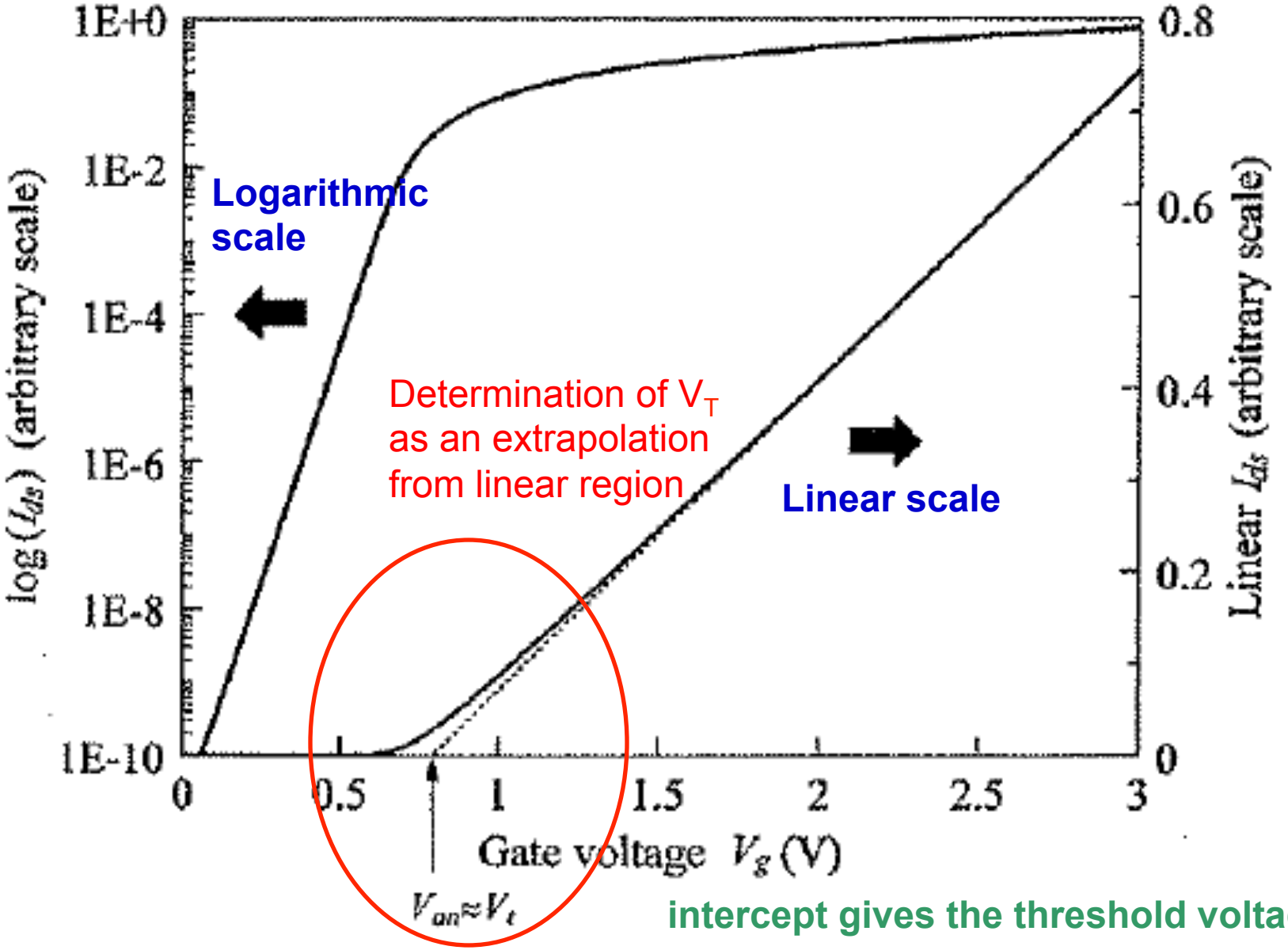


- The inversion channel of a MOSFET can be seen as a resistor.
- Since the charge density inside the channel depends on the gate voltage, this resistance is also voltage-dependent.

Determination of the Threshold Voltage from the Linear Behavior

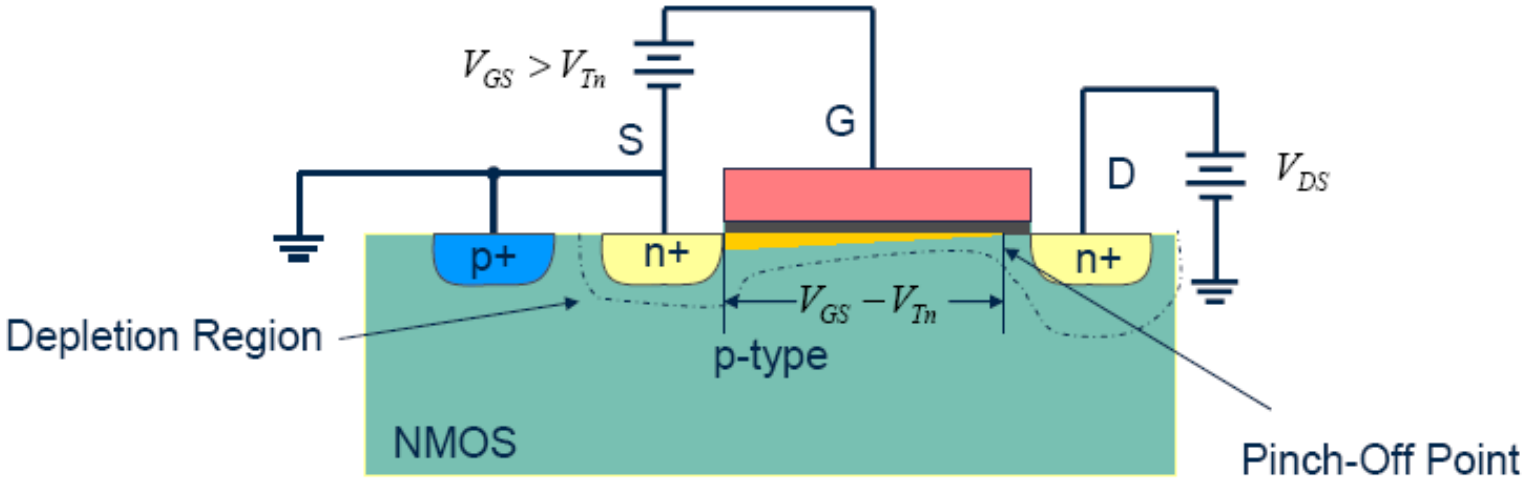
The threshold voltage V_T can be determined by plotting I_{ds} versus V_g at low drain voltages as shown in the figure (see next slide). The extrapolated intercept of the linear portion of $I_{ds}(V_g)$ curve with the V_g-axis gives the approximate value of V_T. In reality such a linearly extrapolated threshold voltage is slightly higher than "$2\psi_B$" due to inversion-layer capacitance and other effects to be discussed later. The $I_{ds}(V_g)$ curve is not linear near threshold voltage, because charge-sheet-approximation" is no longer valid in this regime.

Charge Sheet Approximation Linear Region



Saturation Regime and Pinch-Off

Pinching the MOS Transistors



- When $V_{DS} > V_{DS,sat}$, the channel is "pinched" off at drain end (hence the name "pinch-off region")
- Drain mobile charge goes to zero (region is depleted), the remaining electric field is dropped across this high-field depletion region
- As the drain voltage is increases further, the pinch off point moves back towards source
- Channel Length Modulation: The effective channel length is thus reduced → higher I_{DS}

CLM will be addressed later on

Saturation Regime

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left[(V_g - V_{FB} - 2\psi_B - \frac{V_{ds}}{2})\cdot V_{ds} - \frac{2\sqrt{2\varepsilon_s q N_A}}{3C_{ox}}[(2\psi_B + V_{ds})^{3/2} - (2\psi_B)^{3/2}]\right]$$

For larger values of V_{ds}, the second order terms in the power series expansion are not negligible and must be kept. A good approximation of the I_{ds} current can be obtained as:

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left((V_g - V_T)V_{ds} - \frac{m}{2}V_{ds}^2\right)$$

$$m = 1 + \frac{\sqrt{\varepsilon_s q N_A \psi_B}}{2C_{ox}} = 1 + \frac{C_d}{C_{ox}} = 1 + \frac{3t_{ox}}{W_{\max}} \qquad \frac{\varepsilon_s}{\varepsilon_{ox}} \approx 3$$

m is usually close to unity since the W_{max} is much larger than t_{ox}

Saturation Regime

The V_T can be now expressed in terms of m as

$$V_T = V_{FB} + (2m-1)2\psi_B$$

As V_{ds} increases, I_{ds} follows a parabolic curve, as shown in the figure on the next slide until a maximum (or saturation value is reached. This occurs when $V_{ds}=V_{dsat}=(Vg- V_T)/m$ at which

$$I_{ds} = I_{dsat} = \mu_{eff} C_{ox} \frac{W}{L} \frac{(V_g - V_T)^2}{2m}$$

One often finds a following formula for the saturation current in the textbooks ($m \approx 1$ $C_d << C_{ox}$)

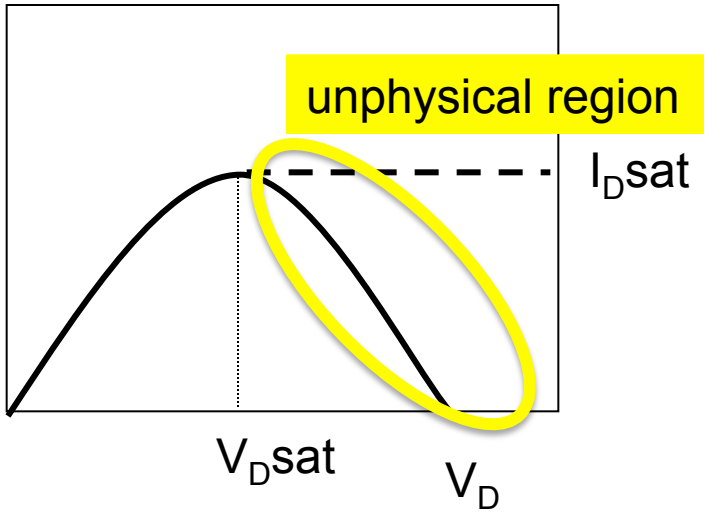
$$I_{ds} = I_{dsat} = \mu_{eff} C_{ox} \frac{W}{2L} (V_g - V_T)^2$$

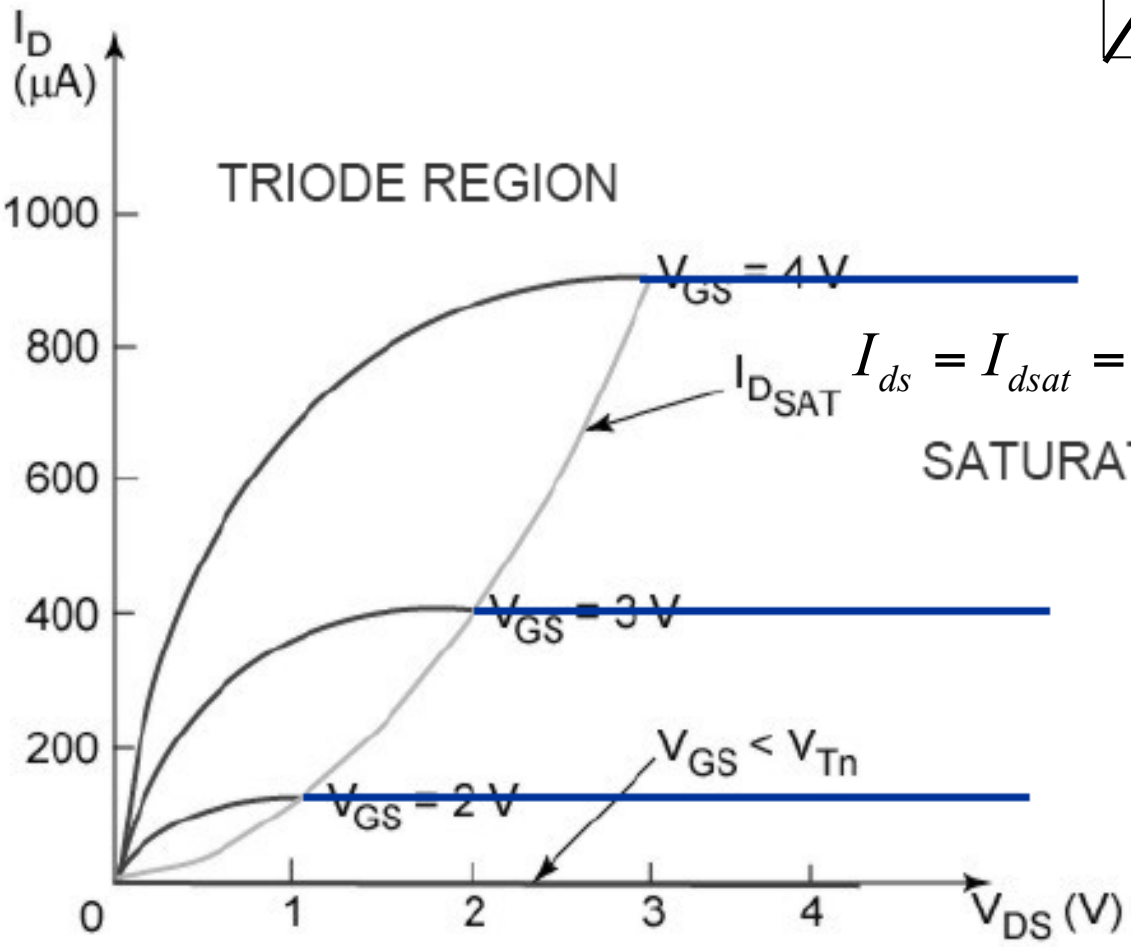
$$m = 1 + \frac{\sqrt{\varepsilon_s q N_A / 2\psi_B}}{C_{ox}} = 1 + \frac{C_d}{C_{ox}} = 1 + 3\frac{t_{ox}}{W_{\max}}$$

$$\varepsilon_s / \varepsilon_{ox} \approx 3$$

Saturation Regime

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left((V_g - V_T)V_{ds} - \frac{m}{2}V_{ds}^2\right)$$





$$I_{ds} = I_{dsat} = \mu_{eff} C_{ox} \frac{W}{2L}(V_g - V_T)^2$$

SATURATION REGION

Saturation Regime – explicit calculation

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left((V_g - V_T)V_{ds} - \frac{m}{2}V_{ds}^2\right)$$

$$\frac{\partial I_{ds}}{\partial V_{ds}} = 0 = \mu_{eff} C_{ox} \frac{W}{L}(V_g - V_T) - mV_{ds}$$

$$for \quad m \approx 1$$

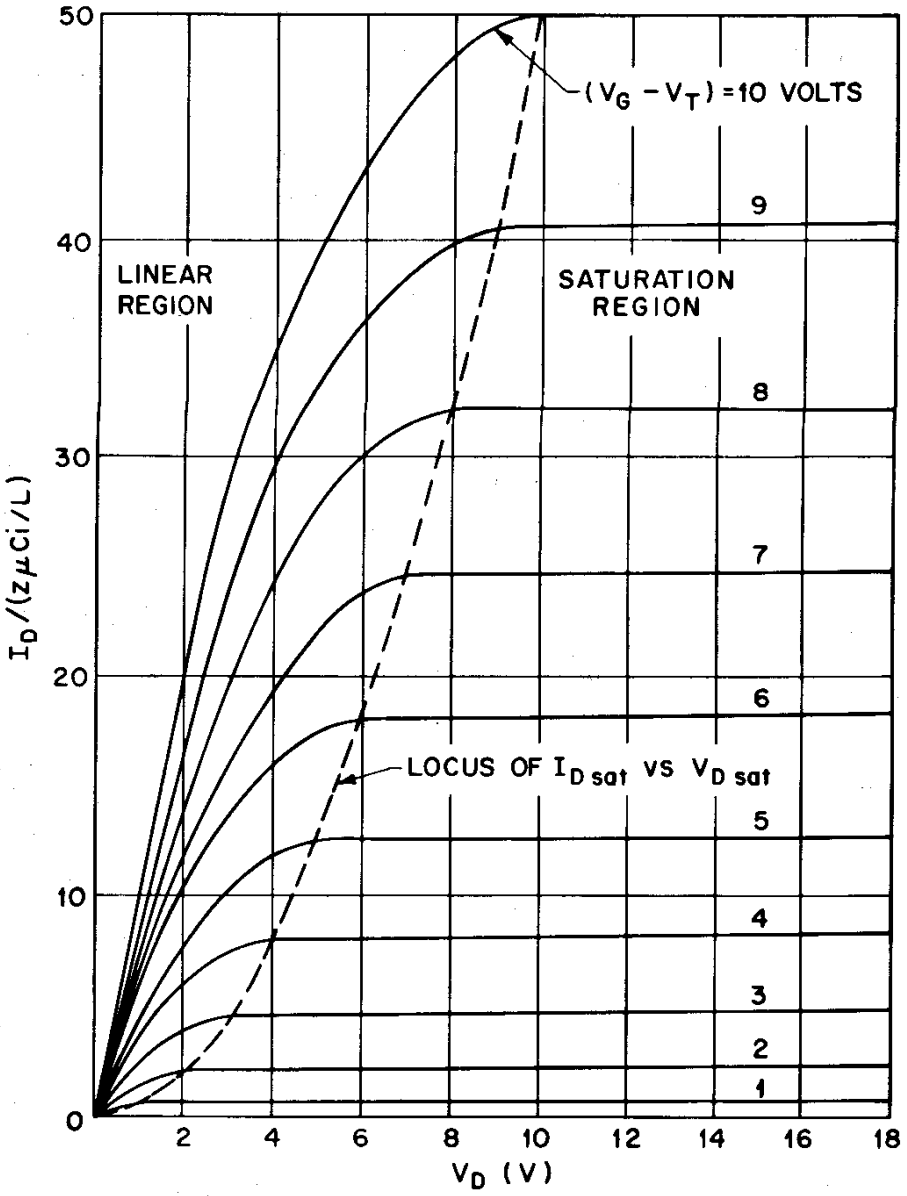
$$\mu_{eff} C_{ox} \frac{W}{L}(V_g - V_T) = V_{ds}$$

$$then$$

$$I_{ds}(V_{ds} = \mu_{eff} C_{ox} \frac{W}{L}(V_g - V_T)) = I_{dsat}$$

$$I_{dsat} = \mu_{eff} C_{ox} \frac{W}{2L}(V_g - V_T)^2$$

Saturation Regime



Ideal Characteristics of n-channel enhancement mode MOSFET

$$I_{ds} = I_{dsat} = \mu_{eff} C_{ox} \frac{W}{2L} (V_g - V_T)^2$$

$$V_{dsat} = (V_g - V_T)$$

$$I_{dsat} = \mu_{eff} C_{ox} \frac{W}{2L} V_{dsat}^2$$

Saturation Regime

$$I_{ds} = I_{dsat} = \mu_{eff} C_{ox} \frac{W}{2L} (V_g - V_T)^2$$

This is valid for low substrate doping where $W_{max} >> t_{ox}$ and then $m \approx 1$.

So beyond V_{dsat}, I_{ds} stays constant at I_{dsat} independent of V_{ds}.

ONSET of Pinch-Off

$$Q_n(V) = Q_s - Q_d =$$

$$= -C_{ox}(V_g - V_{FB} - 2\psi_B - V) + \sqrt{2\varepsilon_s q N_A (2\psi_B + V)}$$

For $V<2\psi_B$ one can expand the square root term into power series in V and keep only the two lowest terms:

$$Q_n = Q_s - Q_d = -C_{ox}(V_g - V_T - mV)$$

Drain current is proportional to the area under the $-Q_n$ curve between V=0 and $V=V_{ds}$. When V_{ds} small, Q_n is only slightly lower than that at the source end. But as the drain voltage increases Q_n decreases and goes to zero at V_{dsat}. Inversion vanishes at the drain – channel saturation. This condition is called pinch-off.

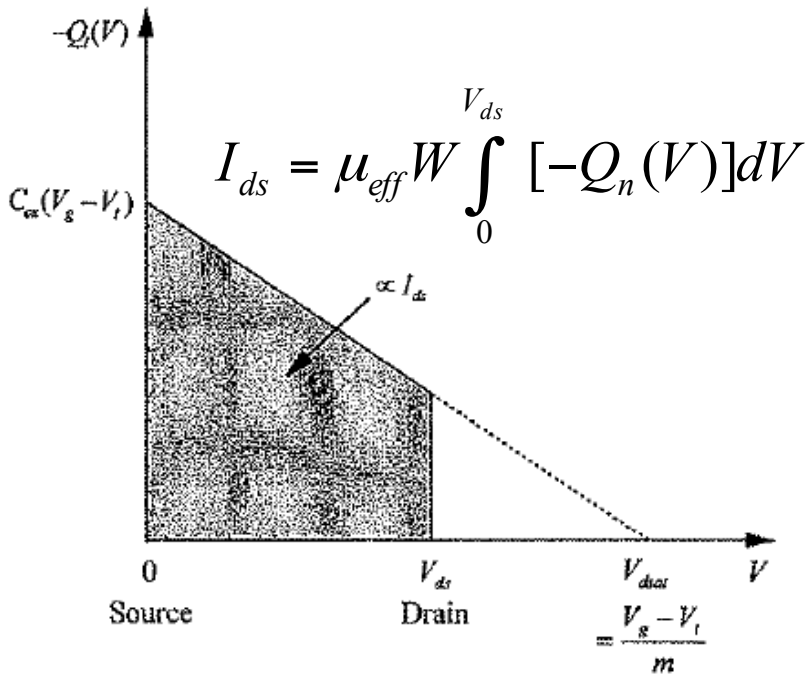


FIGURE 3.5. Inversion charge density as a function of the quasi-Fermi potential of a point in the channel. Before saturation, the drain current is proportional to the shaded area integrated from zero to the drain voltage.

ONSET of Pinch-Off

When V_{ds} increases even further (beyond saturation), the pinch-off point moves toward the source, but the drain current remains essentially the same.

$$\int_0^{L'} I_{ds}\,dy = \mu_{eff} W \int_0^{V_{ds}} [-Q_n(V)]\,dV$$

But note that we have now L' <L. This phenomenon is called **channel length modulation** and will be discussed later.

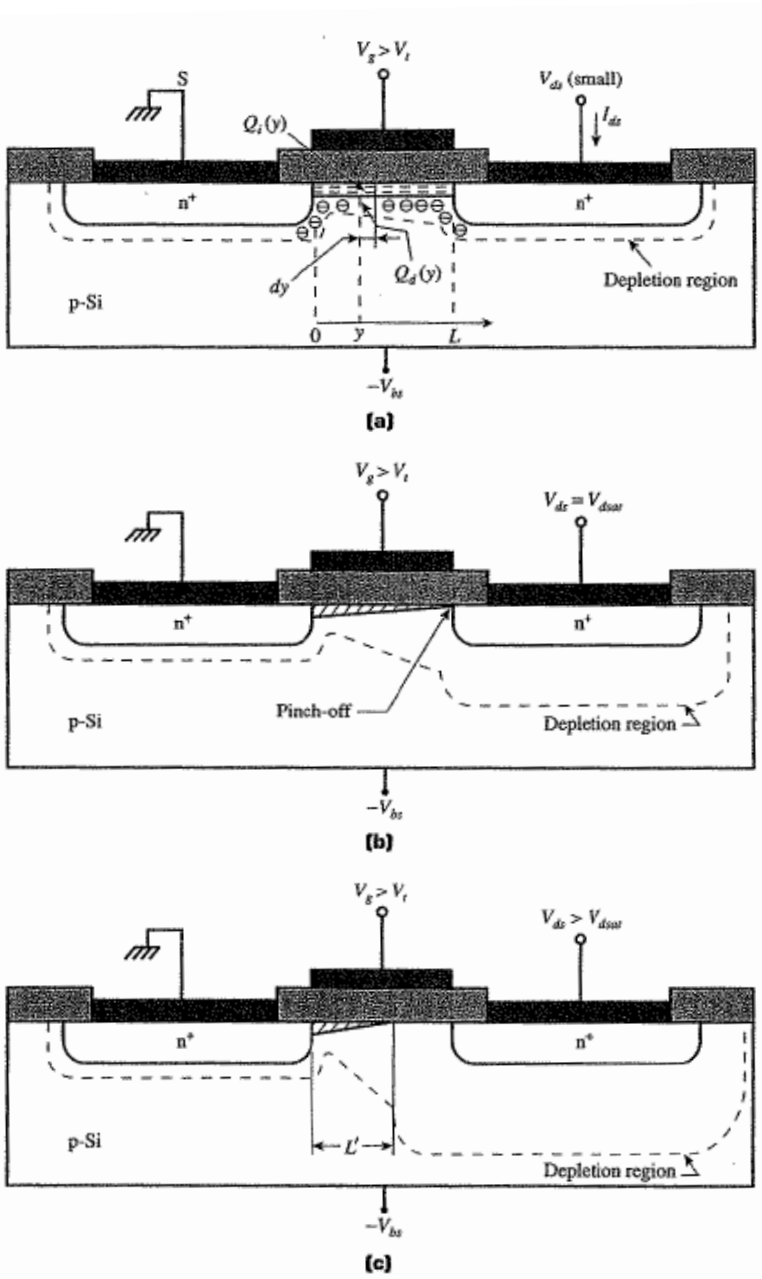


FIGURE 3.6. (a) MOSFET operated in the linear region (low drain voltage). (b) MOSFET operated at the onset of saturation. The pinch-off point is indicated by *Y*. (c) MOSFET operated beyond saturation where the channel length is reduced to *L'*. (After Sze, 1981.)

ONSET of Pinch-Off

We wish to learn explicitly about V(y) behavior as a function of y.
So far we know V(y=0) =0 and V(y=L)=V_{ds}.

Using $$Q_n = Q_s - Q_d = -C_{ox}(V_g - V_T - mV)$$

We can integrate $$\int_0^y I_{ds}\,dy = \mu_{eff} W \int_0^{V(y)} [-Q_n(V)]\,dV$$

And obtain $$I_{ds} \cdot y = \mu_{eff} C_{ox} W (V_g - V_T)V(y) - \frac{m}{2}V(y)^2)$$

Substituting $$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left((V_g - V_T)V_{ds} - \frac{m}{2}V_{ds}^2\right)$$

One can solve the resulting equation for V(y) and obtain

$$V(y) = \frac{V_g - V_T}{m} - \sqrt{\left(\frac{V_g - V_T}{m}\right)^2 - 2\frac{y}{L}\left(\frac{V_g - V_T}{m}\right)V_{ds} + \frac{y}{L}V_{ds}^2}$$

ONSET of Pinch-Off

$$V(y) = \frac{V_g - V_T}{m} - \sqrt{\left(\frac{V_g - V_T}{m}\right)^2 - 2\frac{y}{L}\left(\frac{V_g - V_T}{m}\right)V_{ds} + \frac{y}{L}V_{ds}^2}$$

$$m = 1 + \frac{\sqrt{\varepsilon_s q N_A / 2\psi_B}}{C_{ox}} = 1 + \frac{C_d}{C_{ox}} = 1 + 3\frac{t_{ox}}{W_{\max}}$$

$$\varepsilon_s / \varepsilon_{ox} \approx 3$$

This is a very satisfying result: V(y) depends on the external biases V_g and V_{ds}; it depends also on V_T and on transistor geometry: L, t_{ox}, and W_{max}.

We remind ourselves that the current I_{ds}= (const at each location y) is proportional to the product of $Q_n(y)dV(y)/dy$.

We now plot $Q_n(y)$ and V(y) in the same figure.

ONSET of Pinch-Off

$$V(y) = \frac{V_g - V_T}{m} - \sqrt{\left(\frac{V_g - V_T}{m}\right)^2 - 2\frac{y}{L}\left(\frac{V_g - V_T}{m}\right)V_{ds} + \frac{y}{L}V_{ds}^2}$$

Both V(y) and $-Q_n/(mC_{ox})$ are plotted below:

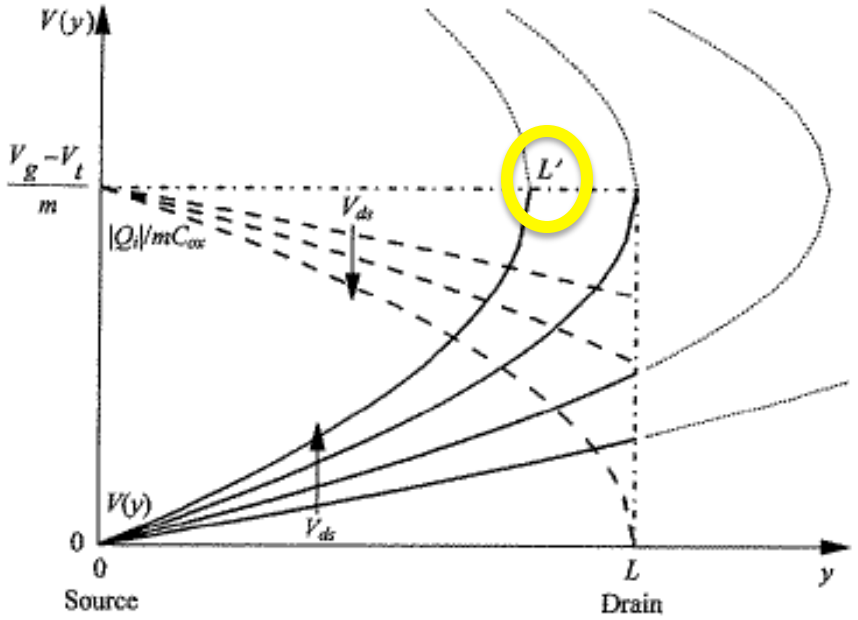


FIGURE 3.7. Quasi-Fermi potential versus distance between the source and the drain for several V_{ds}-values from the linear region to beyond saturation. The dashed curves show the corresponding variation of inversion charge density along the channel. The dotted curves help visualize the parabolic behavior of the characteristics.

At low V_{ds} V(y) varies linearly between source and drain. But as V_{ds} increases, The inversion charge at the drain decreases due to the lowering of the electron quasi Fermi level. This is accompanied by a corresponding increase of dV(y)/dy to maintain the current continuity. When V_{ds} reaches saturation we have: $V_{dsat}=(V_g-V_T)/m$ and $Q_n(y=L)=0$.

This implies that the electric field in the y-direction changes more rapidly than the field in x-direction and the gradual channel approximation breaks down.

ONSET of Pinch-Off

Comparison between

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left[(V_g - V_{FB} - 2\psi_B - \frac{V_{ds}}{2})\cdot V_{ds} - \frac{2\sqrt{2\varepsilon_s q N_A}}{3C_{ox}}[(2\psi_B + V_{ds})^{3/2} - (2\psi_B)^{3/2}]\right]$$

and

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left((V_g - V_T)V_{ds} - \frac{m}{2}V_{ds}^2\right)$$

Note: strictly speaking, when $V_{ds} > 2\psi_B$ we can't expand the square root in power series.
But we can differentiate the above equation $dI_{ds}/dV_{ds}=0$ and we obtain for V_{dsat}:

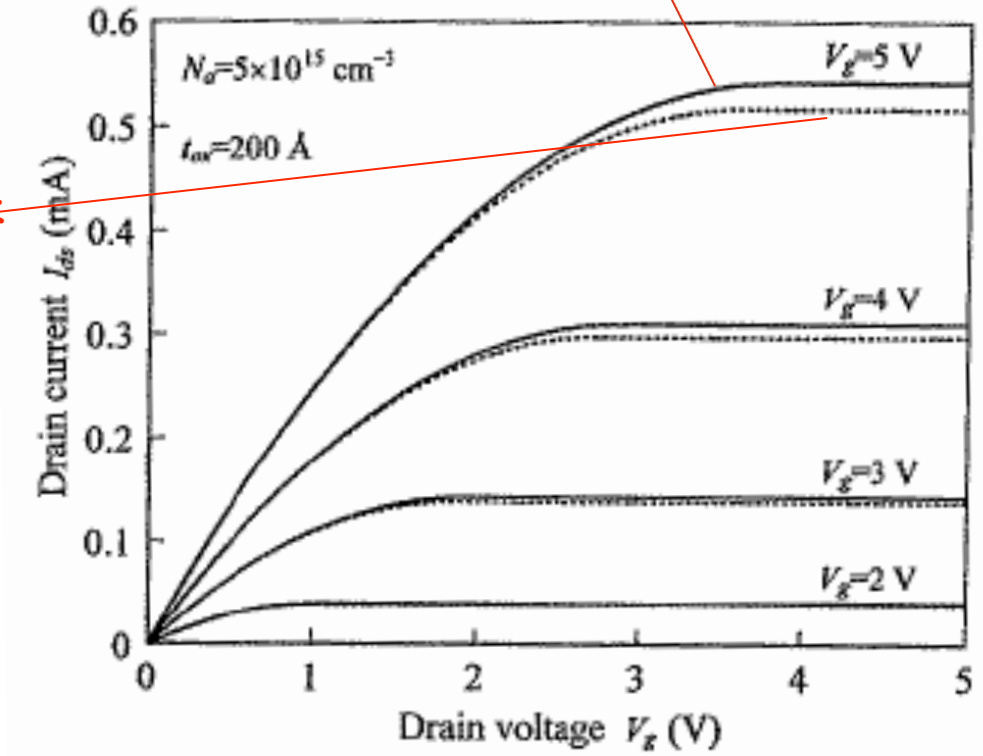


FIGURE 3.8. I_{ds}–V_{ds} curves calculated from the full equation (3.18) (solid curves), compared with the parabolic approximation (3.21) (dotted curves).

ONSET of Pinch-Off

Note: strictly speaking, when $V_{sd} > 2\psi_B$ we can't expand the square root in power series.
But we can differentiate the above equation $dI_{ds}/dV_{ds}=0$ and we obtain for V_{dsat}:

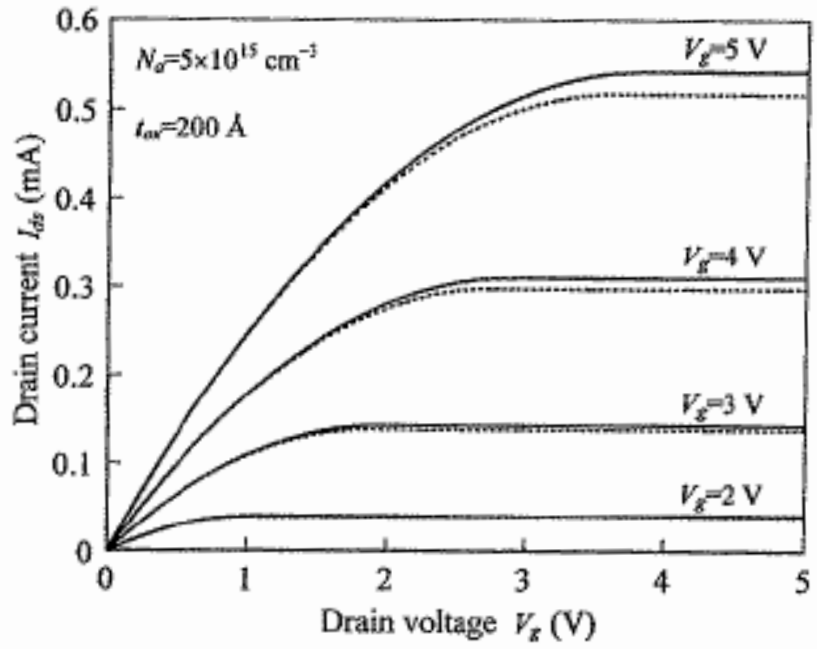


FIGURE 3.8. I_{ds}–V_{ds} curves calculated from the full equation (3.18) (solid curves), compared with the parabolic approximation (3.21) (dotted curves).

$$V_{dsat} = V_g - V_{FB} - 2\psi_B + \frac{\varepsilon_s q N_a}{C_{ox}^2} - \sqrt{\frac{\varepsilon_s q N_a}{C_{ox}^2}\left(V_g - V_{FB}\right) + \frac{\varepsilon_s q N_a}{C_{ox}^2}}$$

The corresponding I_{dsat} can be found by substituting the above equation in

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left[\left(V_g - V_{FB} - 2\psi_B - \frac{V_{ds}}{2}\right)\cdot V_{ds} - \frac{2\sqrt{2\varepsilon_s q N_A}}{3C_{ox}}\left[(2\psi_B + V_{ds})^{3/2} - (2\psi_B)^{3/2}\right]\right]$$

(not shown, usually numerical evaluation)

Simplified Analysis for Linear and Saturation MOSFET Regime

Inversion layer thickness

Inversion layer concentration

$$I_D = Wt \bullet (-q\ n\ v_{drift})$$
$$= W \bullet Q_n \bullet v_{drift}$$

Note: I_D is constant for all positions along channel

Let V_T **defined** to be threshold voltage at Source

$$V_T(\text{average}) \sim V_T + \frac{V_{DS}}{2}$$

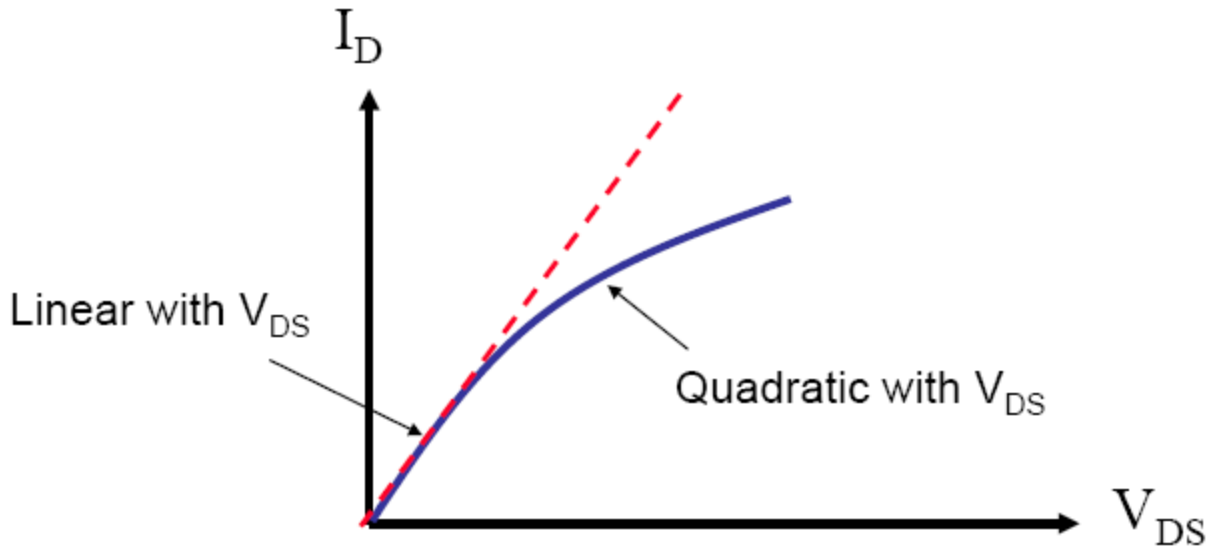
[This is an approximation]

$$Q_n(\text{average}) = C_{OX}\left(V_G - V_T(\text{average})\right)$$
$$= C_{OX}\left(V_G - V_T - \frac{V_{DS}}{2}\right)$$

Simplified Analysis for Linear and Saturation MOSFET Regime

$$\text{With } v_{drift} = -\mu_n E \approx \frac{\mu_n V_{DS}}{L}$$

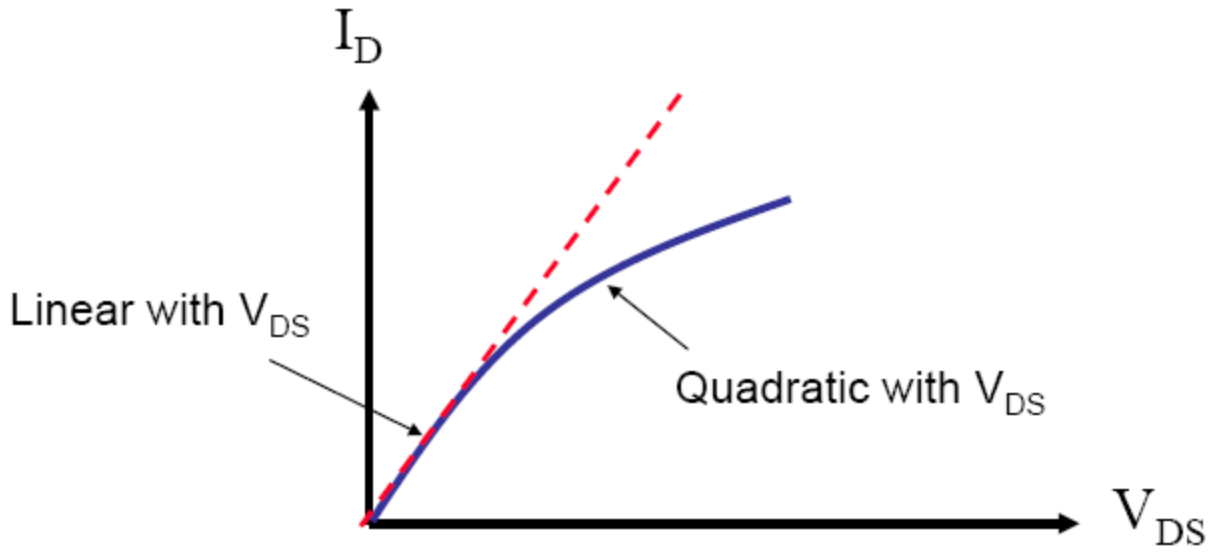
$$I_D = \mu \frac{W}{L} C_{OX} \left(V_G - V_T - \frac{V_{DS}}{2} \right) V_{DS}$$



Simplified Analysis for Linear and Saturation MOSFET Regime

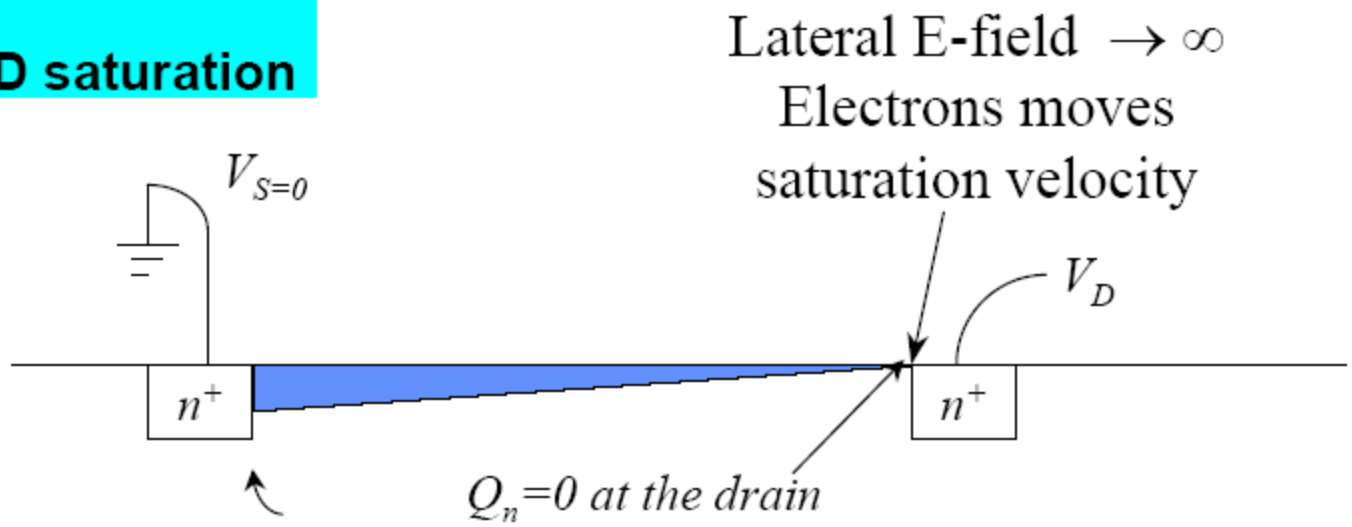
$$\text{With } v_{drift} = -\mu_n E \approx \frac{\mu_n V_{DS}}{L}$$

$$I_D = \mu \frac{W}{L} C_{OX} \left(V_G - V_T - \frac{V_{DS}}{2} \right) V_{DS}$$



Simplified Analysis for Linear and Saturation MOSFET Regime

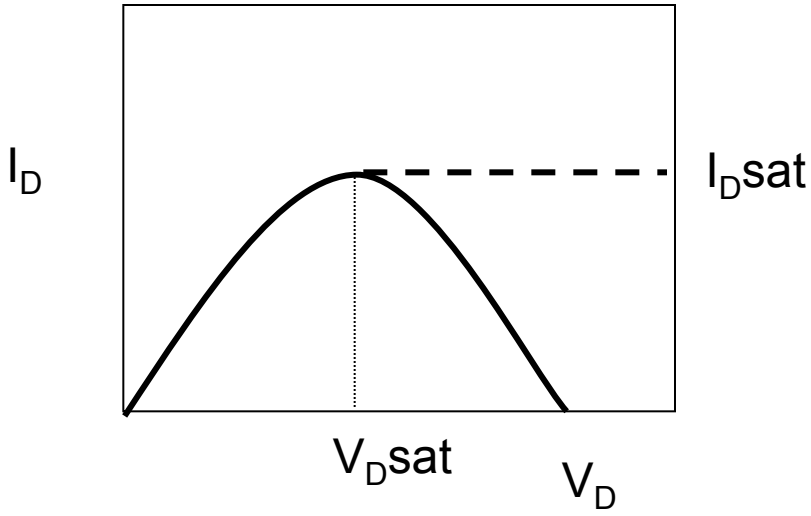
$V_{D\ saturation}$



V_{Dsat} is *defined* to be the value of V_D with Q_n=0 *at drain.*

From $Q_n = C_{ox}(V_G - V_T - V_D)$, we get $V_{Dsat} = V_G - V_T$

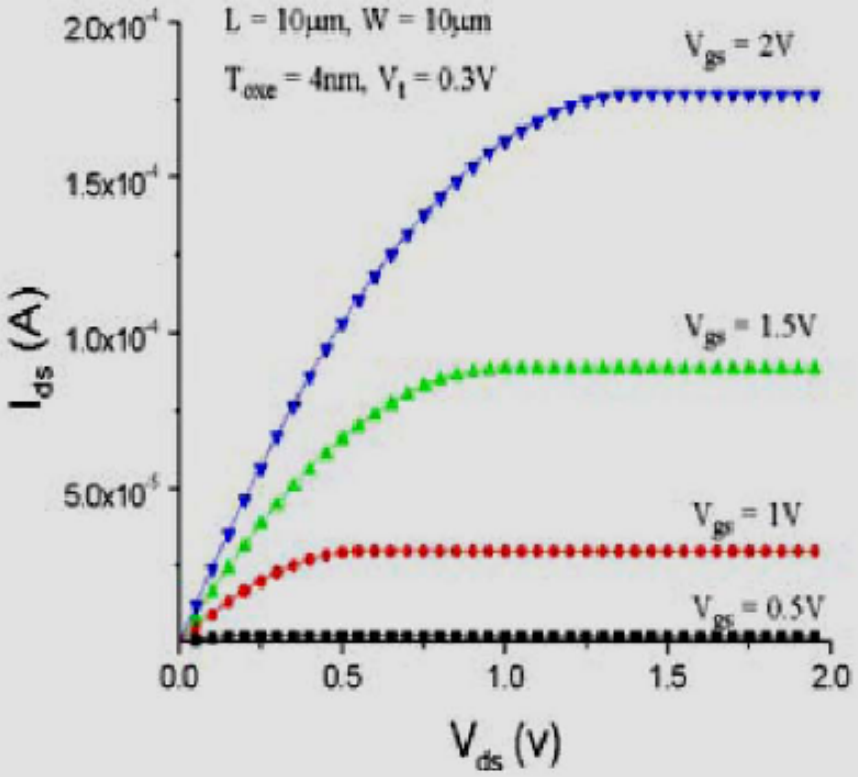
Simplified Analysis for Linear and Saturation MOSFET Regime



This makes it clear that the branch for $V_d > V_{dsat}$ has to be discarded as unphysical because it would imply that for $V_d > V_{dsat}$ we would have at the end of the channel a positive (hole) inversion charge

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left((V_g - V_T)V_{ds} - \frac{m}{2}V_{ds}^2\right)$$

Simplified Analysis Saturation MOSFET Regime



- saturation region:

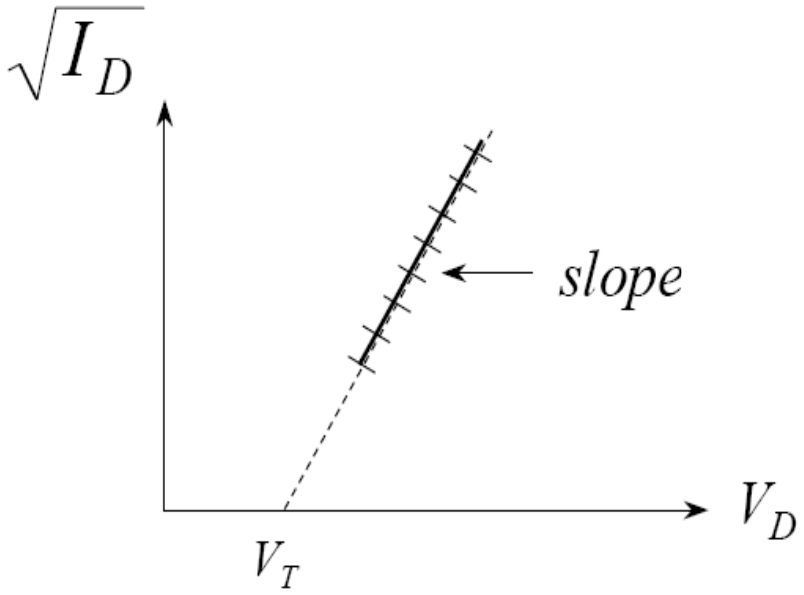
$$V_D \geq V_{Dsat} = V_{GS} - V_T$$

$$I_{Dsat} = \frac{W}{2L} C_{oxe} \mu_{eff} (V_{GS} - V_T)^2$$

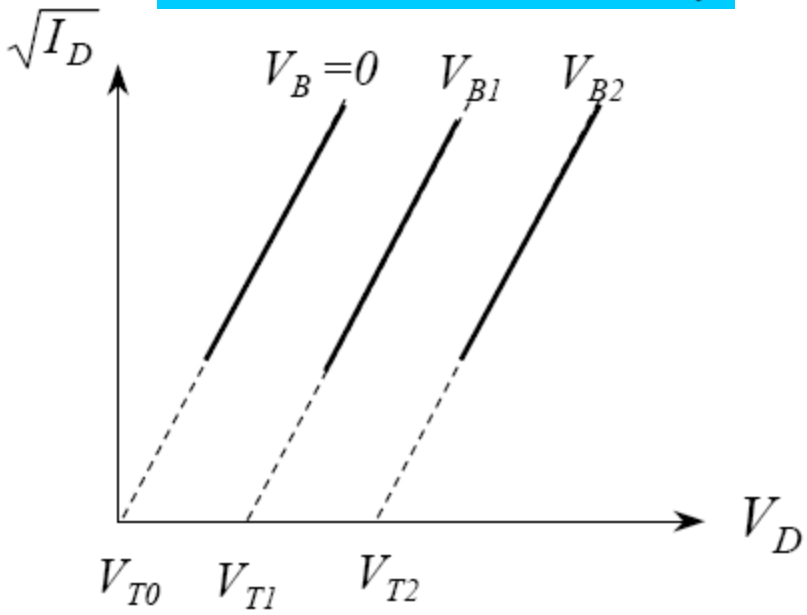
Note that we can extract V_T now from saturation region. In practice, the extractions of V_T from linear and saturation regions differed.

Threshold Voltage Extraction from linear region

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L} [V_g - V_T(V_B)] V_{ds}$$



Body Coefficient γ



Subthreshold Characteristics

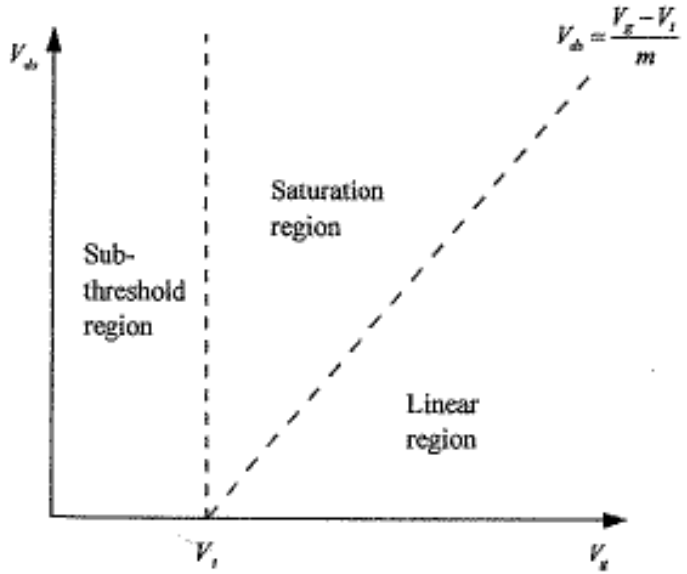


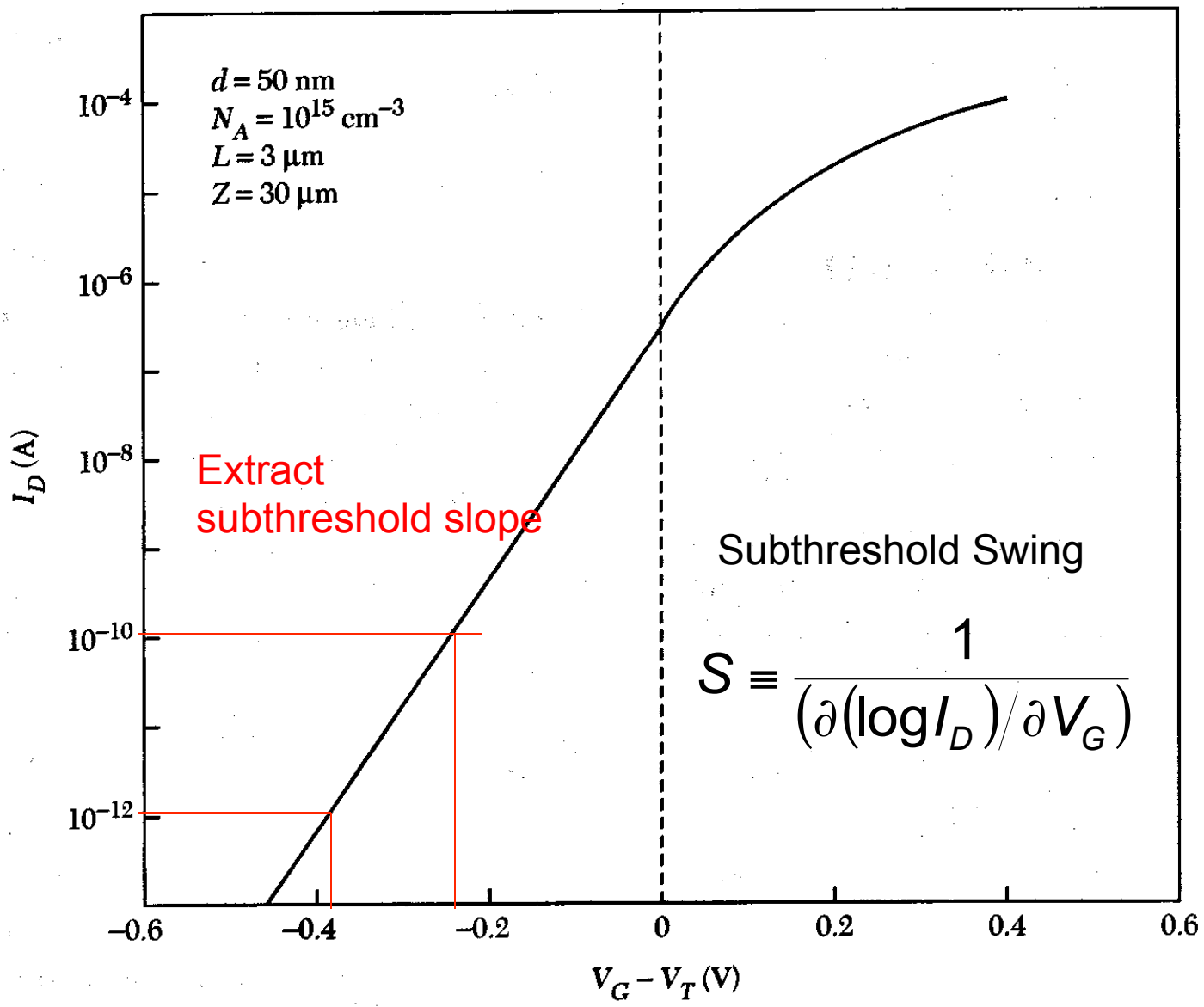
FIGURE 3.9. Three regions of MOSFET operation in the V_{ds}–V_g plane.

On linear scale, the off-current is negligible. However, on a logarithmic scale, the drain current remains at non-negligible levels for several tenths of a volt below V_T.

This is because the inversion charge density does not drop to zero abruptly. Rather, it follows an exponential dependence on ψ_s or V_g.

$$n(x,y) = n(\psi, V) = \frac{n_i^2}{N_A} \exp(\beta[\psi - V\ ])$$

Subthreshold Slope or Swing



Subthreshold Current

Leakage current issues

- Leakage vs. performance trade-off:
  - For high-speed, need small V_T and L
  - For low leakage, need high V_T and large L (to reduce DIBL and V_T roll-off)
- Process scaling
  - V_T reduces with each new process (historically)
  - Leakage increases ~10X!
- One solution: dual-V_T process
  - Low-V_T transistors: use in critical paths for high speed
  - High-V_T transistors: use to reduce power

Subthreshold Characteristics

Subthreshold behavior is of particular importance in low-voltage, low-power applications. The subthreshold is the region immediately below V_T.

($\psi_B < \psi_s < 2\psi_B$) is called the WEAK INVERSION region and has received a lot of attention in the last 15 years.

[Comment: Swiss watch industry, integrated circuits for watches, Prof. Eric Vittoz, University of Lausanne.]

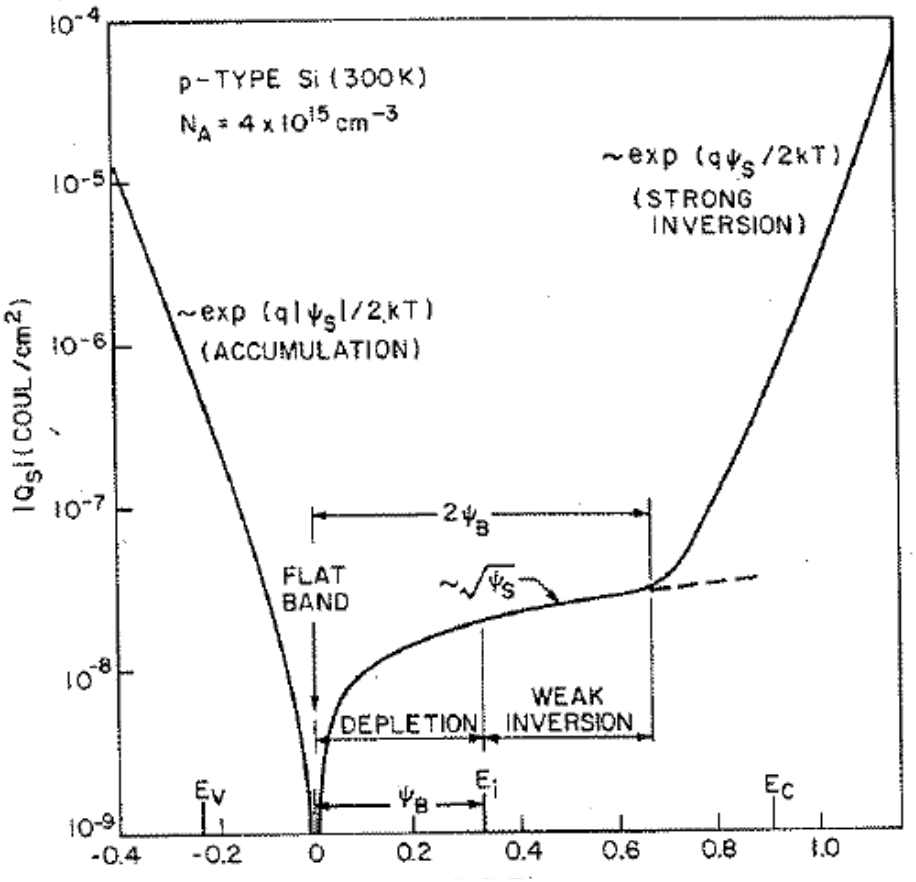
We know already the exact relation between surface potential and electric field at the semiconductor surface valid for any band bending.

$$E_s^2 =$$

$$= \left(\frac{2kTN_A}{\varepsilon_s}\right)\left[\left(\exp(-\beta\psi) + \beta\psi - 1\right) + \frac{n_i^2}{N_A}\left(\exp(-\beta V)\exp(\beta\psi) - \beta\psi - 1\right)\right]^{1/2}$$

Subthreshold Characteristics

recall



$$E_s^2 =$$

$$= \left(\frac{2kTN_A}{\varepsilon_s}\right)\left[\left(\exp(-\beta\psi) + \beta\psi - 1\right) + \frac{n_i^2}{N_A}\left(\exp(-\beta V)\exp(\beta\psi) - \beta\psi - 1\right)\right]^{1/2}$$

Subthreshold Characteristics

Drift and Diffusion Components of the Current

Unlike the strong inversion region, in which the drift current dominates, subthreshold conduction is dominated by the **diffusion current**. Both components are included in the Pao-Sah double-integral formula. In general, current continuity applies to the total current, NOT to its components. In other words, the ratio between drift and diffusion may vary from one point of the channel to another.

At low drain bias it is possible to separate drift and diffusion components using ψ_s (V) relation. When qV/kT<<1, only the first terms of the expansion in V need to be kept.

$$I_{ds}(y) = -\mu_{eff} W \frac{dV(y)}{dy} Q_n(V) \approx -\mu_{eff} W \frac{dV(y)}{dy} Q_n(0)$$

Since the total current is proportional to dV/dy, the drift fraction of the current is given by the change of surface potential (band bending) with respect to the quasi-Fermi potential, i.e. $d\psi_s/dV$.

Subthreshold Characteristics

Drift and Diffusion Components of the Current

This can be evaluated from eq.:

$$V_g = V_{FB} + \psi_s + \frac{\sqrt{2\varepsilon_s kTN_A}}{C_{ox}}\left[\frac{q\psi_s}{kT} + \frac{n_i^2}{N_A}\exp[\beta(\psi_s - V_D)]\right]$$

In the limit for small V we obtain;

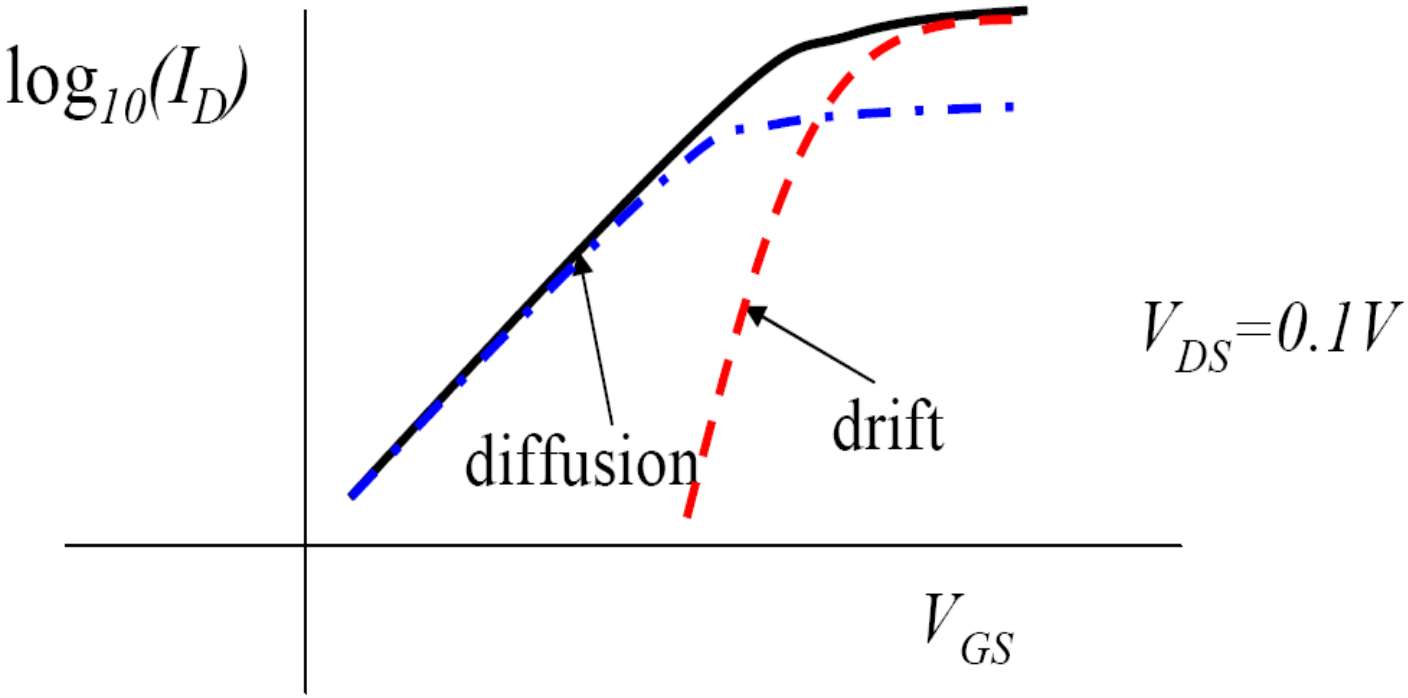
$$\frac{d\psi_s}{dV_D} = \frac{\frac{n_i^2}{N_A^2}\exp[\beta\psi_s]}{1 + \frac{n_i^2}{N_A}\exp[\beta(\psi_s - V_D)] + \frac{C_{ox}^2}{\varepsilon_s qN_A}\frac{|Q_s|}{C_{ox}}}$$

In the weak inversion where ($\psi_B < \psi_s < 2\,\psi_B$), the numerator is much smaller than 1 and the diffusion current dominates.

Conversely, beyond strong inversion, d ψ_s /dV≈1 and the drift current dominates. This is shown in Figure on next page.

Subthreshold Characteristics

Drift and Diffusion Components of the Current



Derivation of Subthreshold Current

$$E_s^2 =$$

$$= \left(\frac{2kTN_A}{\varepsilon_s}\right)\left[\left(\exp(-\beta\psi) + \beta\psi - 1\right) + \frac{n_i^2}{N_A}\left(\exp(-\beta V)\exp(\beta\psi) - \beta\psi - 1\right)\right]^{1/2}$$

From E_s we can calculate Q_s and we keep only the two significant terms in the weak inversion region

$$-Q_s = \varepsilon_s E_s = \sqrt{2\varepsilon_s kTN_A}\left[\frac{q\psi_s}{kT} + \frac{n_i^2}{N_A^2}\exp[q(\psi_s - V)/kT]\right]^{1/2}$$

$$Q_d = -qNAWd = -\sqrt{2\varepsilon_s qN_A\psi_s}$$

Of course the first term is the depletion charge density (we know this from our threshold voltage formula) and the second term gives the (weak) inversion charge density, after expanding into a power series and identifying the zeroth-order term as Q_d, one obtains:

$$-Q_n = \sqrt{2\varepsilon_s qN_A/(2\psi_s)}\,\frac{kT}{q}\,\frac{n_i^2}{N_A^2}\exp[q(\psi_s - V)/kT]$$

Derivation of Subthreshold Current

$$-Q_n = \sqrt{2\varepsilon_s kTN_A/(2\psi_s)}\left[\frac{kT}{q}\frac{n_i^2}{N_A^2}\exp[q(\psi_s - V)/kT]\right]$$

Substituting Q_n in $$I_{ds} = \mu_{eff} W \int_0^{V_{ds}} [-Q_n(V)]dV$$

We obtain

$$I_{ds} =$$

$$\mu_{eff}\frac{W}{L}\sqrt{2\varepsilon_s kTN_A/(2\psi_s)}\left(\frac{kT}{q}\frac{n_i}{N_A}\right)^2 \exp(q\psi_s/kT)[1-\exp(-qV_{ds}/kT)]$$

With $$V_g = V_{FB} + \psi_s + \frac{\sqrt{2\varepsilon_s qN_A\psi_s}}{C_{ox}}$$

The last equation can be solved for ψ_s . After few more transformations we can eliminate ψ_s from the drain current formula and replace it by an expression with V_g.

Derivation of Subthreshold Current

We obtain the final formula for the subthreshold current:

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 \exp(q(V_g - V_T)/kT)[1-\exp(-qV_{ds}/kT)]$$

Note that the subthreshold current depends exponentially on both the gate and drain bias.

The subthreshold current is however independent of V_{ds} once V_{ds} is larger than a few kT/q as we would expect for diffusion-dominated current transport.

The dependence on the gate voltage is exponential and is called *subthreshold slope.*

Subthreshold Slope and Swing

We define the subthreshold **slope** as

$$\ln 10 = 2.3$$

$$S = \left(\frac{d(\log_{10} I_{ds})}{dV_g} \right)^{-1} = \frac{mkT}{q} \ln 10 = 2.3 \frac{kT}{q} \left(1 + \frac{C_d}{C_{ox}} \right)$$

Which is typically for bulk transistors 70-100 mV/decade. Here $m=1+C_d/C_{ox}$.

Subthreshold slope gives the gate voltage decrease to reduce the current by one decade.

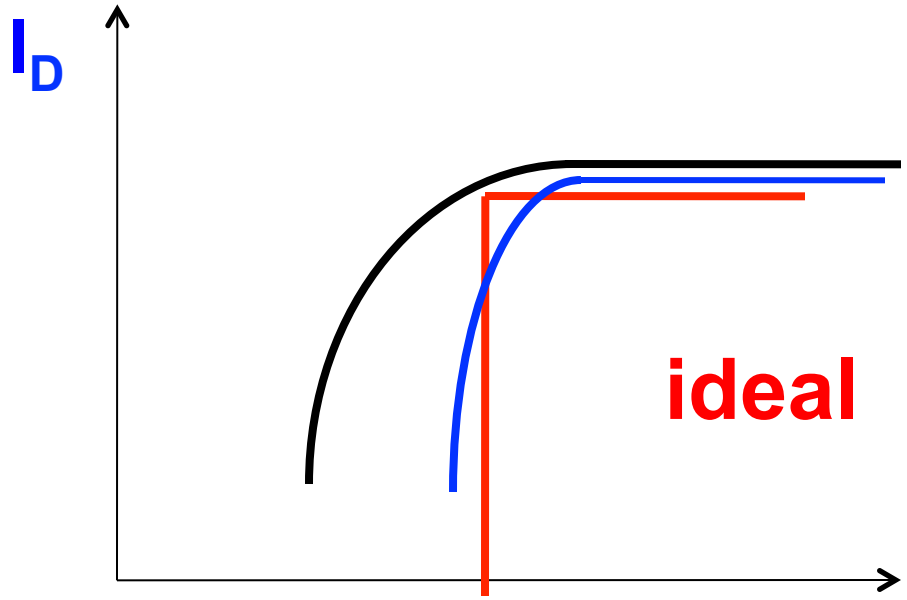
Note that subthreshold slope decreases with smaller depletion capacitance (lighter p-substrate doping → larger depletion width). This runs counter to the scaling trends of the MOSFETs. (Solution: Fully depleted SOI transitors)

Note that the term in parenthesis is the inverse capacitive divider ratio $(C_{ox}+C_d)/C_{ox}$.

Theoretical Limit of Subthreshold Slope

$$S = 2.3\frac{kT}{q}\left(1+\frac{C_d}{C_{ox}}\right) \rightarrow 2.3\frac{kT}{q}$$

$$\xrightarrow{300K} \quad 2.3\times 0.0259 = 60mV/dec$$



Theoretical Limit of Subthreshold Slope

$$S = 2.3\frac{kT}{q}\left(1+\frac{C_d}{C_{ox}}\right) \rightarrow 2.3\frac{kT}{q}$$

$$\overset{300K}{\rightarrow} \quad 2.3\times 0.0259 = 60mV/dec$$

$(1+C_d/C_{ox})$ depends on the oxide thickness d (C_{ox}) and on the maximum depletion width W_d (C_d). The maximum depletion width depends in turn on the substrate doping (assuming uniform substrate doping N_A) and substrate doping can be characterized by the Debey length $L_D(N_A)$. Therefore the factor is often characterized in terms of parameter a defined as:

$$a = \sqrt{2}(\varepsilon_s / L_D)/C_{ox} = \sqrt{2}(\varepsilon_s / \varepsilon_{ox})(d_{ox} / L_D)$$

Subthreshold Slope and Oxide damage

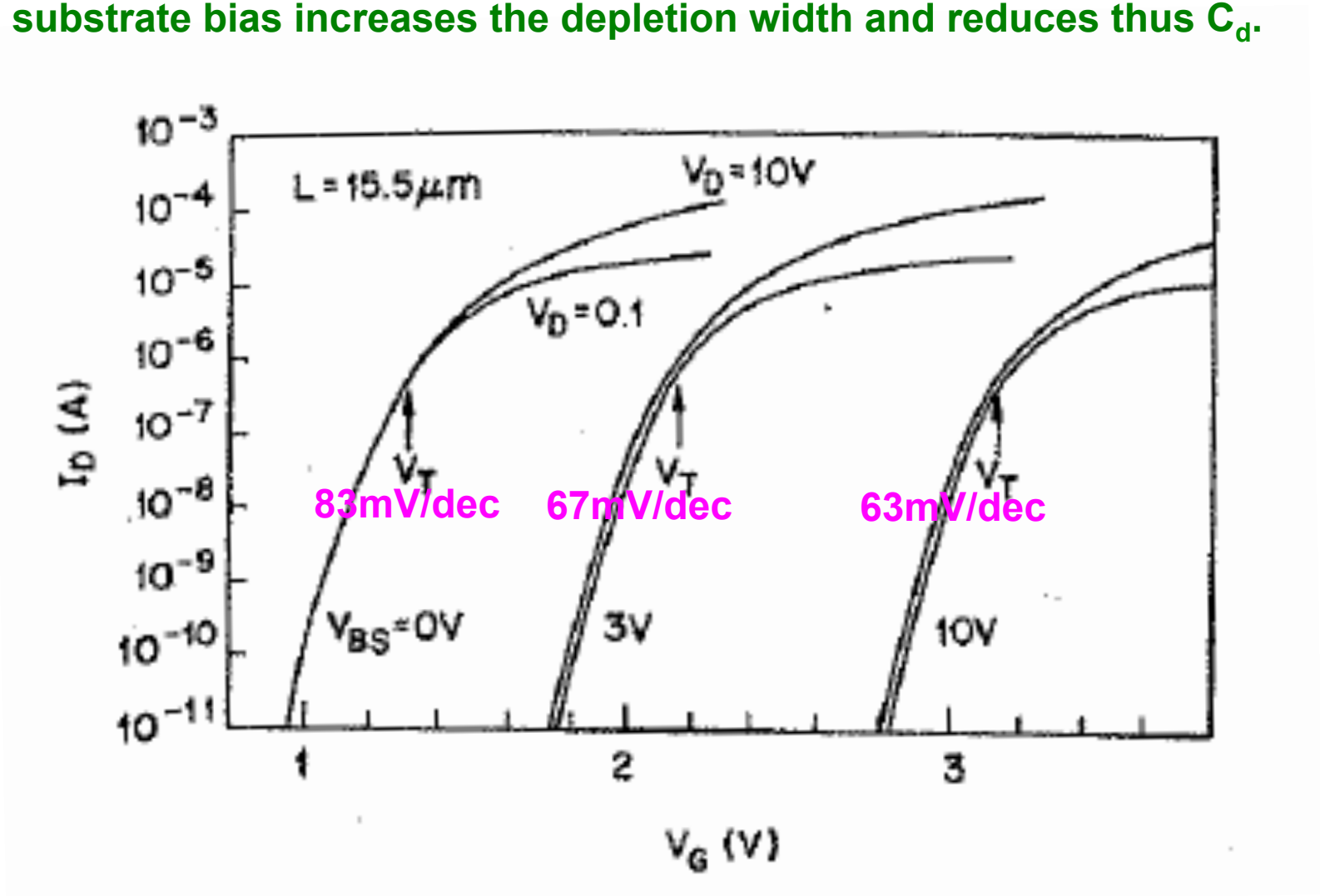
If there is a significant interface-trap density, the capacity C_{it} associated with the interface traps is in parallel with the depletion layer capacitance C_d. So we have to substitute (C_d+C_{it}) for C_d and obtain

$$S = 2.3\frac{kT}{q}\left(1+\frac{C_d}{C_{ox}}\right)\times\frac{1+(C_d+C_{it})/C_{ox}}{1+C_d/C_{ox}}$$

The extra factor is larger than one and **degrades** the subthreshold slope.

Subthreshold Slope and Substrate Bias

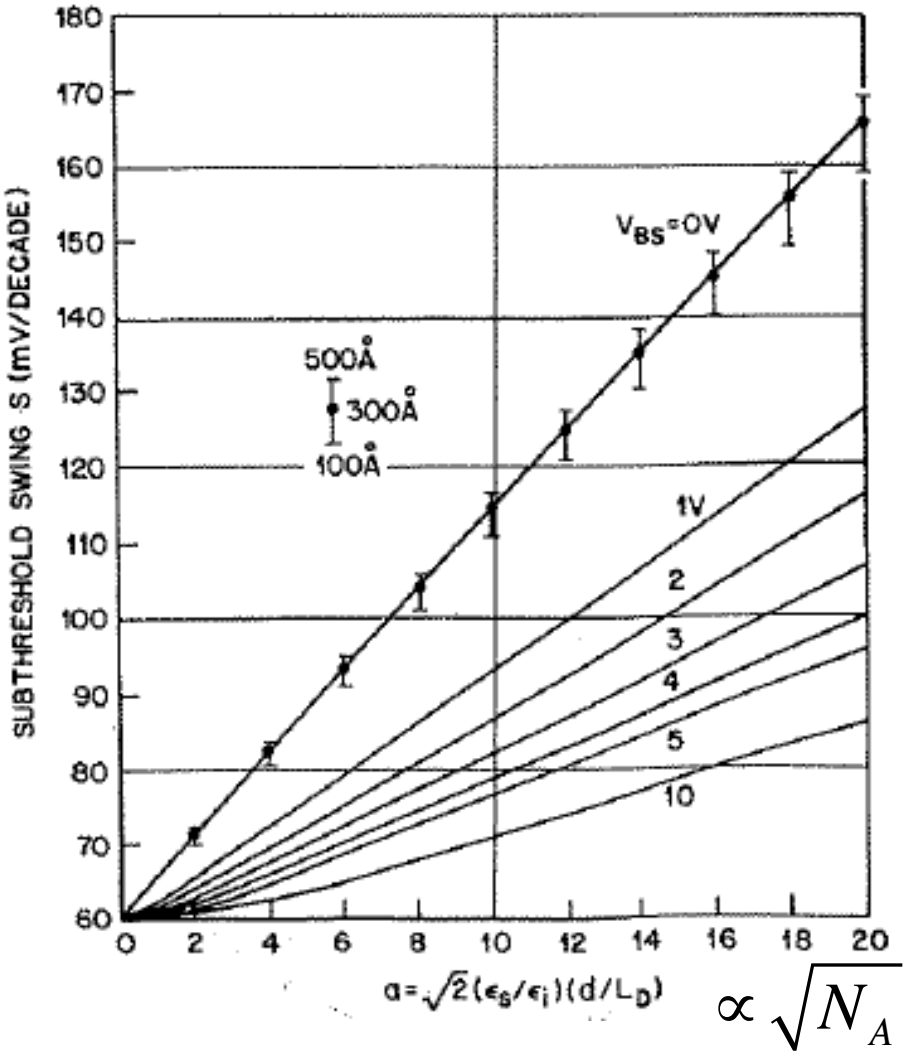
Subthreshold slope can be improved by applying substrate bias because the substrate bias increases the depletion width and reduces thus C_d.



Subthreshold swing versus substrate doping and versus substrate reverse bias.

Subthreshold Slope and Substrate Bias

Subthrehold swing versus substrate doping and versus substrate reverse bias.



$$a = \sqrt{2}(\varepsilon_s / L_D) / C_{ox}$$

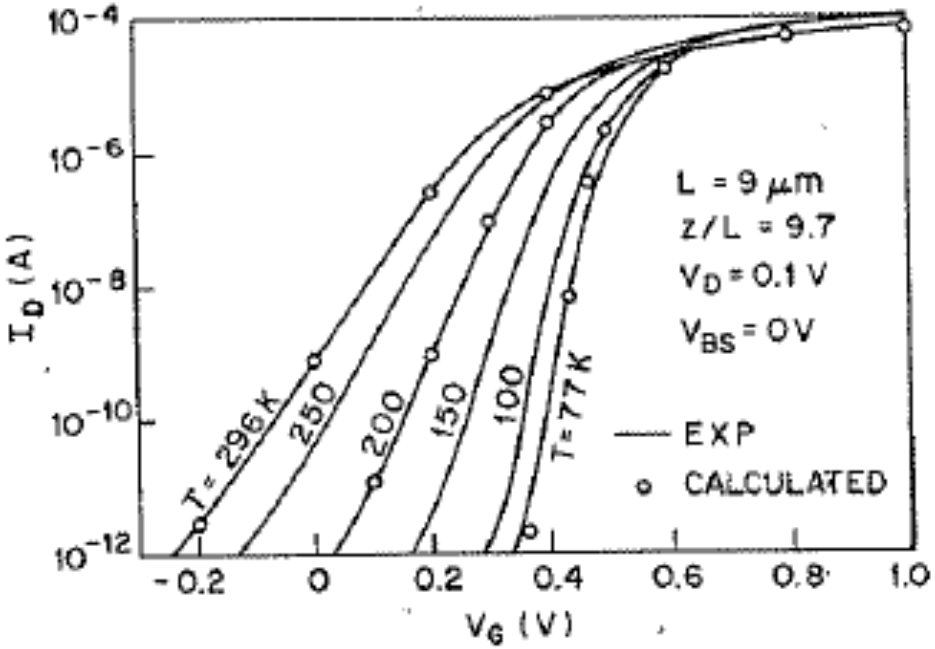
Essentially ratio between flat band capacitance and gate oxide capacitance

Subthreshold Slope and Temperature

$$S = \ln 10 \times \frac{kT}{q}\left(1 + \frac{C_d}{C_{ox}}\right)$$

$$S_{\min} = \ln 10 \times \frac{kT}{q} = 60mV/dec$$

$$@\ \ T = 300K$$



As T decreases, the switching properties of a MOSFET improve. As T decreases from 296K to 77K, the V_T increases from 0.25V to about 0.5V. The most important improvement is the reduction of subthreshold swing from 80mV/dec at 296K to 22mV/decade (a dream for CMOS circuit designers for room temperature applications) at 77K – thus an improvement of almost factor of 4. Of course this improvement is coming mainly from the kT/q term.

Other improvements of low temperature operation include higher mobility, lower leakage current and lower metal resistance.

However MOSFET would have to be immersed in liquid nitrogen (not a very practical proposition).

Alternative (Intuitive) Derivation of Subthreshold Current

We already know that in a weak inversion the current is dominated by diffusion and can be derived in the same way as in pn junction or in the bipolar transistor. Considering the MOSFET as an n-p-n back-to-back junctions, we have

$$I_d = -qAD_n \frac{dn}{dy} = qAD_n \frac{n(0) - n(L)}{L}$$

But

$$n(0) = n_{p0} \exp(\beta \psi_s)$$

$$n(L) = n_{p0} \exp[\beta(\psi_s - V_D)]$$

We have to determine A. A is product of the width W of the device times the thickness of the channel x_j. Because of the exponential dependence of electron density on the potential ψ_s, the effective channel thickness corresponds to the distance in which ψ_s decreases by kT/q. Therefore, the effective channel thickness is kT/qE_s where E_s is the weak-inversion surface field

$$x_i = \frac{kT}{q} \cdot \frac{1}{E_s}$$

$$E_s = -Q_B / \varepsilon_s = \sqrt{2qN_A \psi_s / \varepsilon_s}$$

Alternative (Intuitive) Derivation of Subthreshold Current

$$n(0) = n_{p0}\exp(\beta\psi_s) \qquad E_s = -Q_B/\varepsilon_s = \sqrt{2qN_A\psi_s/\varepsilon_s}$$

$$n(L) = n_{p0}\exp[\beta(\psi_s - V_D)] \qquad x_i = kT/\varepsilon_s E_s$$

Here we make the assumption from the beginning that subthreshold current is due to diffusion.

$$I_d = -qAD_n\frac{dn}{dy} = qAD_n\frac{n(0)-n(L)}{L}$$

Substituting the two first equation in the third one, we obtain:

$$I_d = \mu_n\left(\frac{W}{L}\right)\frac{aC_{ox}}{2\beta^2}\left(\frac{n_i}{N_A}\right)^2(1-\exp(-\beta V_d)\exp(-\beta\psi_s)(\beta\psi_s)^{1/2}$$

$$\psi_s = (V_g - V_{FB}) - \frac{a^2}{2\beta}\left\{\left[1+\frac{4}{a^2}(\beta V_g - \beta V_{FB} - 1\right]^{1/2} - 1\right\}$$

$$a = \sqrt{2}(\varepsilon_s/L_D)/C_{ox}$$

From this the subthreshold swing can be found in the same way as previously.

More complete model – sub-threshold to saturation

- Must include diffusion and drift currents
- Still use gradual channel approximation
- Yields sub-threshold and saturation behavior for long channel MOSFETS
- Exact Charge Model – numerical integration (Pao-Sah Model)

$$I_D = \frac{Z}{L}\frac{\varepsilon_s \mu_n}{L_D}\int_{0}^{V_D}\int_{\psi_B}^{\psi_s} \frac{e^{\beta\psi-\beta V}}{F\left(\beta\psi, V, \frac{n_{p0}}{p_{p0}}\right)}$$