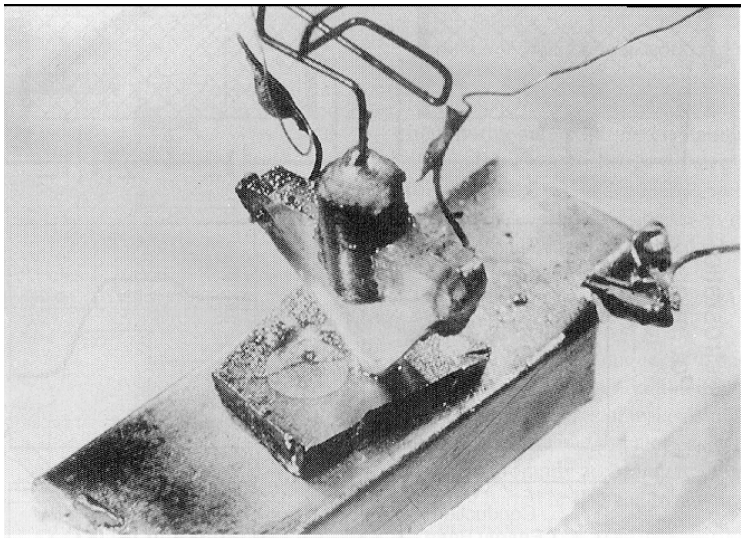
Lecture 6

The MOS Transistor

A bit of History

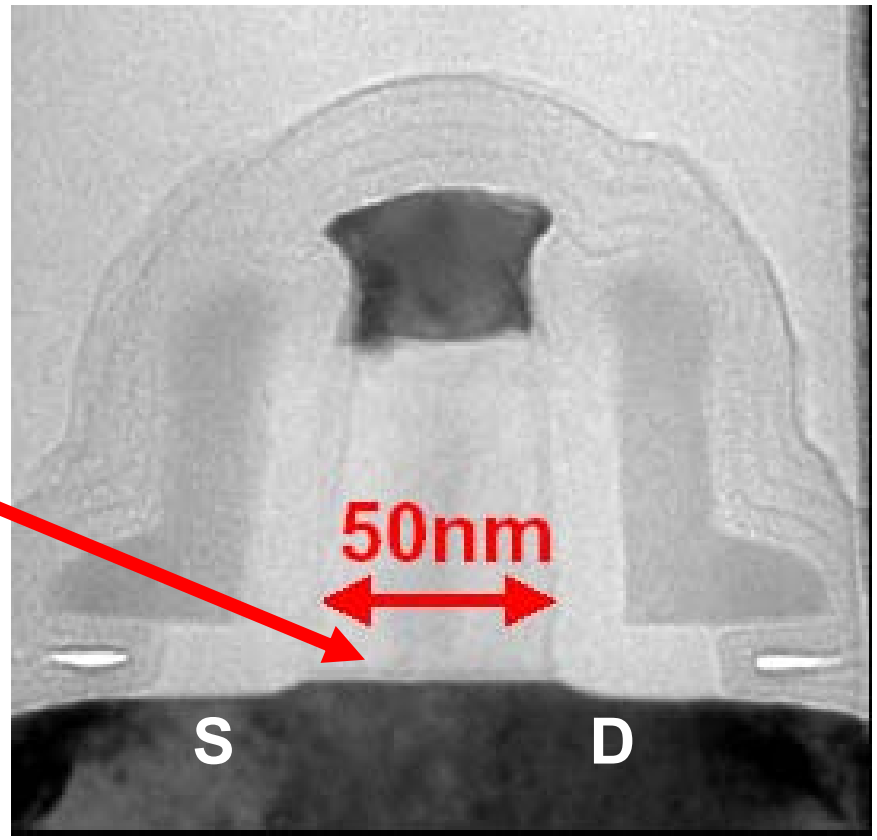
- In the 1970s, a new kind of FET was developed, called metal-oxide semiconductor field-effect transistor (MOSFET).
- MOSFETs have been extremely popular due to some inherent advantages over other types of transistors:
  - They can be made smaller
  - They consume less power
  - Their manufacturing process is well controlled
- Microprocessors and memory chips are today implemented using exclusively MOSFETs.
- There are different types of MOSFETs. We will discuss the ***enhancement-type MOSFET***, which has been adopted by the semiconductor industry. We will refer to this transistor as just "MOS" transistor.



First transistor, Bell Labs, 1948

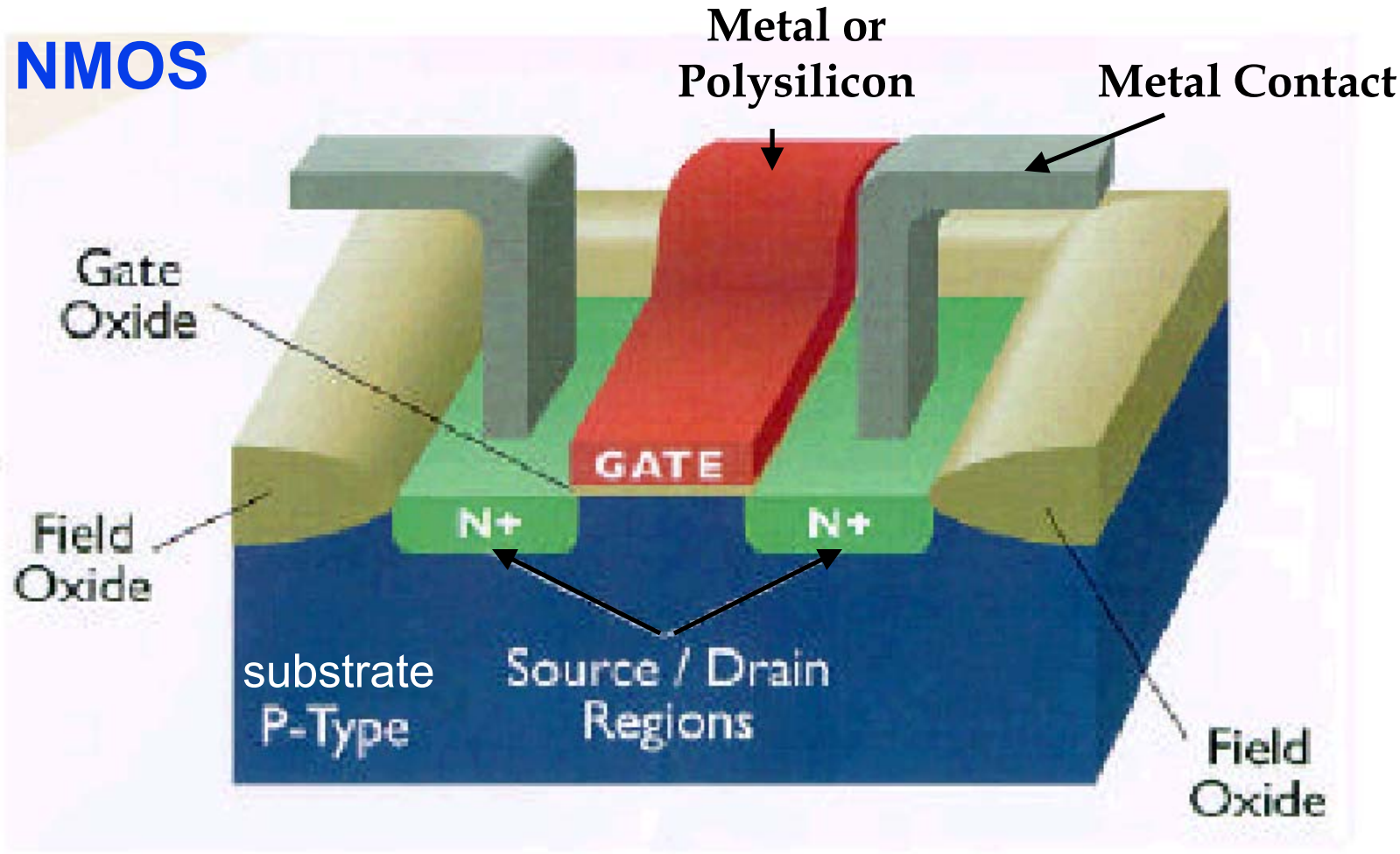
Intel 50nm transistor in production

~2000x smaller than diameter of human hair



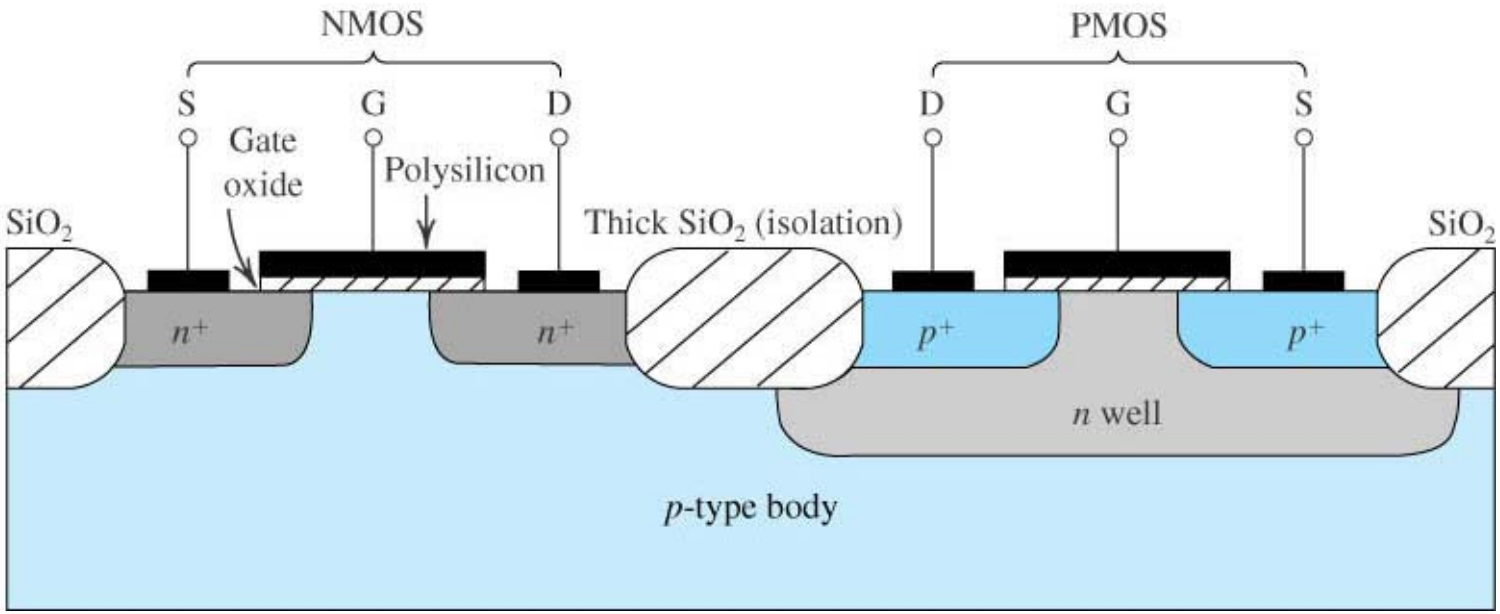
50nm transistor dimension

A Quick Look at the MOS Transistor



For a PMOS transistor, Source/Drain regions are P^+ and substrate is N-type.

CMOS technology



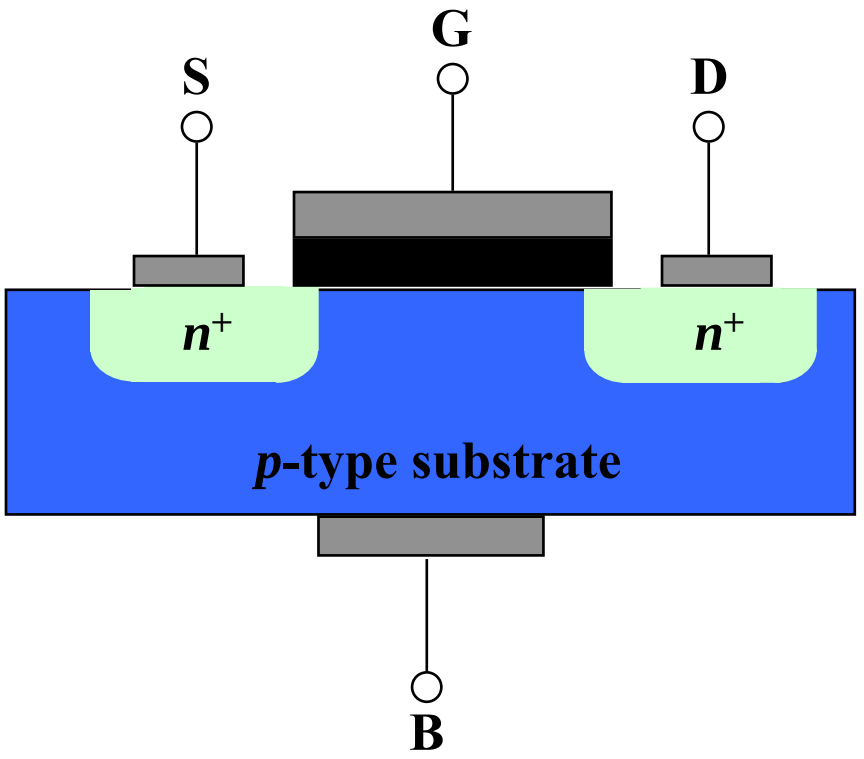
Cross-section of a CMOS integrated circuit. Note that the PMOS transistor is formed in a separate *n*-type region, known as an *n* well. Another arrangement is also possible in which an *n*-type body is used and the *n* device is formed in a *p* well. Not shown are the connections made to the *p*-type body and to the *n* well; the latter functions as the body terminal for the *p*-channel device.

The MOS Terminals

- Purely symmetric device: source and drain are interchangeable depending on the applied voltages;
- To differentiate between source and drain, use the following definitions:

  For NMOS, the *source* terminal is biased at a lower potential than the *drain*;

  For PMOS, the *source* terminal is biased at a higher potential than the *drain* terminal.



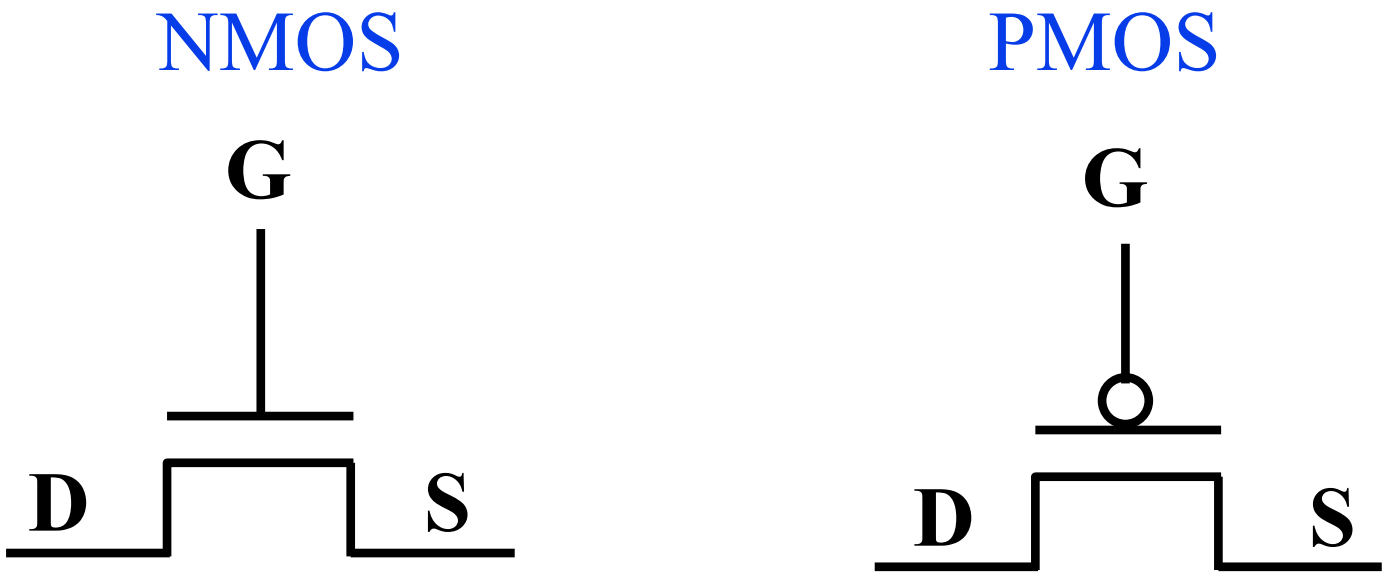
- The BODY (also referred to as Bulk) terminal is usually connected to a fixed potential:

  For NMOS, the BODY is connected to 0V

  For a PMOS, the BODY is connected to the highest potential, that is the power supply V_{DD}. Assume V_{DD} ranging from 2.5 to 5V

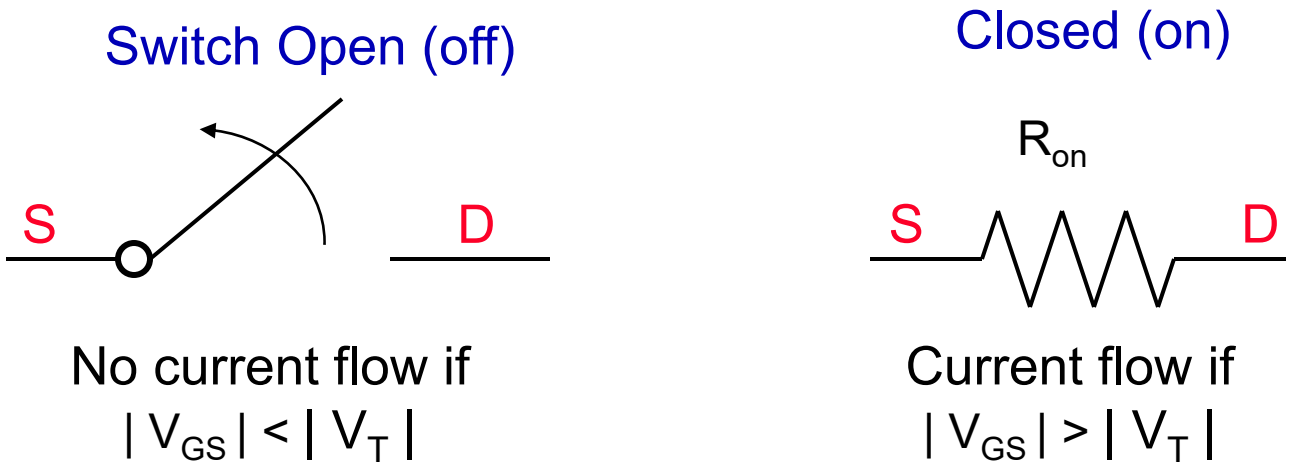
SYMBOLS Used FOR MOS Transistors

- We will use the following symbols to represent MOS transistors:



- Since the bulk voltage is at a fixed potential, we do not represent the body or bulk terminal.

Switch Model of MOS Transistor



- The MOS Transistor can be modeled as a ***voltage controlled switch***.
- V_T is called the threshold voltage of the transistor:

  | V_T |=0.1 to 0.5V depending on the technology used

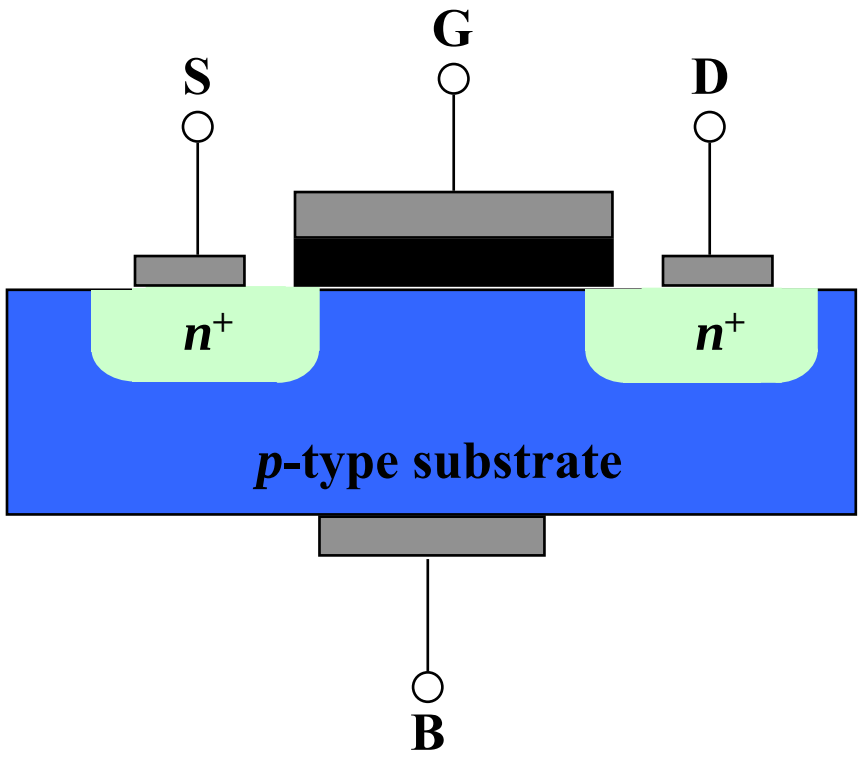
  V_T is positive for NMOS transistors

  V_T is negative for PMOS transistors

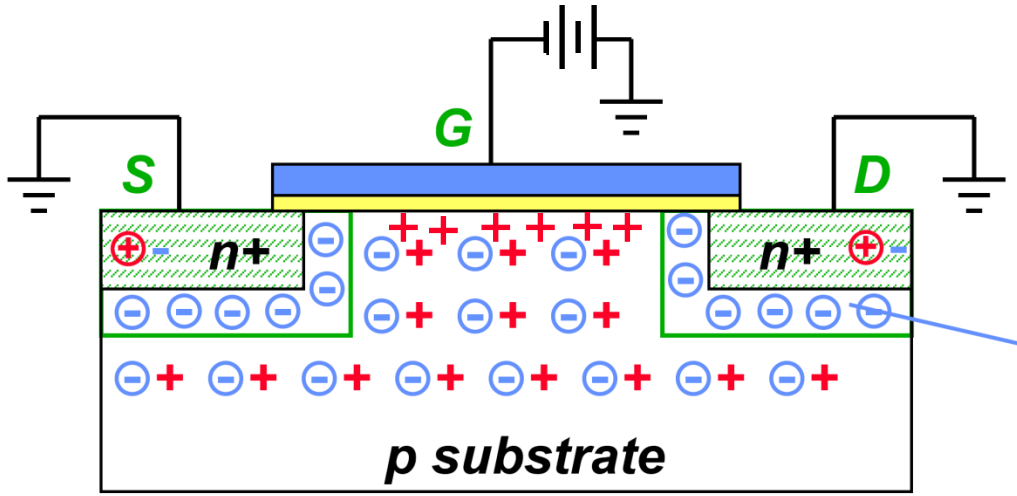
- Subscript convention: $V_{GS} = V_G - V_S$, $V_{DS} = V_D - V_S$, etc.

Operation Principle

- The region underneath the gate oxide between the source and drain is called the Channel.
- Body is commonly tied to Ground for NMOS
- In a MOSFET, Drain/Bulk and Source/Bulk junctions are kept reverse biased.
- Current flows between the diffusion terminals if the voltage on the gate terminal is large enough to create a conducting “channel”, otherwise the diffusion terminals are not connected.



Effect of Negative Gate Bias: $V_{GS} < 0$



+ mobile holes

- mobile electrons

⊖ immobile acceptors

⊕ immobile donors

Depletion layer

Negative gate voltage: $V_{GS} < 0$

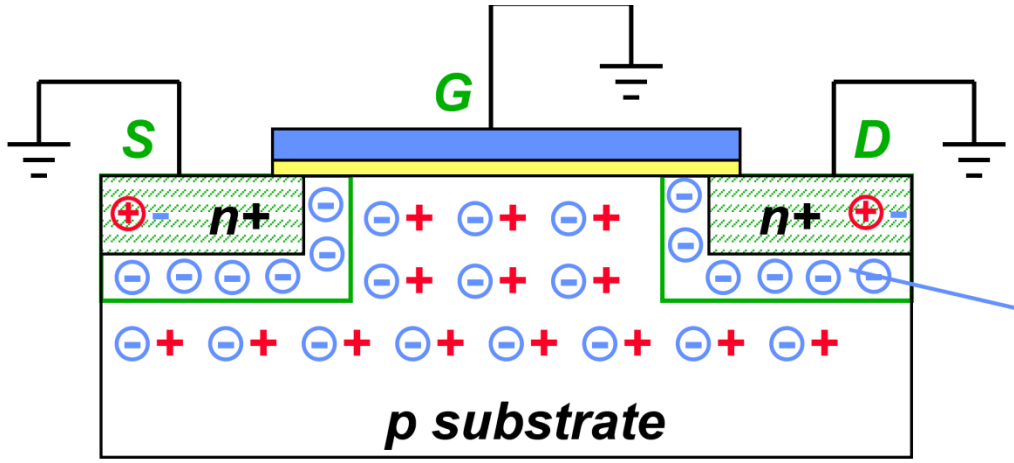
- Initially, consider the case where the source, drain and substrate are all connected to ground. We will examine the effect of applying a gate bias.

- In this case, the MOS transistor operates similarly to a capacitor. The gate acts as one plate of the capacitor, and the surface of the silicon, just under the thin insulating $Si0_2$, acts as the other plate.

- If the gate voltage is very negative, positive charge will be attracted to the channel region. Since the substrate was originally doped p, this negative gate voltage has the effect of simply increasing the channel doping to p^+, resulting in what is called an *accumulated channel.* The n^+ source and drain regions are separated from the p^+ channel region by depletion regions, resulting in the equivalent circuit of two back-to-back diodes.

- Thus, only leakage current will flow through source and drain.

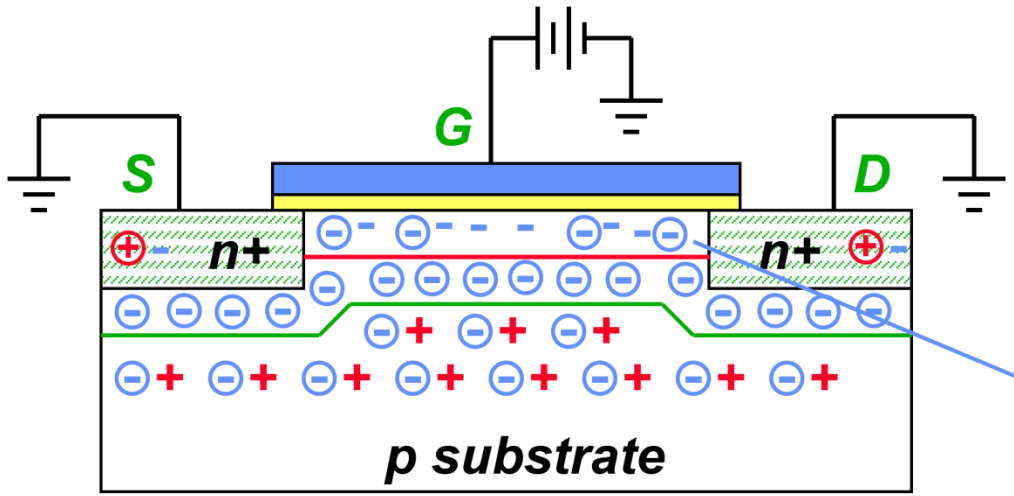
Effect of Gate Bias: $V_{GS} \geq 0$



+ mobile holes
- mobile electrons
⊖ immobile acceptors
⊕ immobile donors

Depletion layer

$V_{GS} < V_T$
Cutoff region



$V_{GS} \geq V_T$

Inversion layer forms

Operation: Positive gate voltage

- Let us connect a positive voltage to the gate. Since source is grounded, the gate voltage appears in effect between the source and gate and is denoted by V_{GS}.
- The positive voltage on the gate has two effects:

  The free holes (which are positively charged) in the substrate, are repelled away from the gate (channel region). These holes are pushed down in the substrate, leaving behind a carrier-depletion region under the gate. The depletion region is now populated by bound negative charge (uncovered because the neutralizing holes have been pushed downward into substrate) associated with the acceptor atoms.

  The positive gate voltage attracts electrons from the n^+ source and drain region into the channel region.

- Because of this two fold effect, when sufficient number of electrons are accumulated under the gate on the surface of substrate, an n region is in effect created, connecting the source (n+) and the drain (n+). It is called ***inversion layer***, as p-type layer has been changed to n-type just by applying a gate voltage.

Operation: Positive gate voltage

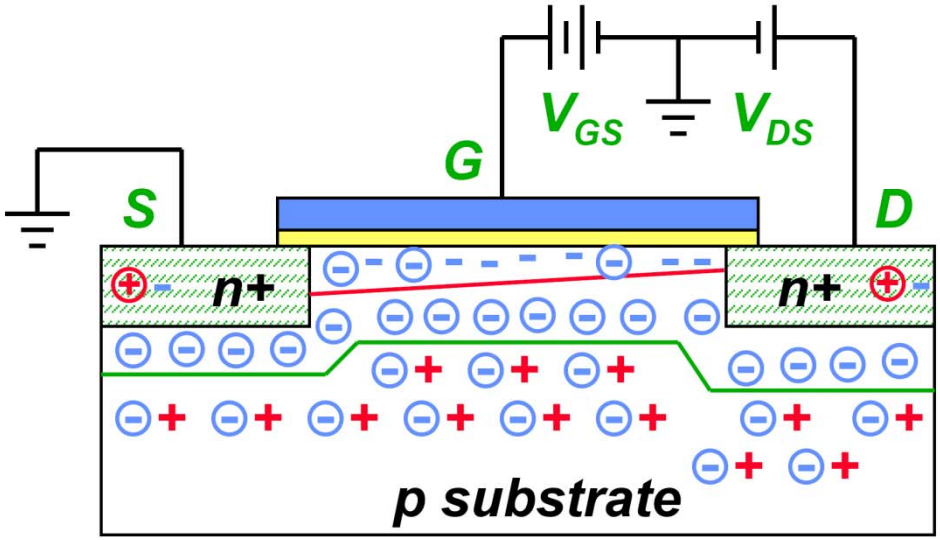
- This induced ***n***-region forms a **channel** between source and drain.
- Now if a positive voltage is applied between the source and the drain, the channel provides a path for the carrier movement, and electric current can flows between the source and drain through this channel.
- Since the channel formed is *n*-type and the carriers flowing through are electrons, this MOSFET is called n-channel MOSFET, alternatively, NMOS transistor.
- The value of V_{GS} at which a sufficient number of electrons are accumulated under the gate to form a conducting channel, is called V_T, which is the **threshold voltage** of the MOS. V_T is the gate-source voltage, for which the concentration of electrons under the gate is equal to the concentration of holes in the p substrate.
- Obviously the V_T is positive for NMOS, and the V_{GS} has to be increased to at least V_T before the MOS will start conducting.
- V_T is controllable during manufacturing of the MOSFET, and is kept typically between 0.35 and 0.7V.

Operation: Positive gate voltage

- For gate-source voltages larger than V_T, there is an ***n*-type channel** present, and conduction between the drain and the source can occur.
- For gate-source voltages less than V_T, it is normally assumed that the transistor is off and no current flows between the drain and the source. However, it should be noted that this assumption of zero drain-source current for a transistor that is off is only an approximation. In fact, for gate voltages around V_T, there is no abrupt current change, and for gate source voltages slightly less than V_T, small amounts of ***subthreshold current*** can flow !!
- When the gate-source voltage, V_{GS}, is larger than V_T, the channel is present. As V_{GS} is increased, the density of electrons in the channel increases. Indeed, the carrier density, and therefore the charge density, is proportional to $V_{GS} - V_T$, which is often called the *effective gate-source voltage* and denoted V_{eff}:

$$V_{eff} = V_{GS} - V_T$$

Linear Mode



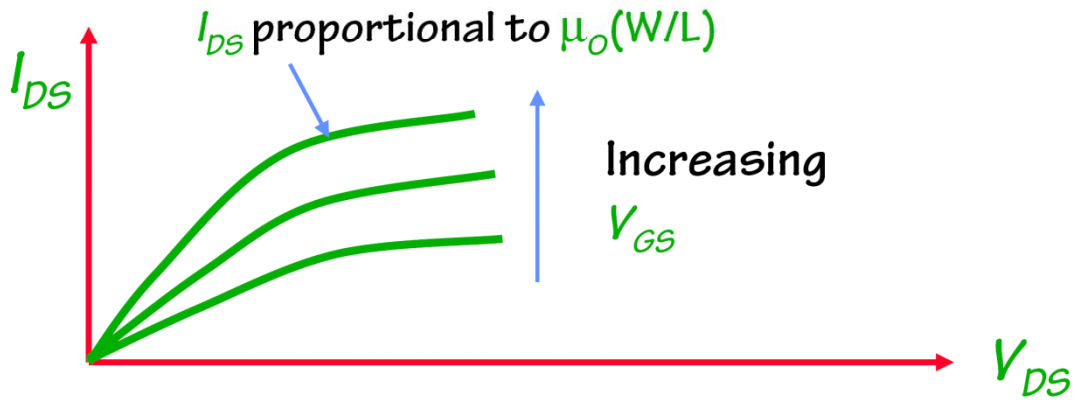
$V_{GS} \geq V_T$

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

Current flows from drain to source

Larger V_{DS} increases I_{DS}

Larger V_{GS} creates deeper channel which increases I_{DS}



Linear Mode

- Having induced the channel, If a small voltage (0.1V or 0.2V) is applied between the drain and source, a current I_{DS} will flow through the induced ***n*-**channel.

- Current is carried by the free electrons traveling from the source to drain through the channel, hence the names source and drain.

- The magnitude of the current I_{DS} depends on the density of electrons in the induced channel, which in turn depends on the magnitude of V_{GS}.

- At $V_{GS} = V_T$, the ***n*-**channel has got just induced. The current flow is negligibly small. As V_{GS} exceeds V_T, more electrons are attracted into the channel. The result is a channel of increased width, and conductance, in other words, decreased resistance. The conductance of channel is proportional to the effective gate-source V_{eff}.

Linear Mode

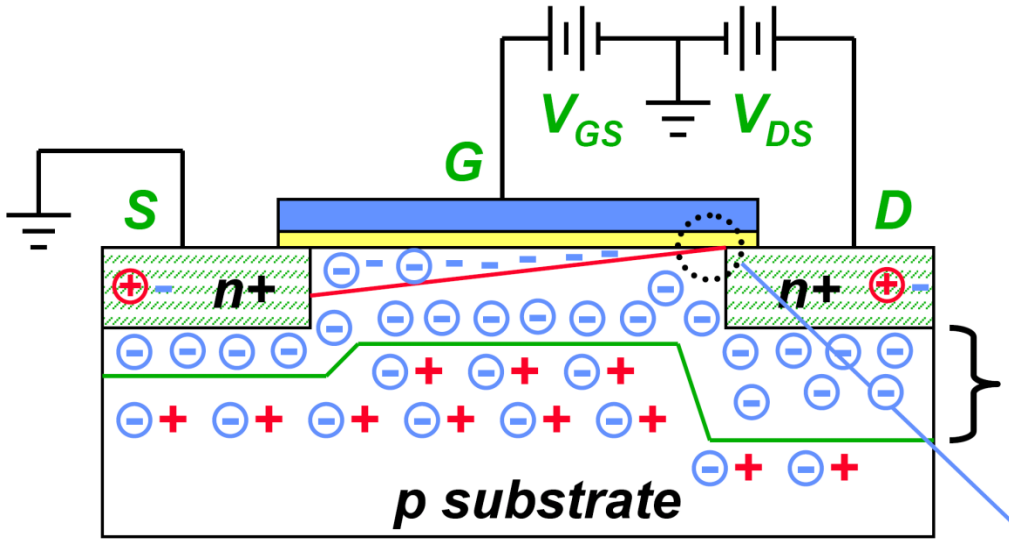
- The current I_{DS} is proportional to V_{eff}, and the voltage V_{DS} that causes the I_{DS} to flow. Thus, MOSFET is operating as a linear resistance, whose value is controlled by V_{GS}. The resistance is infinite for $V_{GS}<V_T$ and decreases for higher V_{GS}.

- Please note the linearity of the curves. NMOS is acting like a resistor, in the linear (or triode) region, whose resistance is controlled by gate voltage V_{GS}, once it is above V_T (V_{DS} is still small).

- Increasing V_{GS} above threshold voltage V_T enhances the channel, hence the name enhancement-type NMOS.

- Most important: The current that enters the drain terminal is same (equal) as that leaves the source terminal. i.e. $I_D = I_S = I_{DS}$.

Saturation Mode



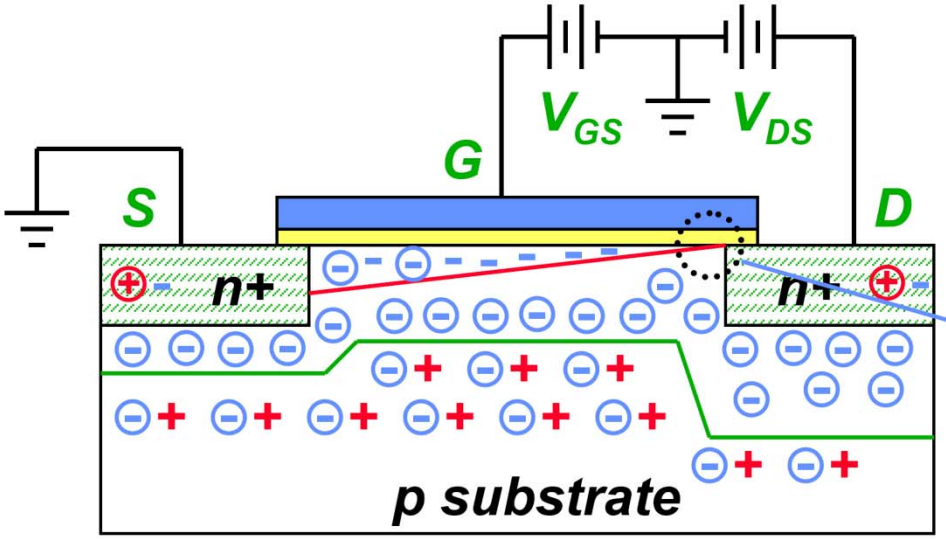
$V_{GS} \geq V_T$

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

Why is the depletion layer deeper near the drain?

Inversion layer is pinched off near the drain

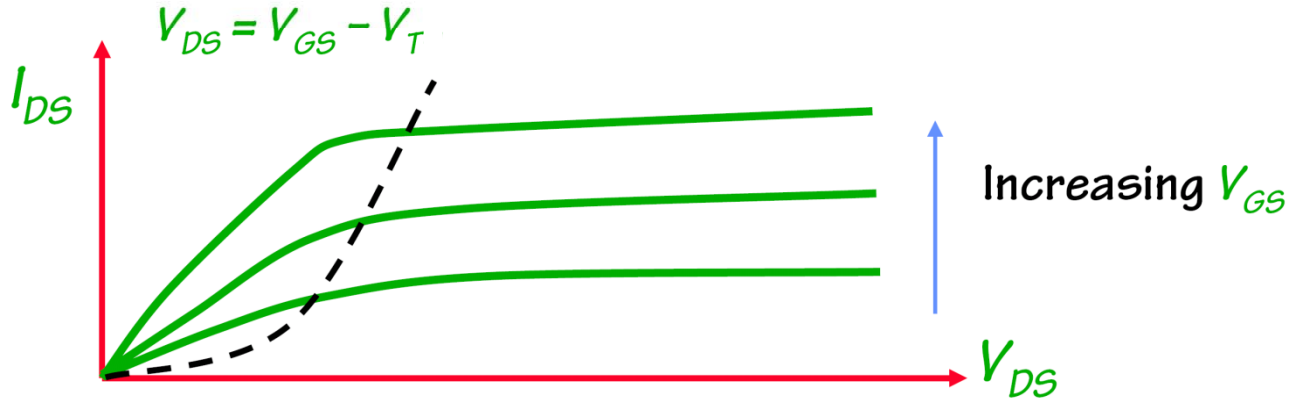
Saturation Mode



Inversion layer is pinched off at the drain

$V_{GS} \geq V_T$

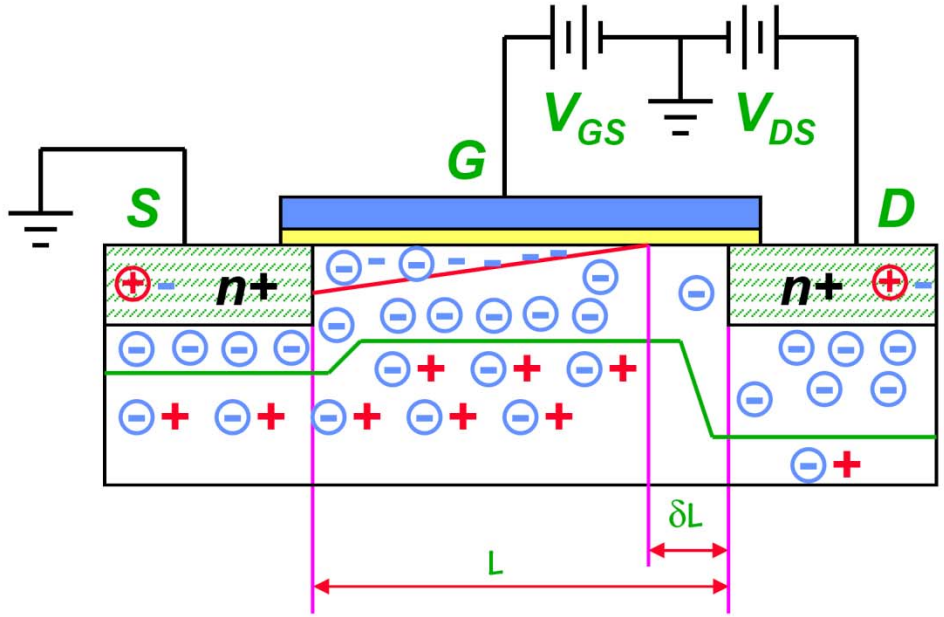
To the first order, once $V_{DS} \geq V_{GS} - V_T$, I_{DS} does not increase



Saturation Region: $V_{DS} > V_{GS} - V_T$

- Let us keep V_{GS} constant, above V_T.
- The voltage, as we travel across the channel, increases from 0V to V_{DS}. Thus the voltage between the gate and points along the channel decreases from V_{GS} at the source end to $V_{GS} - V_{DS}$ at the drain end.
- Since the channel depth depends on this voltage, the channel is no longer uniform, rather will take a tapered shape. As V_{DS} is increased further, channel becomes more tapered. The resistance of the channel increases correspondingly.
- The I_{DS}–V_{DS} curves do not continue as a straight line, but bend as shown. Eventually, when V_{DS} is increased to a value such that $V_{GS} - V_{DS} = V_T$, or in other words, $V_{DS} = V_{GS} - V_T$, the channel depth at the drain end decreases to almost zero.
- The channel is said to be ***pinched off***.
- Increasing V_{DS} beyond this value will have little effect (theoretically, no effect) on the channel shape.

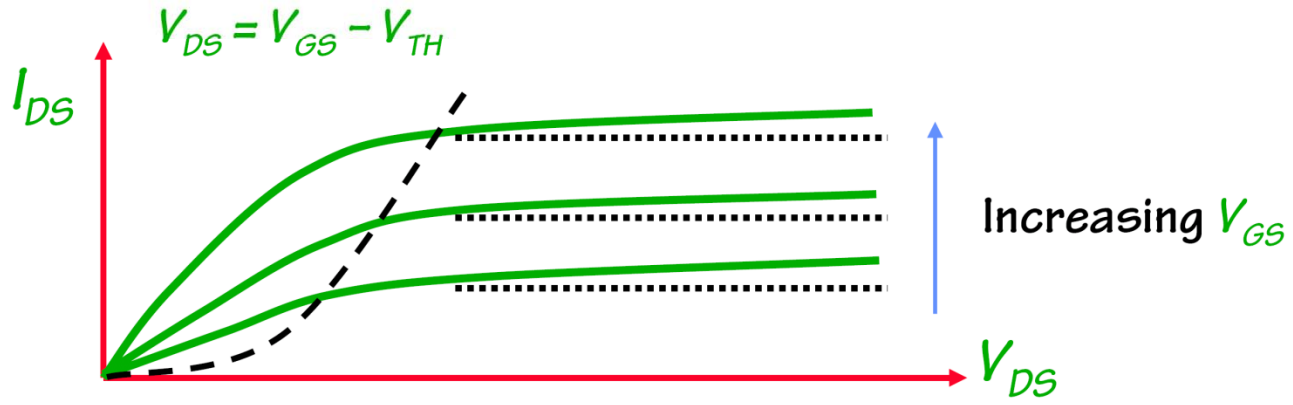
Channel Length Modulation



$V_{GS} \geq V_{TH}$

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

Increased V_{DS} results in effective channel length decreasing, i.e., δL getting larger, which increases I_{DS}



Channel Length Modulation

- As V_{DS} becomes larger than $V_{GS}-V_T$, the depletion region surrounding the drain junction increases its width; this in turn moves the channel pinch-off point away from the drain toward the source.

- Thus the effective channel length is reduced, a phenomenon known as ***channel-length modulation***.

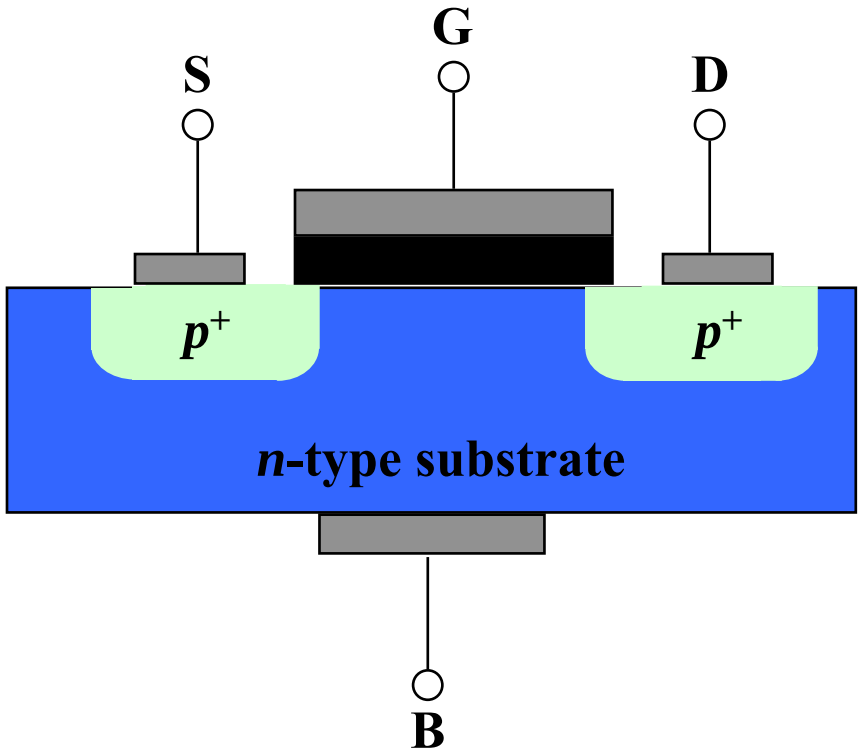
- Now since k is inversely proportional to the channel length, k and, correspondingly, I_{DS} increases with V_{DS}. Channel length modulation can be analytically accounted for by incorporating a factor λ independent of V_{DS} :

$$\mathbf{I_{DS} = I_{DSAT}(1 + \lambda V_{DS})}$$

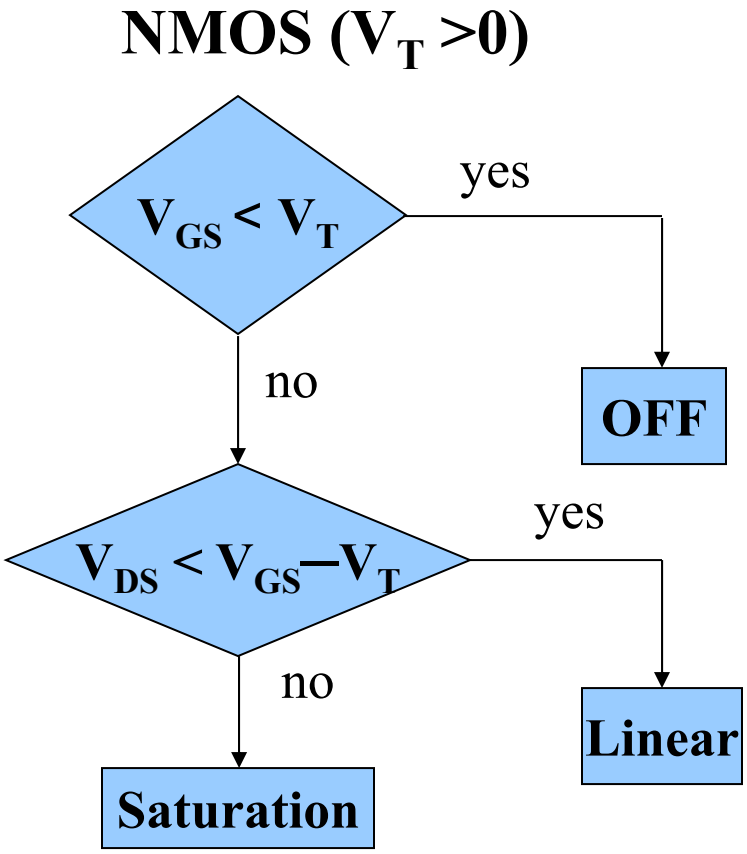
where λ is given by the technology. Typical values range from 0.005 to 0.03 V^{-1} .

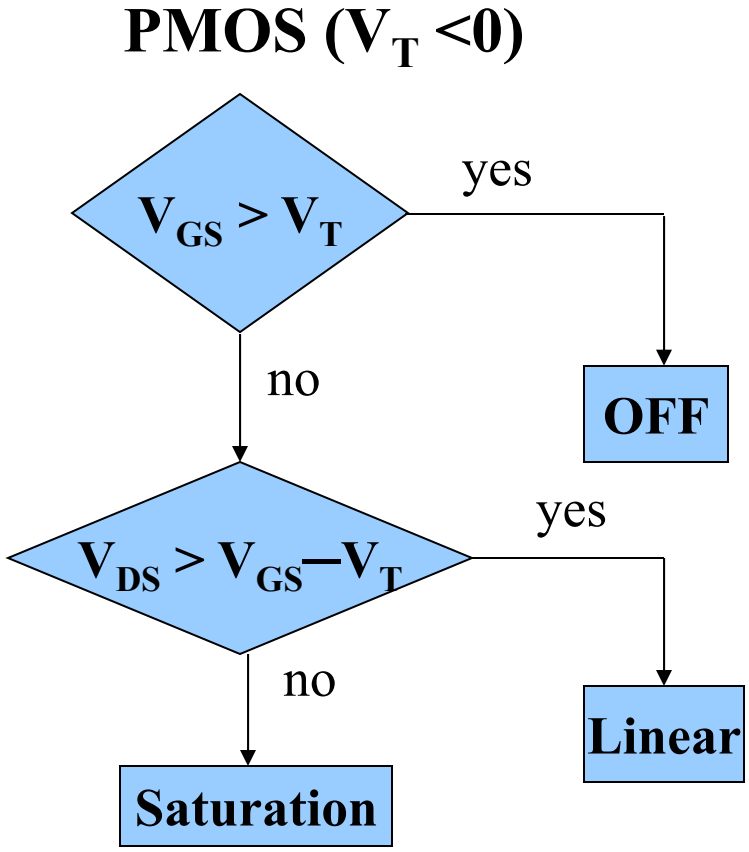
Enhancement type P-Channel MOS

- A p-channel enhancement-type MOSFET is fabricated on n-type substrate with p+-type regions for source and drain.
- Holes are the charge carriers which flow from the source to the drain.
- I_{DS}, V_{GS} and V_{DS} are negative.
- NMOS devices are faster than PMOS due to the fact that electrons have higher mobility than holes ($\mu_n > \mu_p$)

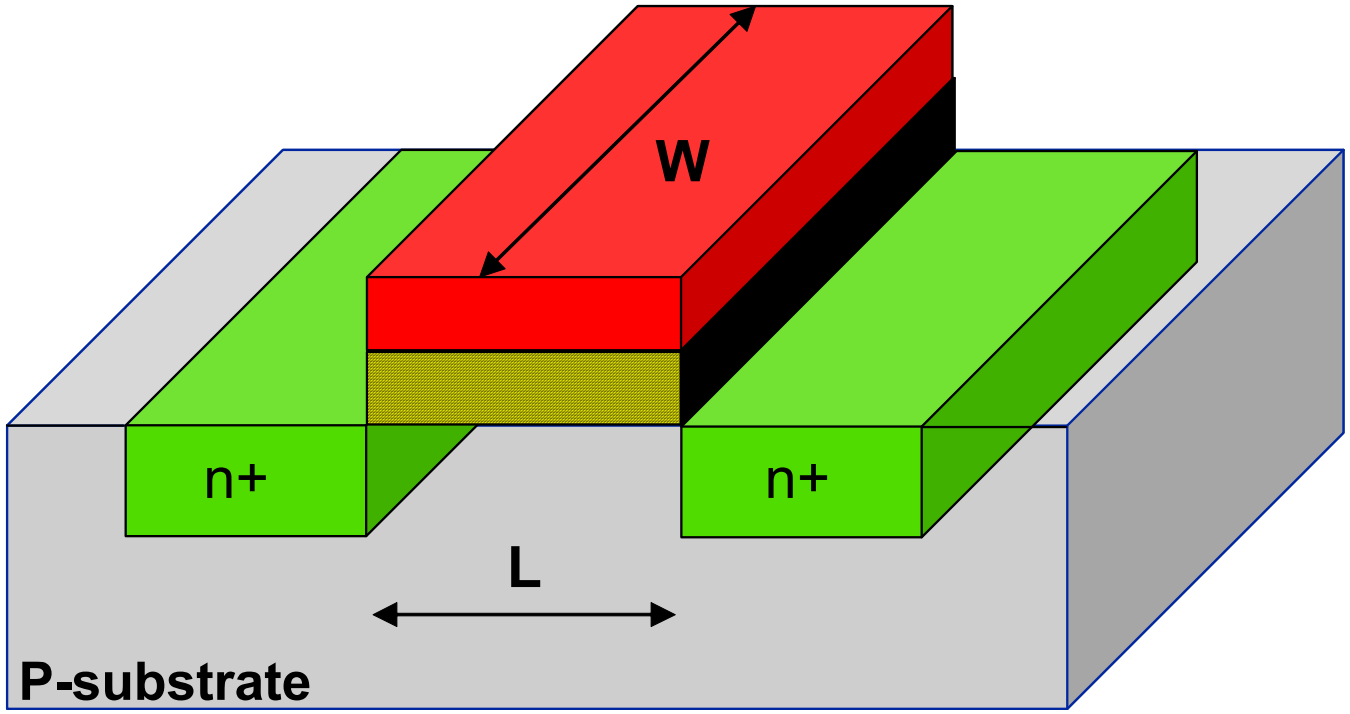


Operating mode of the transistor - Review



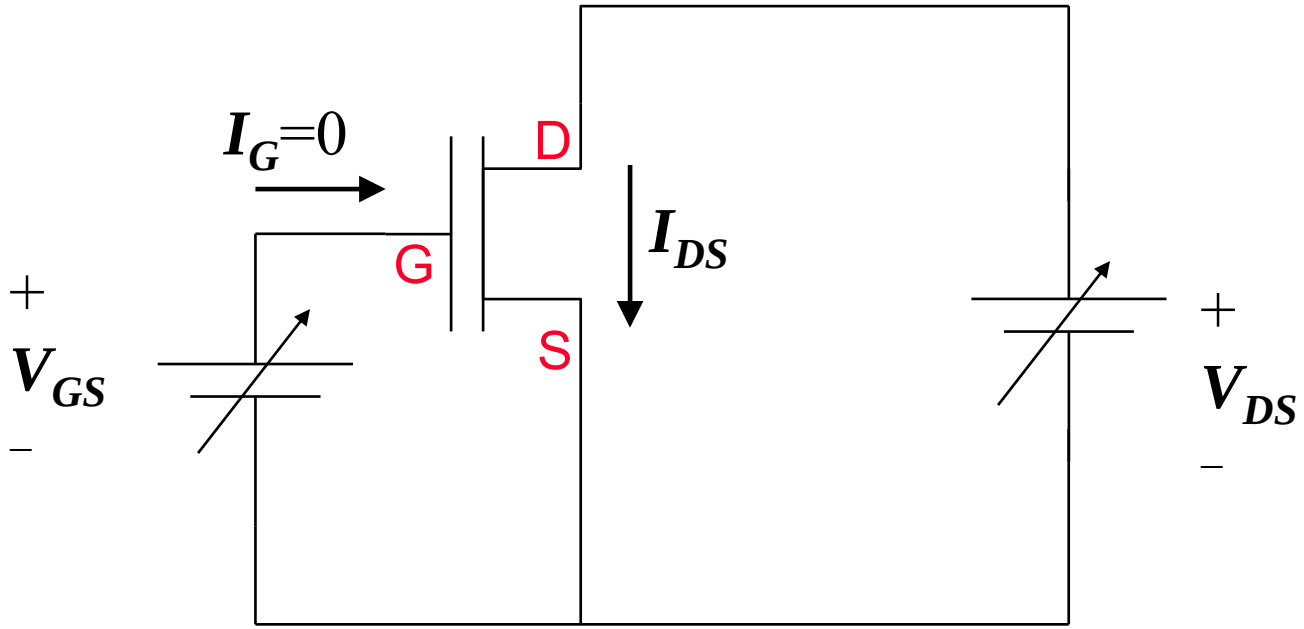


A close look at the NMOS

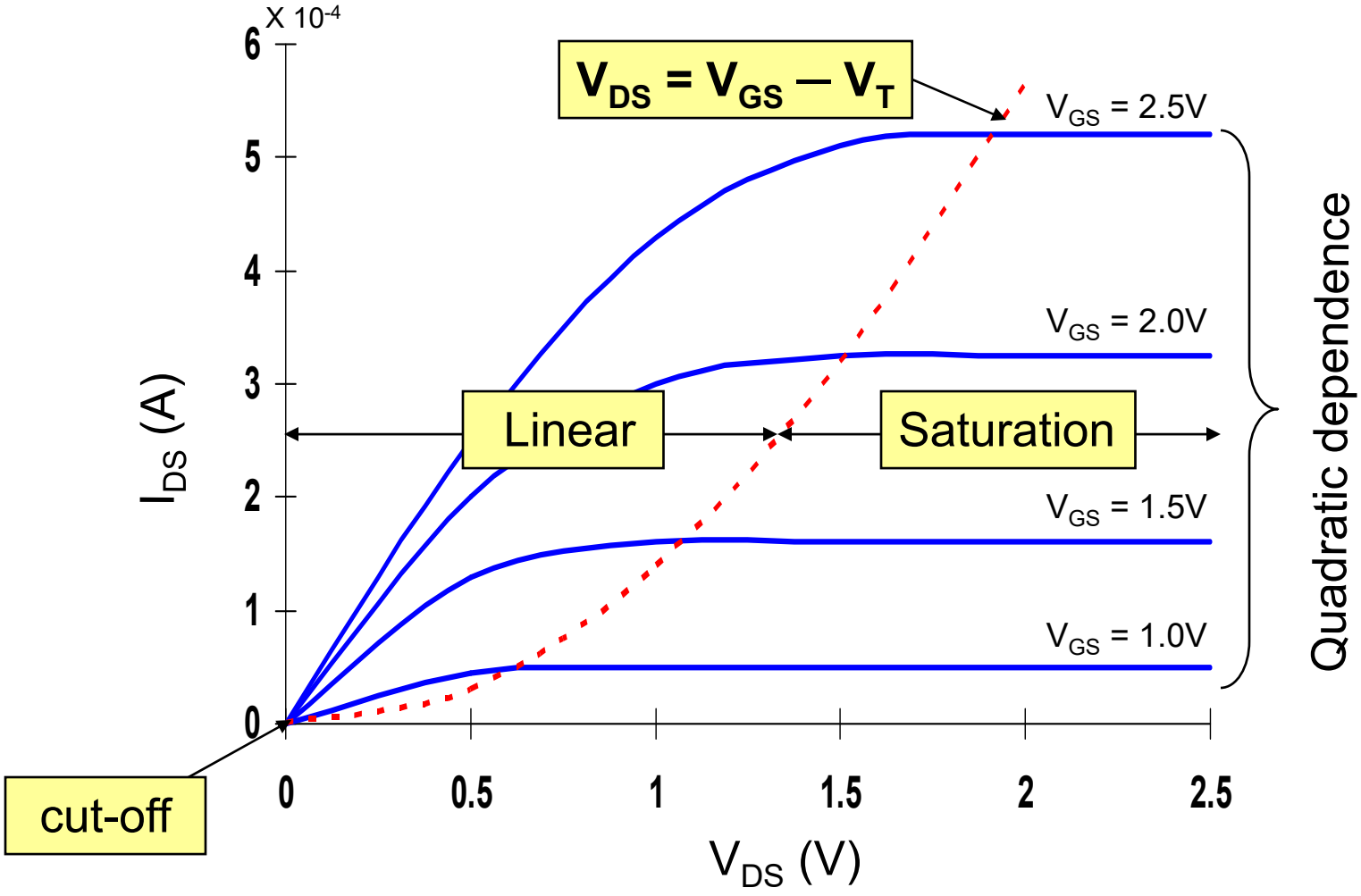


- Two geometric parameters of the MOS:
  - L is the channel length
  - W is the width of the transistor

I-V Characteristics (NMOS)



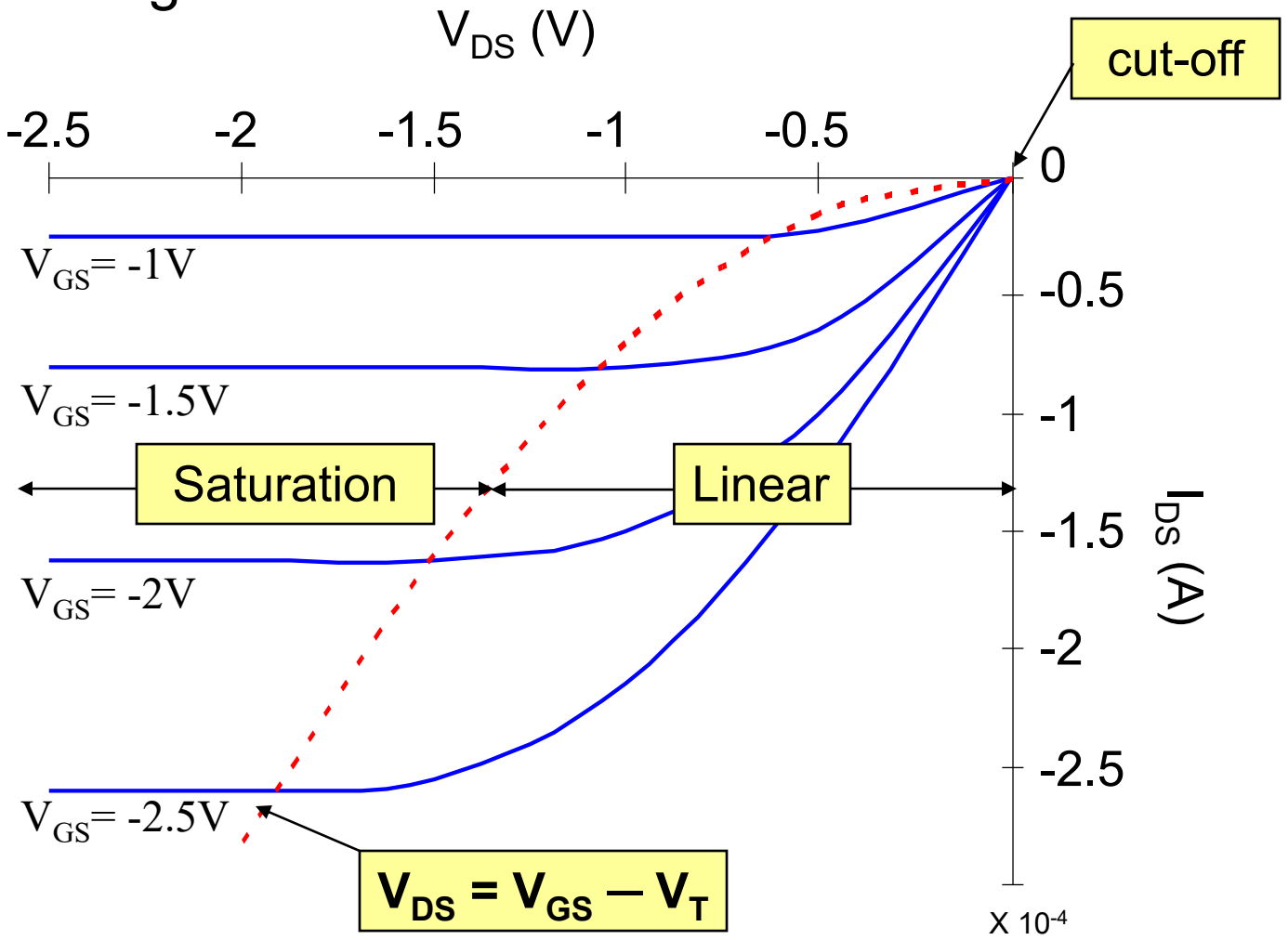
I-V Plot (NMOS)



NMOS transistor, 0.25µm, W/L = 1.5, V_{DD} = 2.5V, V_T = 0.4V

I-V Plot (PMOS)

- Same curves than NMOS but the polarities of all voltages and currents are reversed: I_{DS}, V_{DS}, V_{GS} and V_T are all negative.



Linear Mode (NMOS)

- When $V_{DS} \leq V_{GS} - V_T$ ➡ Linear mode:

$$I_{DS} = k'_n \, W/L \, [(V_{GS} - V_T)V_{DS} - V_{DS}^2/2]$$

- For small V_{DS}, there is a linear dependence between V_{DS} and I_D, hence the name resistive or linear region.

- $k'_n = \mu_n C_{ox} = \mu_n \varepsilon_{ox}/t_{ox}$ is the *transconductance.*

- C_{ox} is the gate oxide capacitance, t_{ox} is the thickness of the oxide (SiO_2) and ε_{ox} is the oxide permittivity.

- μ_n is the carrier mobility ($m^2/V.s$);

Saturation Mode (NMOS)

- When $V_{DS} \geq V_{GS} - V_T$ ➡ Saturation mode:

$$I_{DSAT} = k'_n/2\ W/L\ [(V_{GS} - V_T)^2]$$

- The current remains constant or "saturates" at I_{DSAT}.

- Because of a phenomenon called **channel-length modulation,** the current I_{DSAT} is not actually constant but exhibits a slight linear dependence on V_{DS}. This effect can be analytically accounted for by incorporating a factor λ independent of V_{DS} :

$$I_{DS} = I_{DSAT}(1 + \lambda V_{DS})$$

where λ is given by the technology. Typical values range from 0.005 to 0.03 V^{-1} .

Current Determinates

- For a fixed V_{DS} and V_{GS}, I_{DS} is a function of:

  the distance between the source and drain L

  the channel width W

  the threshold voltage V_T

  the thickness of the SiO_2 t_{ox}

  the dielectric of the gate insulator (SiO_2)

  oxide permittivity $\varepsilon_{ox}=3.97\times\varepsilon_o=3.5\times10^{-11}\mathrm{F/m}$

  the carrier mobility

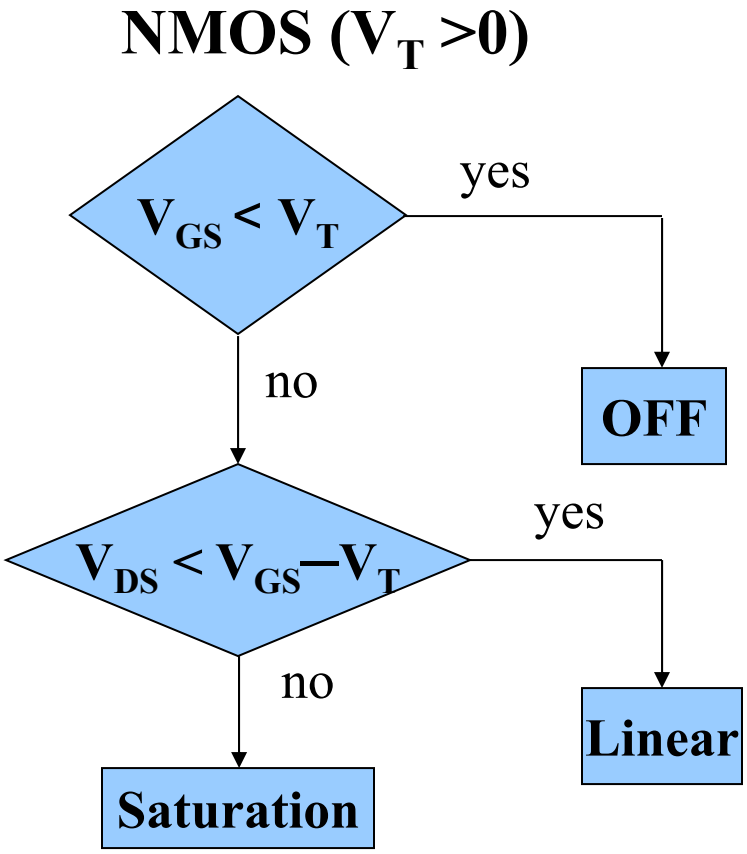
  - for NMOS: μ_n = 500 $cm^2/V.s$
  - for PMOS: μ_p = 180 $cm^2/V.s$

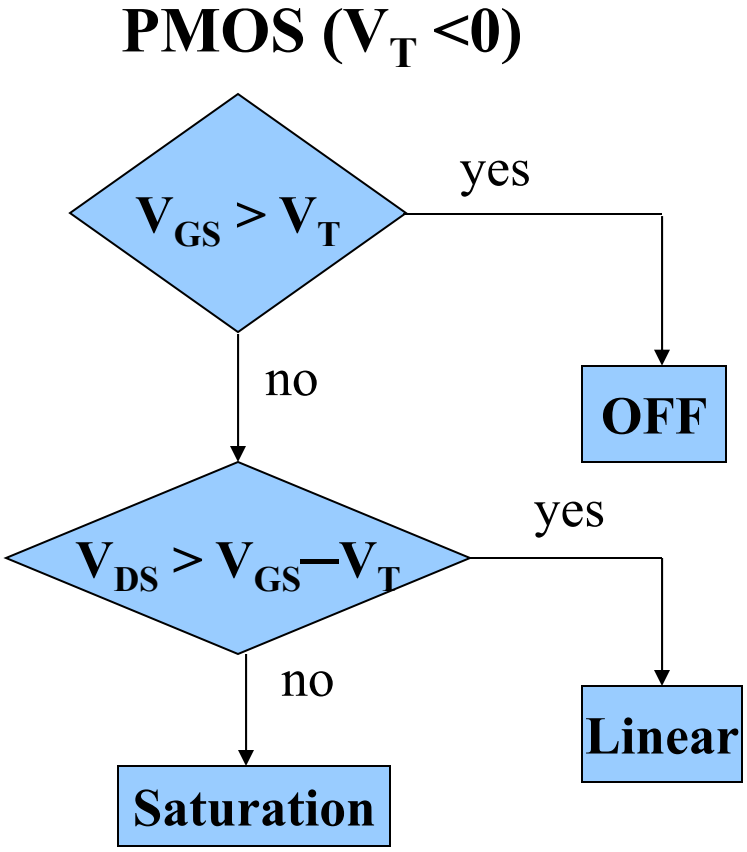
SUMMARY OF MOS EQUATIONS

NMOS Transistor	PMOS Transistor
$K_n' = \mu_n C_{ox} = \mu_n \frac{\varepsilon_{ox}}{t_{ox}}$	$K_p' = \mu_p C_{ox} = \mu_p \frac{\varepsilon_{ox}}{t_{ox}}$
The cutoff region ($V_{GS}<V_t$): $I_{DS}=0$	The cutoff region ($V_{GS}>V_t$): $I_{DS}=0$
The linear region ($0<V_{DS}<V_{GS}-V_t$) $I_{DS} = K_n' \frac{W}{L}[(V_{GS}-V_t)V_{DS} - \frac{V_{DS}^2}{2}]$	The linear region ($0>V_{DS}>V_{GS}-V_t$) $I_{DS} = -K_p' \frac{W}{L}[(V_{GS}-V_t)V_{DS} - \frac{V_{DS}^2}{2}]$
The saturation region ($0< V_{GS}-V_t <V_{DS}$) $I_{DS} = \frac{K_n'}{2}\frac{W}{L}(V_{GS}-V_t)^2(1+\lambda V_{DS})$	The saturation region ($0> V_{GS}-V_t >V_{DS}$) $I_{DS} = -\frac{K_p'}{2}\frac{W}{L}(V_{GS}-V_t)^2(1+\lambda V_{DS})$

- For NMOS: V_{DS}, V_{GS}, V_T, λ are all positive
- For PMOS: V_{DS}, V_{GS}, V_T, λ are all negative
- I_{DS} refer to the current flowing through the drain to the source.

Operating mode of the transistor - Summary





Example

- Various NMOS and PMOS transistors are measured in operation, as shown in the table below. For each transistor, find the value of $\mu C_{ox} W/L$ and V_t that apply and complete the table with V in Volts, I in μA, and $\mu C_{ox} W/L$ in $\mu A/V^2$.

Case	Transistor	V_S	V_G	V_D	I_D	Type	Mode	$\mu C_{ox} W/L$	V_t
a	1	0	2	5	100				
	1	0	3	5	400				
b	2	5	3	-4.5	50				
	2	5	2	-0.5	450				
c	3	5	3	4	200				
	3	5	2	0	800				
d	4	-2	0	0	72				
	4	-4	0	-3	270				

The Threshold Voltage

where $V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|})$

V_{T0} is the threshold voltage at $V_{SB} = 0$ and is mostly a function of the manufacturing process

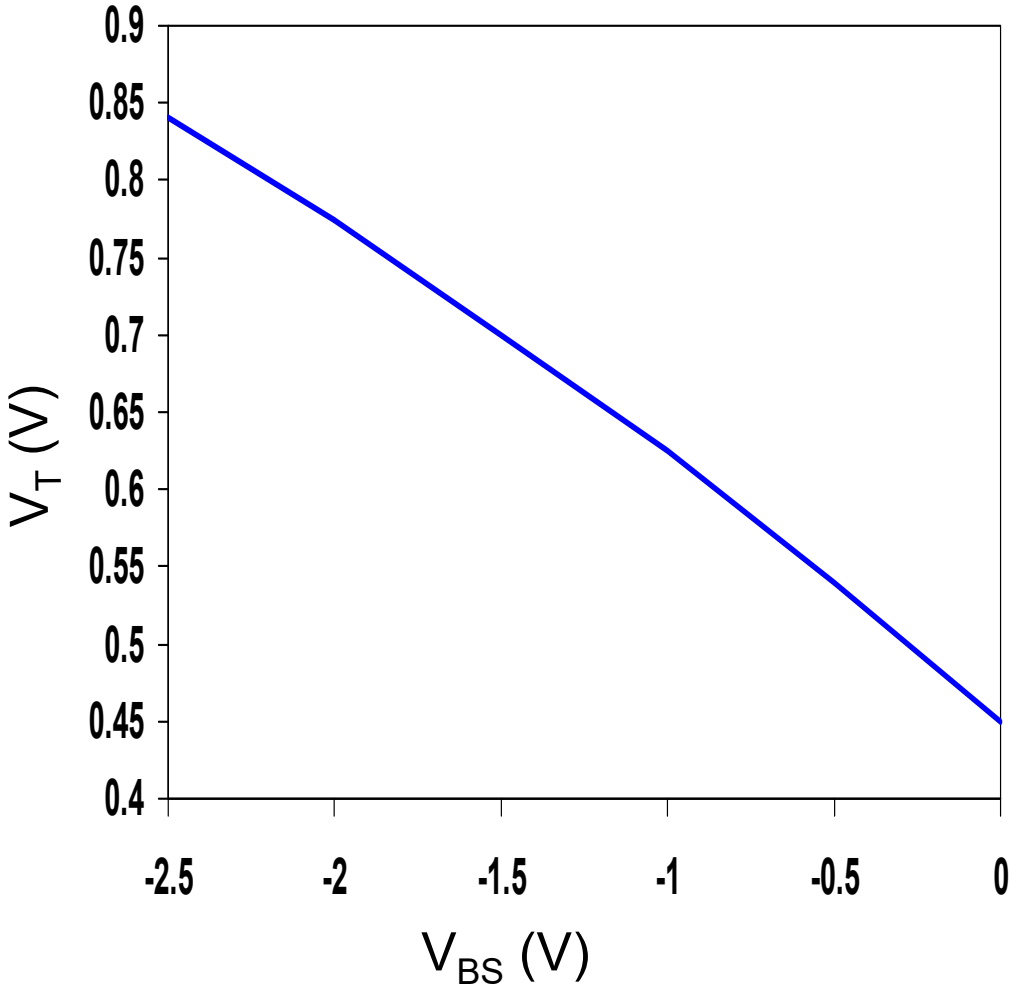
Difference in work-function between gate and substrate material, oxide thickness, Fermi voltage, charge of impurities trapped at the surface, dosage of implanted ions, etc.

V_{SB} is the source-bulk voltage

$\phi_F = -\phi_T \ln(N_A/n_i)$ is the Fermi potential ($\phi_T = kT/q = 26mV$ at 300K is the thermal voltage; N_A is the acceptor ion concentration; $n_i \approx 1.5 \times 10^{10}$ cm^{-3} at 300K is the intrinsic carrier concentration in pure silicon)

$\gamma = \sqrt{(2q\varepsilon_{si}N_A)}/C_{ox}$ is the body-effect coefficient (impact of changes in V_{SB}) ($\varepsilon_{si}=1.053 \times 10^{-10}$F/m is the permittivity of silicon; $C_{ox} = \varepsilon_{ox}/t_{ox}$ is the gate oxide capacitance with $\varepsilon_{ox}=3.5 \times 10^{-11}$F/m)

The Body Effect



V_{SB} is the substrate bias voltage (normally positive for n-channel devices with the body tied to ground)

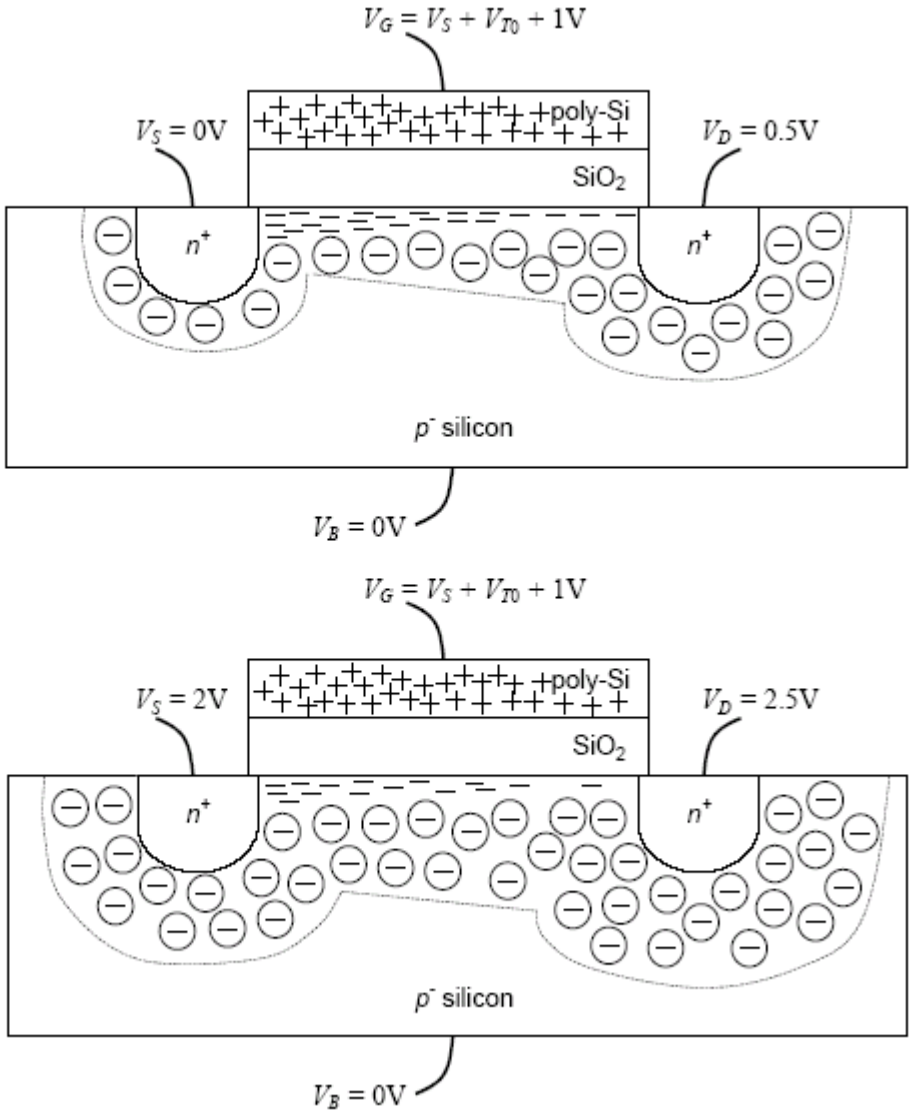
A negative bias causes V_T to increase from 0.45V to 0.85V

The Body Effect

What happens if the source is not at the same potential as the substrate?

The figure shows the situation as we keep V_{DS} constant but raise V_S>0.

Now that the source and drain are more strongly reverse-biased with respect to the substrate, the depletion region widens and the depletion charge increases. Now that there is more channel charge from the depletion region, there need not be as must inversion layer charge to balance the gate charge. Thus, the inversion layer weakens and the current drops.

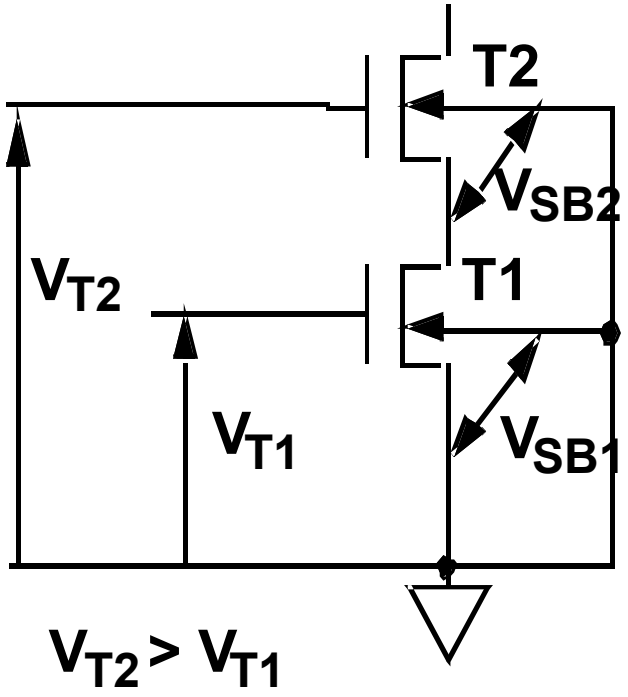


Body Effect

- In normal situation, the substrate voltage of all devices is normally equal. However, in some arrangement of the devices to form circuit functions it might be necessary to connect several devices in series.

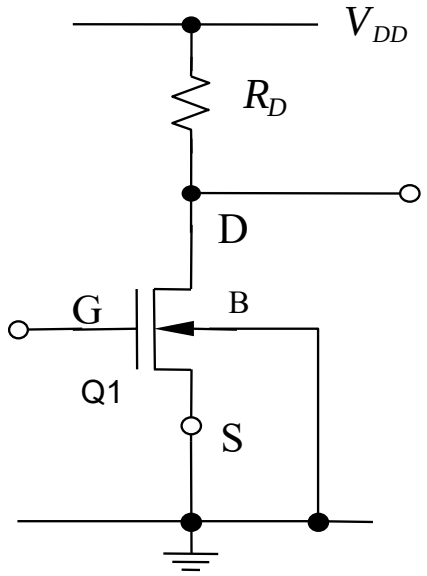
- This may result in an increase in source-to-substrate voltage as we proceed vertically along the series chain (V_{SB1} =0, $V_{SB2} \neq 0$). The overall result is ($V_{T2} > V_{T1}$).

- Every stacked transistor will have slightly higher V_T. That is every stacked transistors will be increasingly harder to turn on.



MOSFET as a switch - example

Low power: digital logic operates a MOS FET in either the cut-off or the ohmic region. In both regions the product of the source current and drain to source voltage can be made very small.



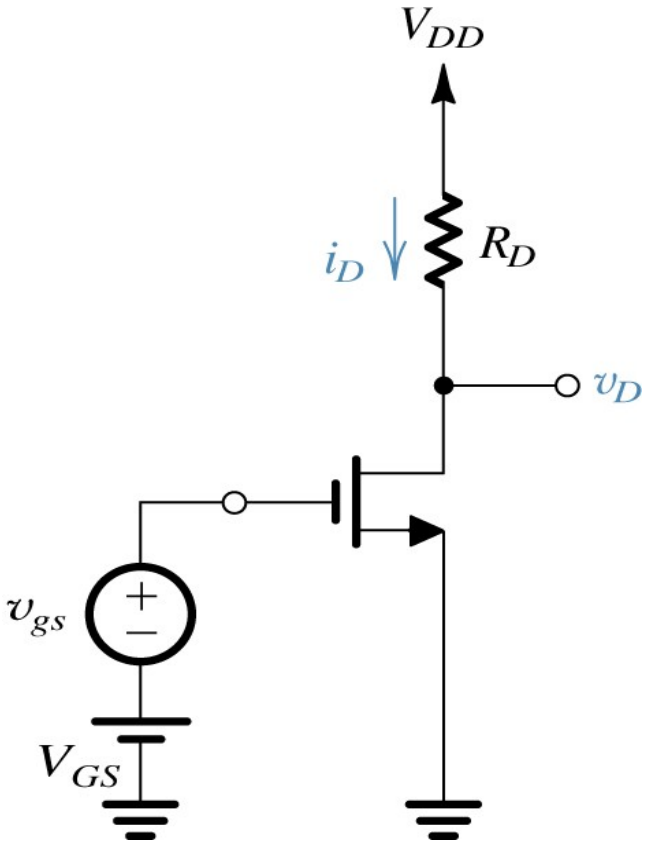
Input = V_{DD} Output = ?	~ 0
Input = 0 Output = ?	~ V_{DD}

Switch ON: Gate to Source potential difference positive a channel exists and current can pass between source and drain

Switch OFF: Gate to Source potential difference zero a channel does not exist and current cannot pass between source and drain

MOSFET as an amplifier - example

- Small signal "linear amplifiers" usually operate a **MOSFET in the saturation region.**



$$
\begin{aligned}
v_D &= V_{DD} - R_D i_D \\
v_D &= V_{DD} - R_D(I_D + i_d) \\
v_D &= V_D - R_D i_d \\
v_d &= -R_D i_d = -g_m R_D v_{gs} \\
\frac{v_d}{v_{gs}} &= -g_m R_D
\end{aligned}
$$

Acknowledgments

- Lecture slides are mainly based on the lecture materials from F. Bousaid (UWA).
- Credit is acknowledged where credit is due. Please refer to the full list of references.