Lecture #2
The MOS Transistor

Suppose we build a parallel-plate capacitor where one plate is metal, one plate is a semiconductor (e.g., weakly-doped silicon), and the insulator is SiO2. Such a device is called a MOS (metal-oxide-semiconductor) capacitor. The metal plate is called the **gate,** and is not always built out of metal. Nowadays, gates are made from heavily-doped polycrystalline silicon (or "polysilicon," or just "poly"). Polysilicon does not have a rigid crystal lattice, and conducts current freely, acting almost like a metal. The MOS capacitor has several distinct regions of operation. For our example, we will consider MOS capacitors built with p - silicon. With no external voltage placed across it, the MOS capacitor is in the **flat band** region (ignoring work functions and implanted fixed charge). If we lower the voltage on the gate (by introducing negative charges), we attract "extra" holes (above the background level of holes in the p - silicon), which **accumulate** on the surface of the silicon. This creates a thin p + region near the silicon-oxide interface (Figure 1(b)).

In the accumulation regime, MOS capacitors act as linear capacitors since the p + region acts as a highly conductive "bottom plate." If instead we raise the gate voltage (by introducing positive charges), we "scare away" holes in the p - silicon. This **depletes** the surface of holes, creating a depletion region with exposed negative dopant ions (Figure 2(b)).

As we have seen, the depletion capacitance is nonlinear, so a MOS capacitor operating in this regime is not very linear. It can be modeled as two capacitors in series: a linear oxide capacitance and the nonlinear depletion capacitance. Thus, the total capacitance is **less** than the oxide capacitance. Due to contact potentials between the substrate and gate, MOS capacitors are typically in the depletion region when the gate and substrate are at the same potential. As we continue to raise the gate voltage, the depletion region cannot provide enough negative charges to match all the positive charges we are putting on the gate. Eventually, these positive charges begin to "pull" electrons from bonds, producing mobile electrons, which balance the charge. (The missing bond electrons are quickly replaced from below.)

The production of mobile electrons begins to **invert** the silicon at near the surface – it changes it from p-type silicon to n-type silicon! When the number of mobile electrons (the **inversion charg**e) is much lower than the number of exposed dopant ions in the depletion region (the **depletion charg**e), we are in the regime of **weak inversio**n (Figure 3(a)).

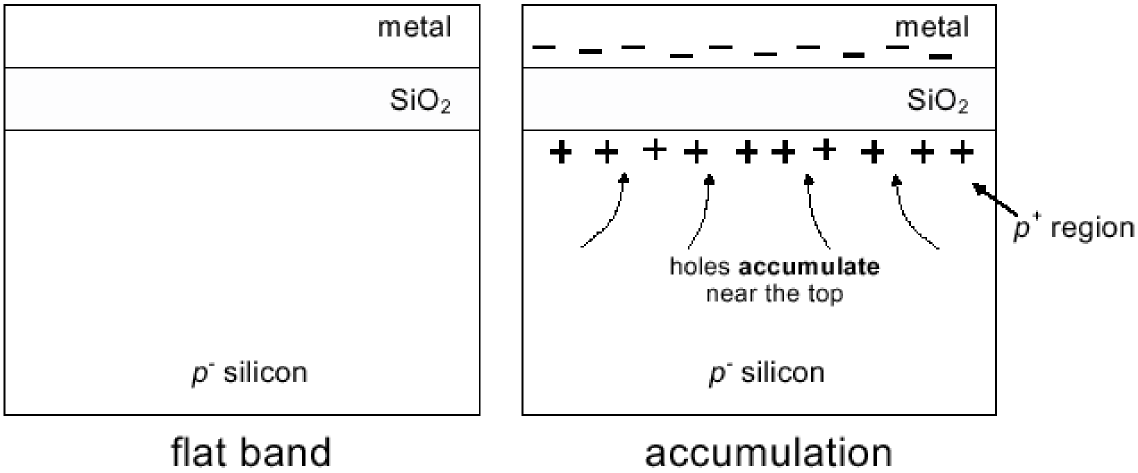

Figure 1. (a) Flat band, (b) Accumulation

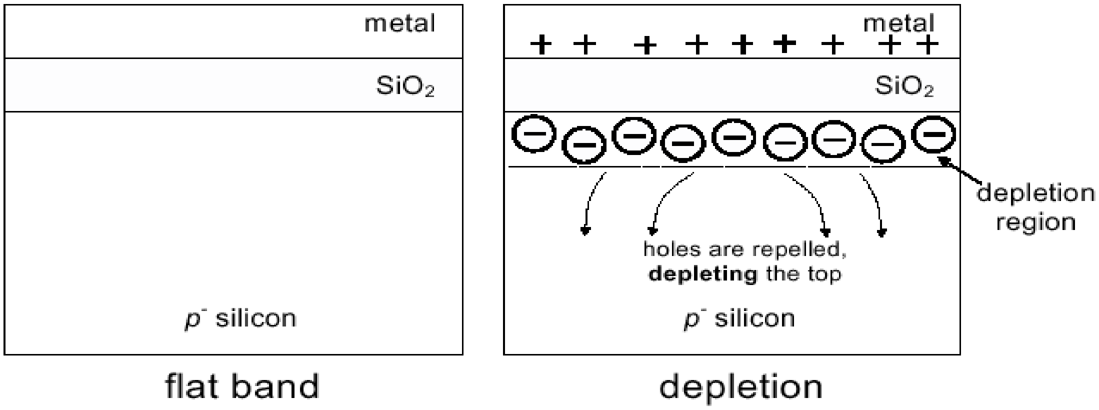

Figure 2. (a) Flat band and (b) Depletion

When the inversion charge greatly exceeds the depletion charge, we have **strong inversio**n, and a conductive n + layer forms at the surface of the semiconductor (Figure 3(b)). When the inversion charge and depletion charge are comparable, we are in a regime know as **moderate inversio**n (Figure 3(c)).

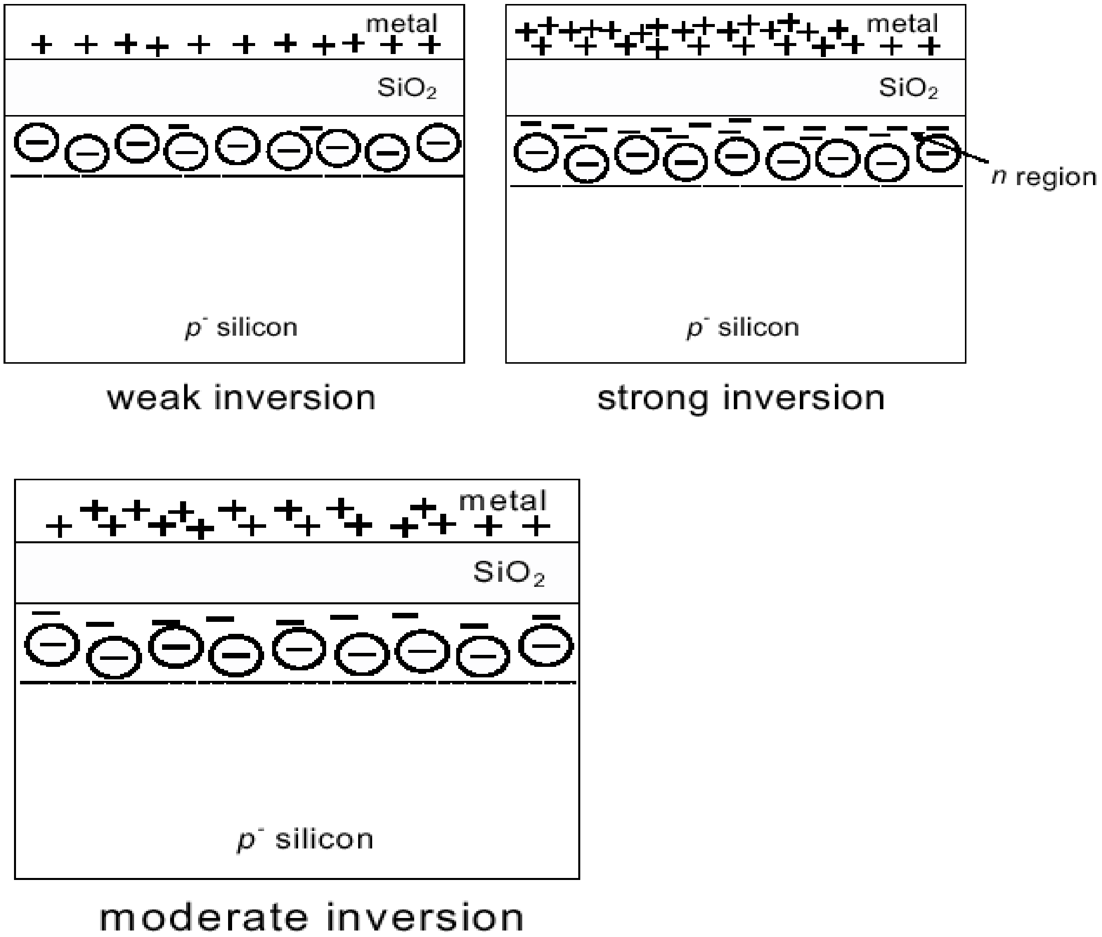

Figure 3. (a) weak inversion, (b) Strong inversion, (c) Moderate inversion

The MOS Transistor in Strong Inversion

In this section, we shall explore the behavior of the MOS transistor when the area under the gate – the **channel** – is strongly inverted. Strong inversion is shown in Figure 4 for a MOS capacitor. The gate-to-bulk voltage *VGB* can be decomposed into the potential across the oxide (ψ_{ox}) and the **surface potential** of the silicon substrate (ψ_s). Ignoring any implanted charge or contact potential effects; the charge on the gate (Q_G) must be balanced out by the charge in the channel. The channel charge consists of the fixed depletion or "bulk" charge *QB* and the mobile inversion layer charge Q_I:

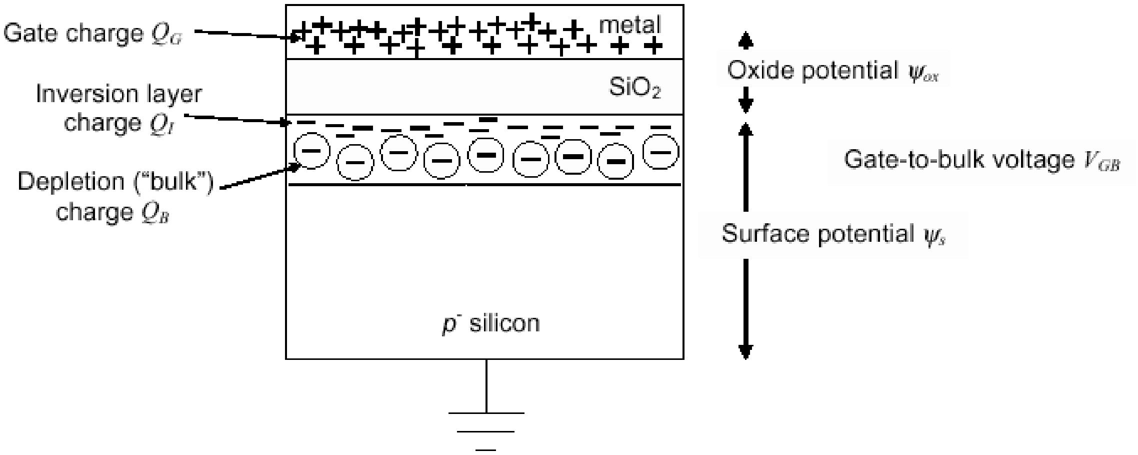
Gate charge Q_G
Inversion layer charge Q_I
Depletion ("bulk") charge Q_B
metal
SiO₂
p^- silicon
Oxide potential ψ_{ox}
Gate-to-bulk voltage V_{GB}
Surface potential ψ_s

Figure 4. Strong Inversion

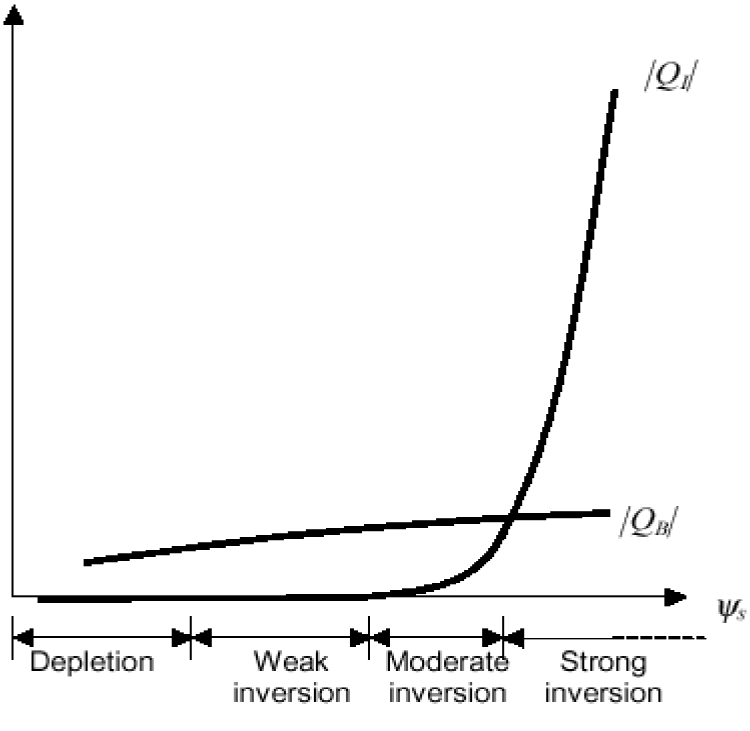

Figure 5. Charge variation with surface potential

In strong inversion, the gate charge is balanced out primarily by the inversion layer charge. The voltage at which inversion layer charge dominates is called the **threshold voltage V_T**. The symbol V*T0* will often be used, and this indicates the threshold voltage when the source voltage equals zero (more on this later).

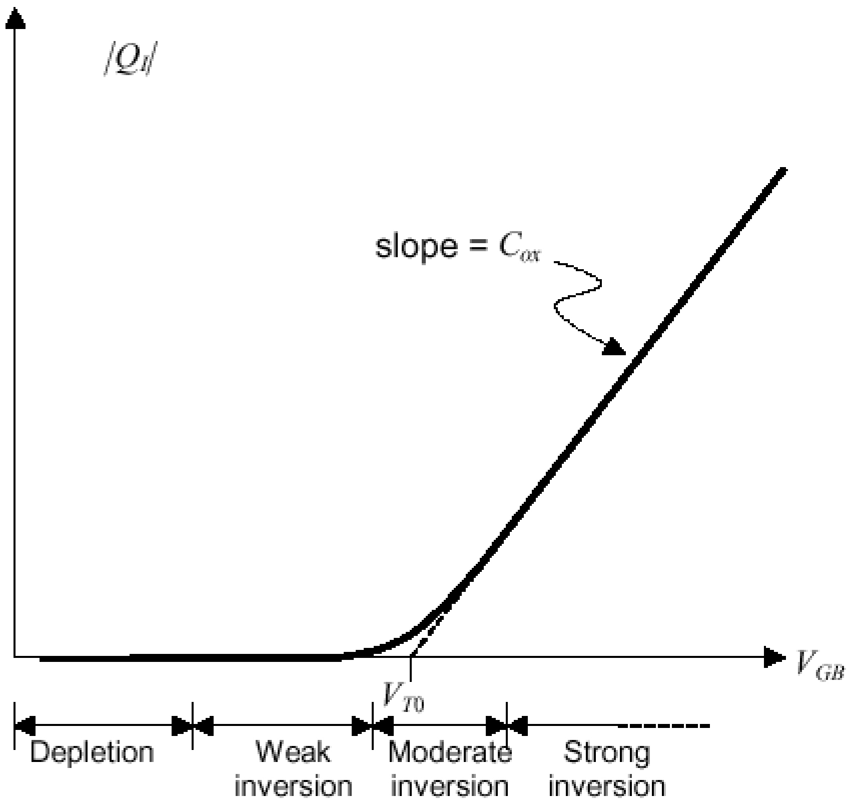

Figure 6. Variation of Inversion layer charge with gate voltage

Of course, to make a transistor we need more than a gate; we also need a source and a drain. Charge carriers flow through the channel (under the gate) from the source to the drain. An n-channel MOSFET ("nMOS transistor" or "nFET") is shown in Figure 7. Notice that an n-channel FET actually has a channel made of weakly-doped p-type silicon. However, when the transistor conducts current, this channel is inverted, and is thus n-type.

Electrons carry charge in nFETs (since the inverted channel is n-type). Since electrons flow from the source to the drain, current flows from the drain to the source. The channel has a width *W* and a length L. The width-to-length ratio ("*W/L* ratio") is an important parameter in MOSFET operation, as we shall see.

The complementary type of MOSFET is a p-channel MOSFET ("pMOS transistor" or "pFET"). This is shown in Figure 8. Although the channel is made from n-type silicon, it becomes p-type when inverted.

In a complementary-MOS (CMOS) process where both nFETs and pFETs can be built, one of the two devices must be built in a well whose doping is opposite that of the substrate. In most processes nowadays, p – substrates and n - wells are used, so that pFETs reside inside wells. The substrate is always set at the lowest potential used in the circuit (usually called ground or V_{SS}) so that the nFET source/drain regions stay reverse- biased. Likewise, the wells are usually set to the highest potential used in the circuit (usually called V_{DD}) so that the pFET source/drain regions stay reverse-biased.

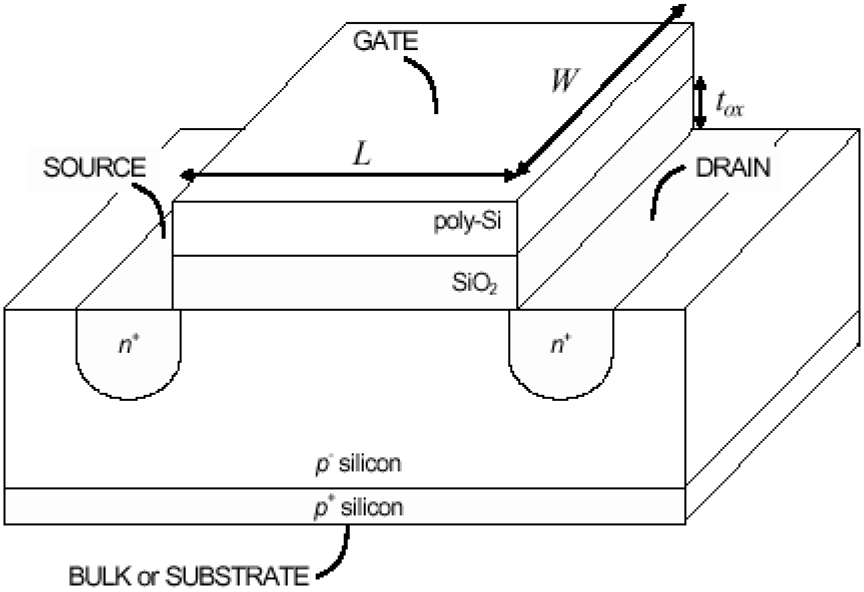

Figure 7. n-channel MOSFET (nMOS)

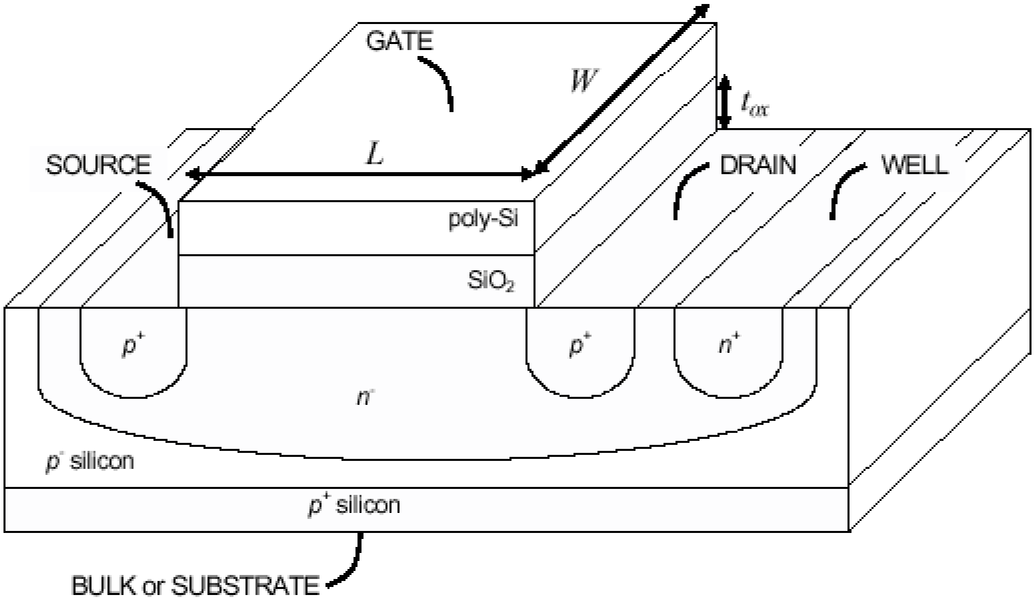

Figure 8. p-channel MOSFET (pMOS)

If we place a metal wire (aluminum is used in most VLSI processes) directly against a semiconductor, we get a rectifying contact – a **Schottky barrier diod**e. This is obviously not what we want when we tie a wire to our transistor. One way to avoid this diode is to connect metal only to heavily- doped (n + or p +) regions. This forms a nonrectifying **ohmic contact** with a typical resistance of a few tens of ohms. This explains why we connect to the well via a n + **well contact** region and connect to the substrate via a p + **substrate contact** region.

Now let's take a look at an nFET with a gate voltage below V_T (subtheshold). The only charge in the channel is due to the depletion region, which also extends around the source and drain p-n junctions as shown in Figure 9. (For now, we will neglect the small number of mobile electrons present in this condition.)

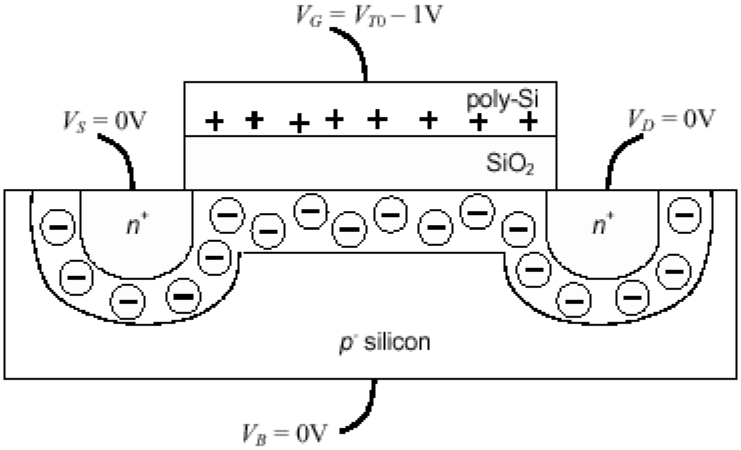

Figure 9. Gate voltage $V_G > V_T$

Now, if we raise the gate voltage above the threshold voltage V_T, we create an inversion layer of mobile electrons that balance out most of the gate charges as shown in Figure 10.

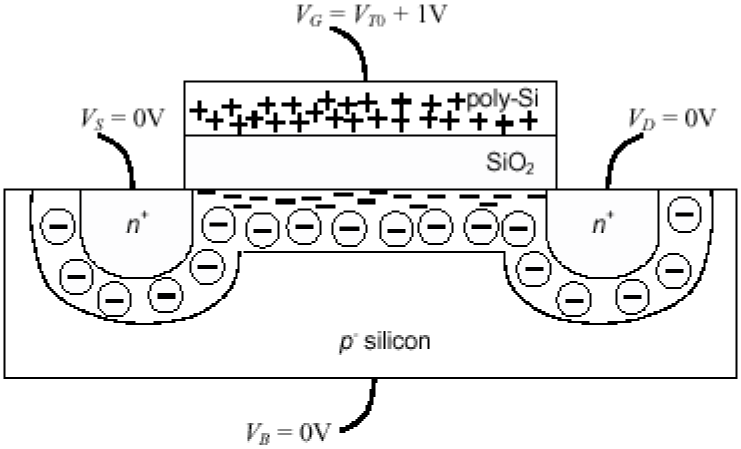

Figure 10. Gate voltage $V_G > V_T$

The charge in the inversion layer is proportional to the **effective voltage V_{eff}.**

$V_{eff} = V_{GS} - V_T$, and $Q_I' = -C_{OX}' V_{eff}$ in strong inversion.

Now let's return to our cross-section of an actual nFET and observe the channel in strong inversion when V_D>V_S (Figure 11). Notice that the depletion layer widens towards the drain since the drain p-n junction is more strongly reverse-biased than the source p-n junction. More importantly, the inversion layer contains more charge towards the source. Why is this? Near the source, the channel potential (surface potential) V_C is approximately equal to V_S. Near the drain, the channel potential is approximately equal to V_D, Since V_S is lower than V_D, the voltage across the oxide is greater near the source, and charge is proportional to voltage in a parallel-plate capacitor. The drain current can be written as:

$$I_D = \mu_n C'_{ox} \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right]$$

This is the familiar MOS transistor equation valid in strong inversion $V_{GS} > V_T$.

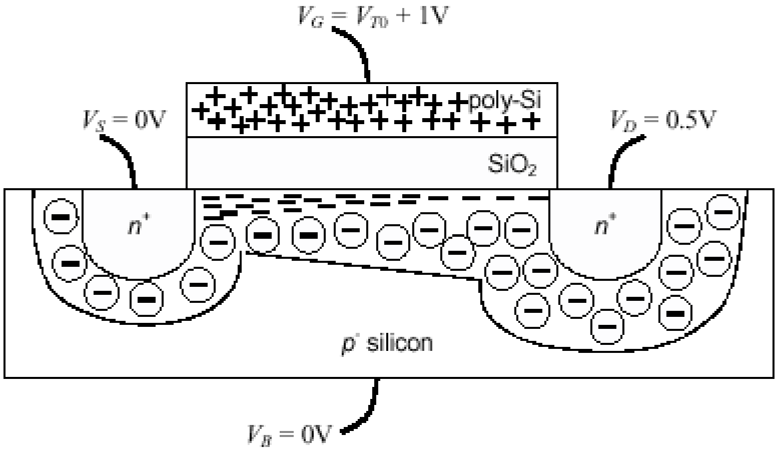

Figure 11. Channel in strong inversion

For small V_{DS}, that is, $V_{DS} << 2(V_{GS} – V_T)$, the V_{DS}^2 term is very small, and can make the following approximation:

$$I_D \approx \mu_n C'_{ox} \frac{W}{L}(V_{GS} - V_T)V_{DS} \qquad \approx \frac{V_{DS}}{R_{on}}$$

where

$$R_{on} = \frac{1}{\mu_n C'_{ox} \frac{W}{L}(V_{GS} - V_T)}$$

In other words, for small drain-to-source voltages, the MOSFET behaves like a resistor. The resistor's value decreases as we increase the gate voltage. This is not at all surprising since that increases the amount of inversion charge. Now let's return to the original MOSFET equation:

$$I_D = \mu_n C'_{ox} \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right]$$

When $V_{DS} = V_{GS} – V_T$, I_D is maximum:

$$I_{D,\max} = \frac{1}{2}\mu_n C'_{ox} \frac{W}{L}(V_{GS} - V_T)^2$$

Here, Q_I goes to zero at the end of the channel since $V_D = V_{eff}$. Wait a minute: How can the transistor conduct if there is zero charge at the end of the channel? Well, the charge doesn't go exactly to zero, but it is quite small, **which means the charge is moving extremely fast – the electric field is very strong here**. In fact, this condition in transistors is called **pinch of**f.

What happens if we increase V_D further? When $V_D > V_{eff}$, do positive charges accumulate in the channel? No. In fact, very little change occurs beyond this point. The drain becomes "disconnected" from the channel, meaning that changes in drain voltage no longer affect channel current (to first order). So the equation we derived for MOSFET operation is only valid up to this point. After this point, the current is "frozen" at that maximum value; it does not follow the parabola described by the equation. So we can describe the current flowing into the drain of an nMOS transistor using two equations (plus one useful approximation):

$$I_D = \mu_n C'_{ox}\frac{W}{L}\left[(V_{GS}-V_T)V_{DS}-\frac{1}{2}V_{DS}^2\right] \qquad V_{DS} < V_{GS}-V_T \ \ (\textbf{triode}\text{ region})$$

$$I_D \approx \mu_n C'_{ox}\frac{W}{L}(V_{GS}-V_T)V_{DS} \qquad V_{DS} << V_{GS}-V_T \text{ (deep triode region)}$$

$$I_D = \frac{1}{2}\mu_n C'_{ox}\frac{W}{L}(V_{GS}-V_T)^2 \qquad V_{DS} \geq V_{GS}-V_T \ \ (\textbf{saturation}\text{ region})$$

For pMOS transistors, we just use the hole mobility instead of electron mobility, and insert a minus sign since the charge carriers are now holes (or you can just think of the current as flowing out of the drain instead of into it). Also, pFETs have negative threshold voltages (but V_{GS} will also be negative in normal operation).

$$I_D = -\mu_p C'_{ox}\frac{W}{L}\left[(V_{GS}-V_T)V_{DS}-\frac{1}{2}V_{DS}^2\right] \qquad |V_{DS}| < |V_{GS}-V_T| \ \ (\textbf{triode}\text{ region})$$

$$I_D \approx -\mu_p C'_{ox}\frac{W}{L}(V_{GS}-V_T)V_{DS} \qquad |V_{DS}| << |V_{GS}-V_T| \text{(deep triode region)}$$

$$I_D = -\frac{1}{2}\mu_p C'_{ox}\frac{W}{L}(V_{GS}-V_T)^2 \qquad |V_{DS}| \geq |V_{GS}-V_T| \ \ (\textbf{saturation}\text{ region})$$

This behavior is shown graphically in Figure 12. When a MOSFET has a low voltage across it (low V_{DS}), it acts like a voltage-controlled resistor. The resistance is inversely proportional to V_{eff}. When a MOSFET has a sufficiently high voltage across it (high V_{DS}), it acts like a voltage-controlled current source. The current is proportional to the square of V_{eff}.

Why does the current depend on the square of the effective gate voltage? Because when we increase the gate-to-source voltage, we affect the current in two ways: we increase the channel charge and we increase the electric field in the channel. Notice that much of our supply voltage can be used up just for keeping MOSFETs in the active region!

Channel-Length Modulation

When we discussed pinch-off and the saturation region, we said that after pinch-off occurred, the drain current was constant. Well, that's not quite true due to an effect called **channel-length modulatio**n. As we increase the drain voltage in the saturation region, the drain depletion region widens. This has the effect of widening the pinch-off region and thus shrinking the channel by a small amount (Figure 13).

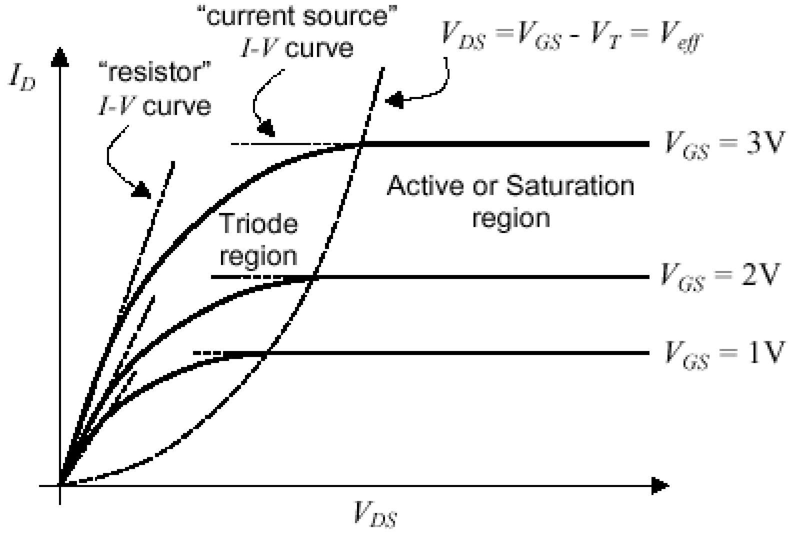

Figure 12. $V_{DS} - I_D$ characteristics of nMOS

Since the current in a MOSFET is proportional to *W/*L, a shrinking channel length increases the current through the device. In "long-channel" devices (which usually means at least 3 times the minimum length allowed by the process), the slight increase in current as the drain voltage increases is nearly linear. If we plot I_D vs V_{DS} for several values of V_{GS} and extrapolate all the current traces backwards, they tend to converge at the same point on the V_{DS} axis. This voltage is called the **EARLY VOLTAGE** (V_A), and is the characteristic voltage in the first-order model of channel length modulation.

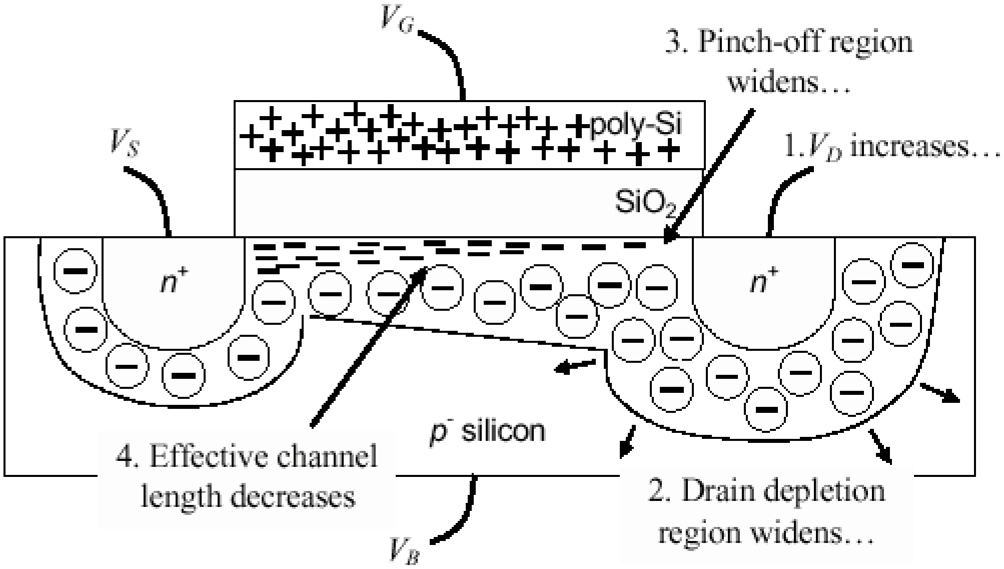
V_G
3. Pinch-off region widens…
poly-Si
V_S
1.V_D increases…
SiO$_2$
n^+
n^+
p^- silicon
4. Effective channel length decreases
2. Drain depletion region widens…
V_B

Figure 13. $V_{DS} >> V_{GS} - V_T$

We can thus modify our expression for current in the saturation region by adding an extra term:

$$I_D = \frac{1}{2}\mu_n C'_{ox}\frac{W}{L}(V_{GS} - V_T)^2 \cdot \left(1 + \frac{V_{DS}}{V_A}\right)$$

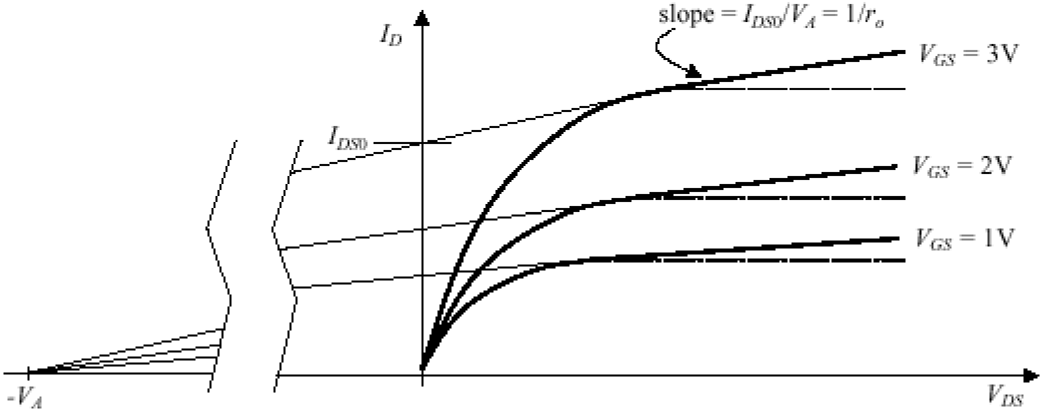

Figure 14. V_{DS} – I_D plot for long channel nMOS

Some people use the **channel length modulation coefficient** λ instead of V_A:

$$I_D = \frac{1}{2}\mu_n C'_{ox}\frac{W}{L}(V_{GS} - V_T)^2 \cdot (1 + \lambda V_{DS})$$

where $\lambda = 1/V_A$.

The circuit symbols shown in Figure 15 are commonly used to represent nMOS and pMOS transistors in circuit diagrams. In this class we will use the first symbols, but the book uses the later symbols (with the arrows). The arrows make a rather artificial distinction between the source and drain in a completely symmetric device.

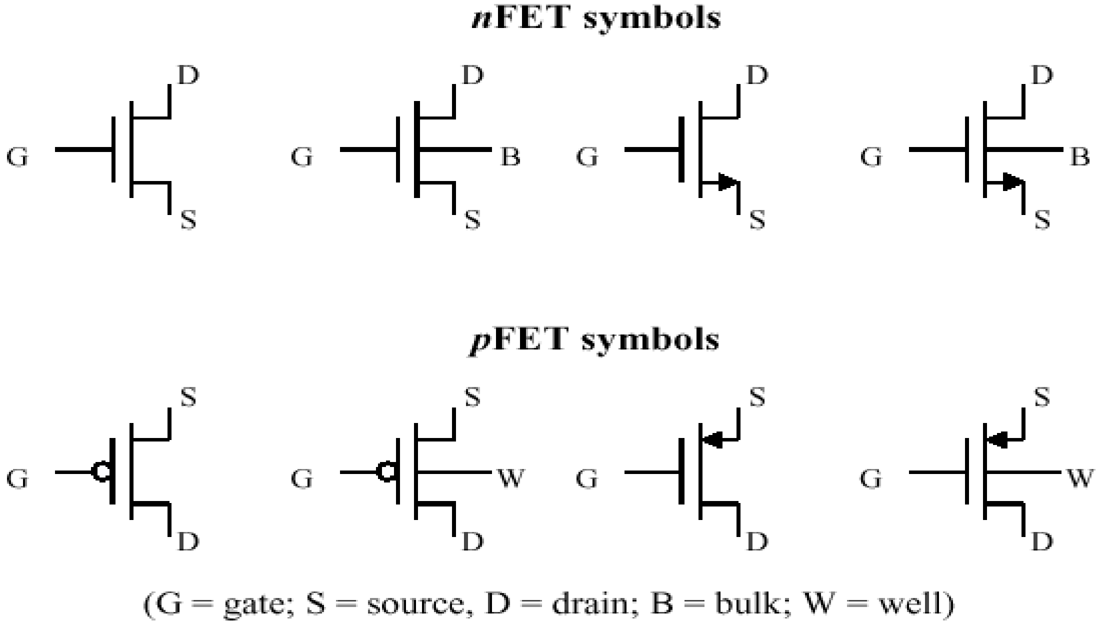

Figure 15. Circuit symbols for MOS transistors

References:

1. CMOS VLSI Design, Neil Weste and David Harris, Addison Wesley, 2004.
2. Digital Integrated Circuits, Jan M. Rabaey et. al., Prentice Hall, 2003.
3. Digital Integrated Circuit Design, Ken Martin, Oxford University Press, 2000.
4. CMOS Digital Integrated Circuits, Sung-Mo Kang and Yusuf Leblebici, McGraw Hill, 2003.
5. Introduction to VLSI Circuits and Systems, Joh P. Uyemura, John Wiley & Sons, 2002.