ECE 255, MOSFET's

6 February 2017

MOSFET stands for metal-oxide-semiconductor field effect transistor. It is a highly important transistor since it is widely used in **digital circuits**.

- Its fabrication process is simpler than BJT, and is more predictable.
- It draws little current, and hence consumes less power.

Because of its simplicity of fabrication, digital transistors are reaching to be over 7 billions per chip now!

On the other hand, MOSFET are not as linear a device as BJT and hence, does not guarantee high-fidelity amplification. Nevertheless, MOSFET is widely used in digital circuits, and some analogue circuits. When digital circuits are mixed with **analog circuits**, they are known as **mixed signal circuits**.

Printed on March 14, 2018 at 10 : 35: W.C. Chew and S.K. Gupta.

1 Device Structure and Physics

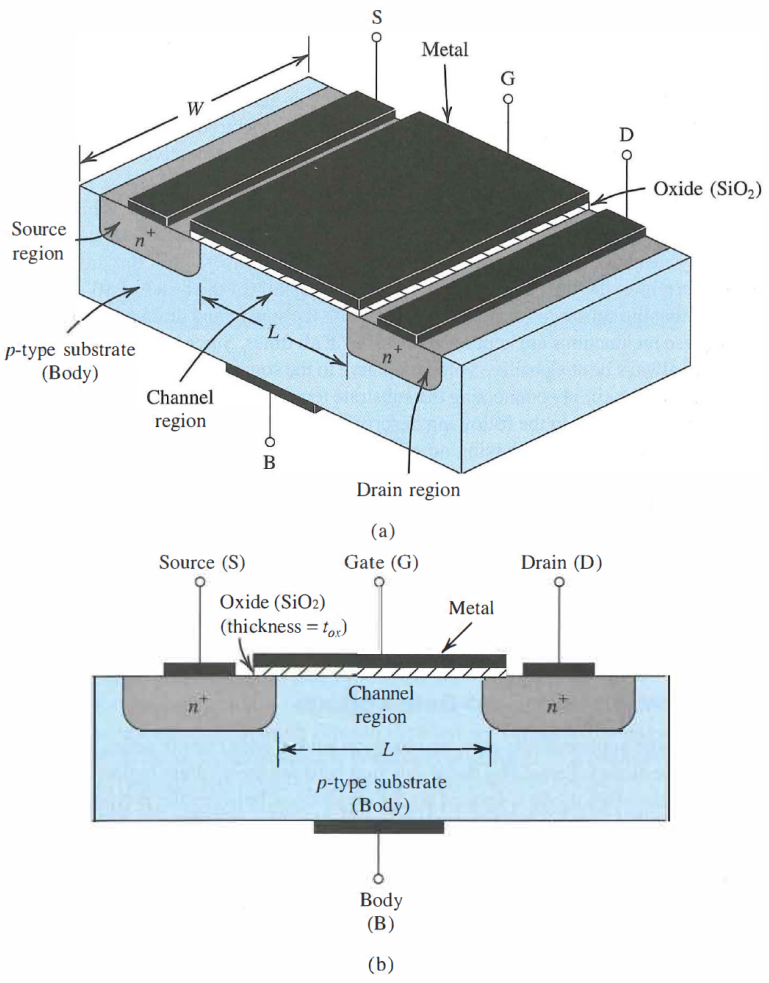

Figure 1: Physical structure of an enhancement-type MOSFET. Typically, $L = 0.03\,\mu$m to 1 μm, $W = 0.05\,\mu$m to 100 μm, and the thickness of the oxide layer (t_{ox}) is on the order of 1 to 10 nm (Courtesy of Sedra and Smith).

The physical structure of a MOSFET is shown in Figure 1. In general, it consists of two heavily doped n^+ **source** (S) and **drain** (D) regions. A p region separates the source and drain regions. A current flows from the source to the drain, via the **channel** region, and its flow is controlled by an applied voltage at the **gate** (G). The voltage at G creates a voltage drop between the gate (G) and the **body** (B) electrodes creating an internal electric field.

There are mainly two types of MOSFET.

- **enhancement-type MOSFET**. These MOSFET are off in the quiescent state, but are turned on by a gate voltage.
- **depletion-type MOSFET**. They are on in the quiescent state, but are turned off by a gate voltage

MOSFET are also called **IGFET** where IG stands for insulated gate, since the gate is separated from the substrate by an insulating oxide layer.

1.1 Creating a Conducting Channel

The source and the drain n^+ regions, interfacing with the p region, form two pn junctions with depletion zones. They are like two diodes in series opposition, allowing no current flow from the source to the drain when a voltage is applied. This holds true when no gate voltage is applied.

When a gate voltage is applied, as shown in Figure 2, the gate voltage and the internal electric field generated push away the holes from the channel region creating a depletion region. In the first instance, the electron carriers from the source and the drain regions that are attracted to the p region fill the vacant bonds of the region, and the electrons fall into the valence band. The p region or the depletion region becomes negatively charged or is said to be "uncovered".

If even higher gate voltage is applied, then carriers from the source and drain regions, with plentiful supply of electrons, are attracted to the channel region. These electrons are in the conduction band, and hence, can conduct electric current when a voltage is applied between the source and the drain. These "extra" electrons make the channel look like an n region with mobile electrons around. Such a MOSFET is also called an **n-channel MOSFET**, or an **NMOS** transistor. Since the conducting channel is obtained by converting a p region into an n region, the induced channel is also called an **inversion layer**.

The voltage of v_{GS} at which a p region becomes an n region is called the **threshold voltage**, V_t. The device can be fabricated such that this threshold voltage is 0.3 V to 1.0 V. Because the conducting channel is induced by the vertical electric field developed between gate and body electrodes, hence the name field-effect transistor.

Since the voltage has to be above V_t, the threshold voltage, to generate the conduction electrons, the excess voltage is

$$v_{OV} = v_{GS} - V_t \tag{1.1}$$

This is also called the **overdrive voltage** or the **effective voltage**. It is this effective voltage that gives rise to free carriers in the channel or **inversion layer**, making it behave like a conductor.

When the channel behaves like a conductor, then the channel together with the gate form a parallel plate capacitor. The charge created in the channel is, according to $Q = CV$ formula, given by

$$|Q| = C_g v_{OV} \tag{1.2}$$

where the gate capacitance is given by

$$C_g = C_{ox} W L \tag{1.3}$$

Here, C_{ox} is the capacitance per unit area, and WL is the area of the gate electrode. Using the familiar formula for parallel plate capacitance that

$$C = \frac{\epsilon A}{d} \tag{1.4}$$

where A is its area of one of the plates, and d is the separation of the plates, then the per unit area capacitance of the oxide layer is

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{1.5}$$

Typically, $\epsilon_{ox} = 3.9\epsilon_0$, and $t_{ox} = 4$ nm. Then

$$C_g = C_{ox} W L = 1.1\,\text{fF} \tag{1.6}$$

2 Small v_{DS} Model for MOSFET

When the applied voltage between the source and the drain is small, the current can be found approximately. The charge of the parallel plate capacitor between the gate and the channel is given by $Q = CV$, or in terms of the variables used in this place,

$$|Q| = C_{ox} W L v_{OV} \tag{2.1}$$

Then defining a charge per unit length,

$$Q_n = \frac{|Q|}{L} = C_{ox} W v_{OV} \tag{2.2}$$

where L is the channel length. Then the electric field along the channel is

$$|E| = \frac{v_{DS}}{L} \tag{2.3}$$

The drift velocity of the electron is given by

$$v_{\text{drift}} = \mu_n |E| = \mu_n \frac{v_{DS}}{L} \tag{2.4}$$

where μ_n is the mobility of the electrons at the surface of the channel.

The current is then charge per unit length times the drift velocity yielding

$$\begin{aligned} i_D = Q_n v_{\text{drift}} = C_{ox} W v_{OV} \mu_n \frac{v_{DS}}{L} &= \left[\mu_n C_{ox} \frac{W}{L} v_{OV} \right] v_{DS} \\ &= \left[\mu_n C_{ox} \frac{W}{L} \left(v_{GS} - V_t \right) \right] v_{DS} \end{aligned} \tag{2.5}$$

The conductance can be found as

$$g_{DS} = \mu_n C_{ox} \frac{W}{L} v_{OV} = \mu_n C_{ox} \frac{W}{L} (v_{GS} - V_t) \tag{2.6}$$

Therefore the conductance is a function of v_{OV}, or v_{GS}.

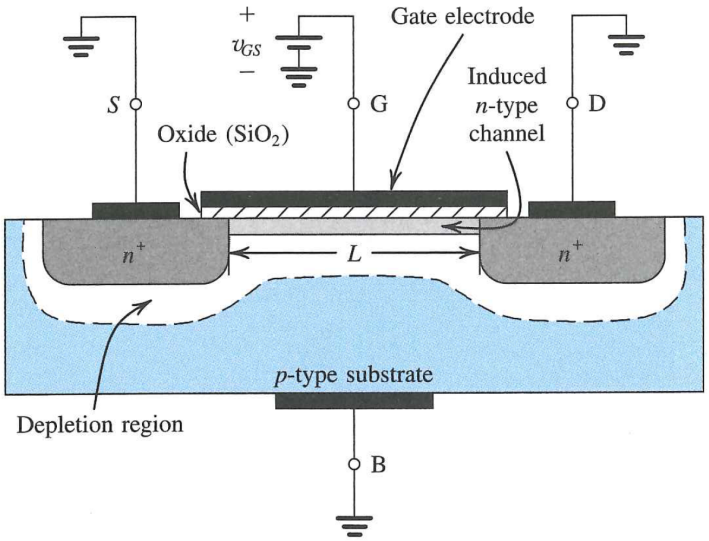

Figure 2: The enhancement-type MOSFET or NMOS with a positive gate voltage. Typically, $L = 0.03\,\mu$m to 1 μm, $W = 0.05\,\mu$m to 100 μm, and the thickness of the oxide layer (t_{ox}) is on the order of 1 to 10 nm (Courtesy of Sedra and Smith).

The factor

$$k'_n = \mu_n C_{ox} \tag{2.7}$$

is called the **process transconductance** which can be affected by manufacturing process. The second factor that affects the conductance is the aspect ratio $\frac{W}{L}$. For instance, for 14 nm technology, the channel length L cannot be smaller than 14 nm.

The factor

$$k_n = k'_n \frac{W}{L} = \mu_n C_{ox} \frac{W}{L} \tag{2.8}$$

is the **MOSFET transconductance parameter**. Moreover,

$$r_{DS} = \frac{1}{g_{DS}} = \frac{1}{\mu_n C_{ox}(W/L) v_{OV}} = \frac{1}{\mu_n C_{ox}(W/L)(v_{GS} - V_t)} \tag{2.9}$$

Hence, the conductance is dependent on v_{GS} as shown in Figure 4.

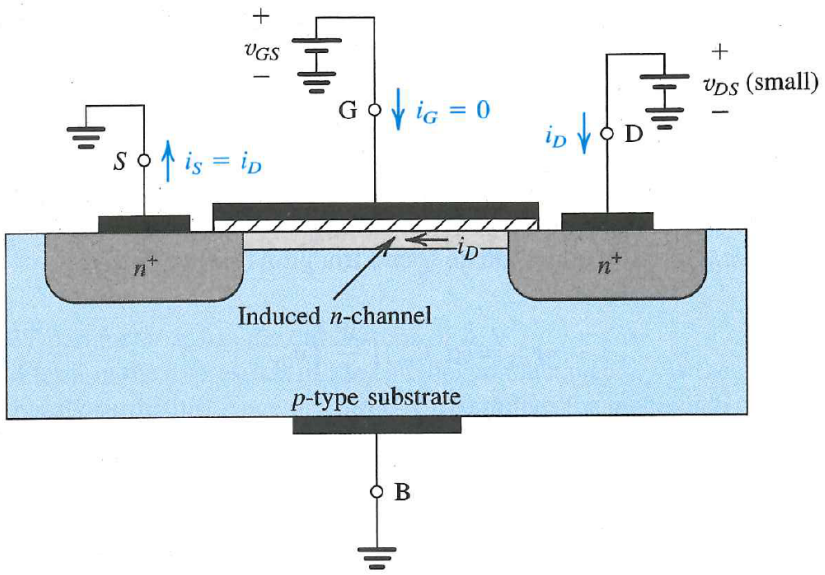

Figure 3: An NMOS transistor in operation with a positive gate voltage $v_{GS} > V_t$ where V_t is the threshold voltage. Depletion region is omitted for simplicity (Courtesy of Sedra and Smith).

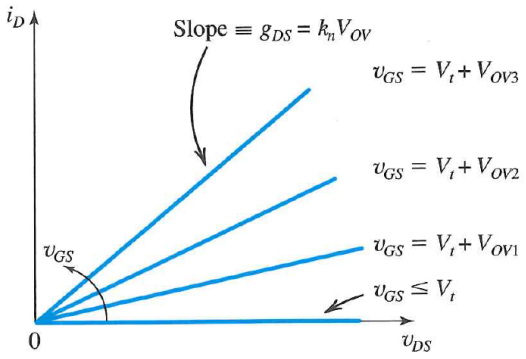

Figure 4: The slope of the curve which is g_{DS} is dependent on v_{GS} since $V_{OV} = v_{GS} - V_t$ (Courtesy of Sedra and Smith).

3 Large v_{DS} Model for MOSFET

Assume that V_{OV} is constant,[1] then as v_{DS} increases, the induced channel width will be different. Remember that the n carriers are drawn from the n^+ regions. Therefore, different v_{GS} will draw different amount of n carrier into the channel creating a non-uniform depth of the channel as shown in Figure 5.

The channel depth is proportional to v_{GS} at the source end or

$$v_{GS} = V_t + V_{OV} \tag{3.1}$$

and the channel depth is proportional to v_{GD} at the drain end or

$$v_{GD} = V_t + V_{OV} - v_{DS} \tag{3.2}$$

or that the effective overdrive voltage at the drain end is

$$V'_{OV} = V_{OV} - v_{DS} \tag{3.3}$$

To keep V_{OV} constant in this study, v_{GS} is kept constant in accordance with (3.1). Assuming that the voltage drop is linearly decreasing, then this is illustrated in Figure 6.

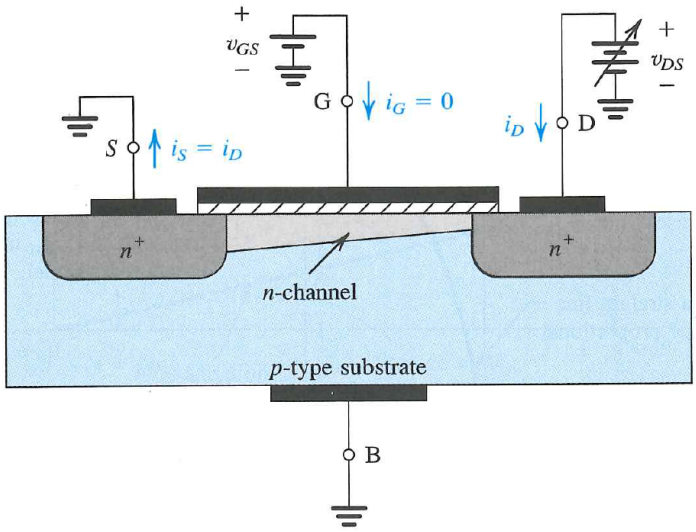

Figure 5: Unequal channel depth is induced when a larger v_{DS} is applied across the transistor giving rise to non-uniform channel depth (Courtesy of Sedra and Smith).

[1] In the textbook, capital letter is usually used to denote a DC value or non-time-varying value.

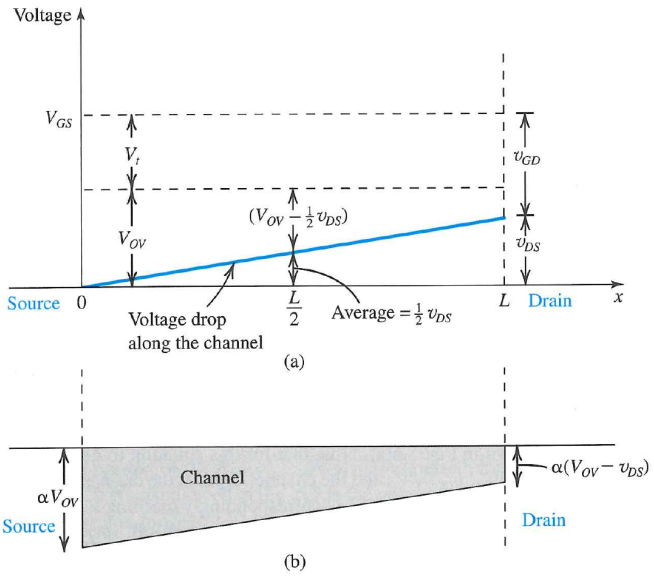

Figure 6: The linear proportionality of the channel depth is shown here. The channel depth at the source end is proportional to V_{OV}, while at the drain end, it is proportional to $V_{OV} - v_{DS}$ (Courtesy of Sedra and Smith).

The charge in the channel region Q is proportional to Cv_{OV}. If v_{OV} is a function of x, then this charge Q is proportional to the area of the channel.[2]

Without tapering of the charge, if v_{OV} is independent of x, the area under the curve is $v_{OV}L$. But with tapering, by using the area of a trapezoid, this area evaluates to $(V_{OV} - \frac{1}{2}v_{DS})L$. So if the analysis was previously done without tapering, one should replace v_{OV} with $V_{OV} - \frac{1}{2}v_{DS}$ for the tapered case. In (2.5), reproduced below, the current analyzed without tapering the channel region is

$$i_D = k'_n \frac{W}{L} v_{OV} v_{DS} \tag{3.4}$$

If the charge is tapered now, it should be (see Appendix for a more rigorous derivation)

$$i_D = k'_n \frac{W}{L} \left(V_{OV} - \frac{1}{2} v_{DS} \right) v_{DS} \tag{3.5}$$

Alternatively, we can write this current as

$$i_D = k'_n \frac{W}{L} \left(V_{OV} v_{DS} - \frac{1}{2} v_{DS}^2 \right) = k'_n \frac{W}{L} \left[(v_{GS} - V_t) v_{DS} - \frac{1}{2} v_{DS}^2 \right] \tag{3.6}$$

Notice that when v_{DS} is small, the quadratic term can be dropped and the small v_{DS} equation (3.4) is retrieved.

[2]See Appendix for a detail derivation.

4 Channel Pinch-Off and Saturation

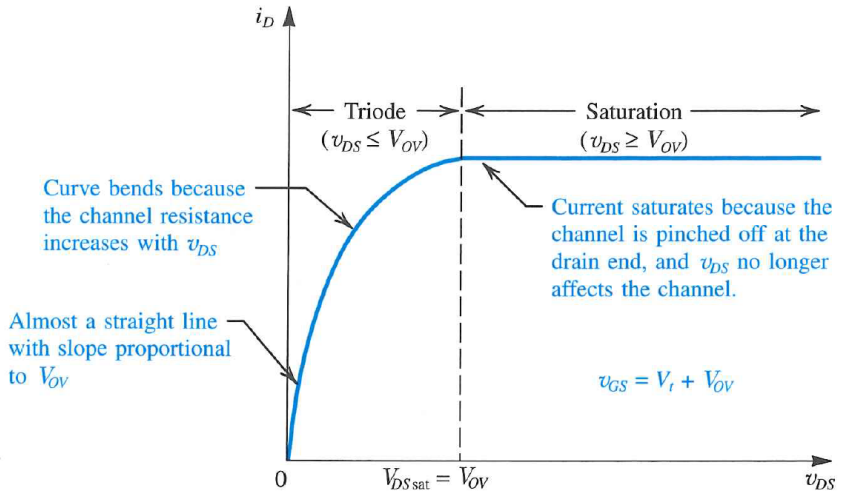

Figure 7: The i_D-v_{DS} relation of an NMOS transitor including the triode and saturation regimes (Courtesy of Sedra and Smith).

In the previous model, notice that when $v_{DS} = V_{OV}$, then the channel depth is zero near the drain end.[3] At this junction, increasing v_{DS} does not increase the drain current i_D. Then

$$i_{D\text{sat}} = \frac{1}{2} k_n' \frac{W}{L} V_{OV}^2 \tag{4.1}$$

The saturation voltage is given by

$$V_{DS\text{sat}} = V_{OV} = V_{GS} - V_t \tag{4.2}$$

When v_{DS} is further increased beyond the saturation regime,[4] the depletion regions around the pn junction further increase preventing the increase of i_D. The extra voltage of v_{DS} is also dropped across the depletion region.

The **triode region** refers to the operation region before the transistor reaches saturation. In general, as a function of v_{OV}, in the saturation region

$$i_D = \frac{1}{2} k_n' \frac{W}{L} v_{OV}^2 = \frac{1}{2} k_n' \frac{W}{L} (v_{GS} - V_t)^2 \tag{4.3}$$

In the saturation regime, the relation between i_{DS} and v_{GS} is nonlinear. MOSFET becomes a nonlinear device in this regime. Notice that the gate voltage

[3] One can show that this is also the inflexion point.

[4] Please be noted that the definition of the saturation regime for a MOSFET is very different from the saturation regime of a BJT.

v_{GS} controls the drain current i_D in a MOSFET similar to that the base-to-emitter voltage v_{BE} in a BJT controls the collector current i_C. But there is no exponential relation between the drain current i_D and the gate voltage v_{GS} in a MOSFET. In a BJT, the exponential relation can be replaced with a constant-voltage-drop model, or straight lines, whereas it is harder to do so with MOSFET. Therefore, a BJT is more linear compared to a MOSFET.

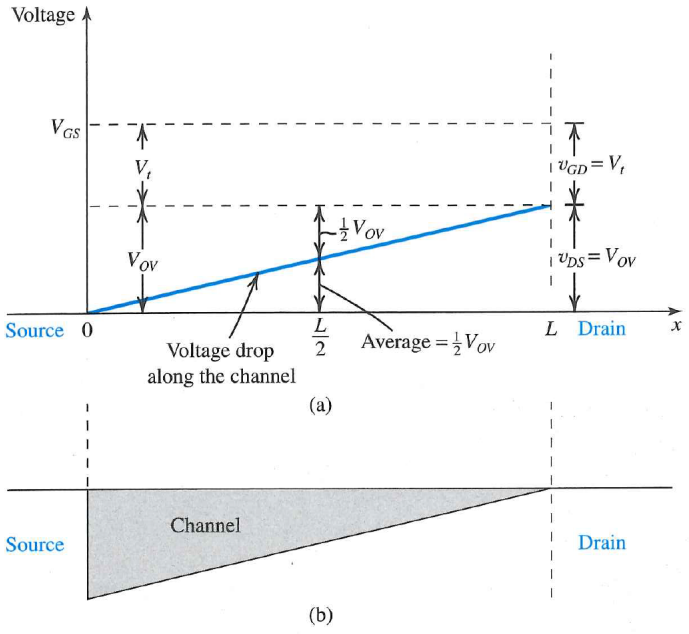

Figure 8: The shape of the channel of NMOS just before it reaches the saturation regime or pinch-off regime (Courtesy of Sedra and Smith).

5 The p-Channel MOSFET

The complementary device to NMOS can be made by replacing p region with n region and vice versa as shown in Figure 9.

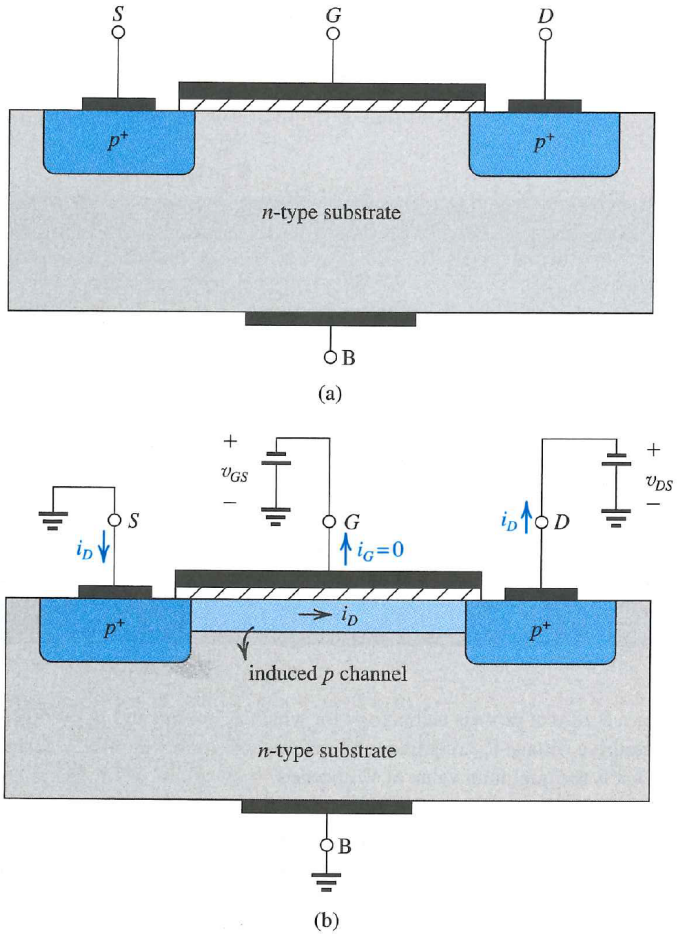

Figure 9: The physical structure of a PMOS, which is a complementary transistor to NMOS. Notice that the threshhold voltage $V_{tp} < 0$ and $V_{OV} < 0$ for such transistors (Courtesy of Sedra and Smith).

6 Complementary MOS of CMOS

Complementary MOS (CMOS) where NMOS and PMOS transistors work in tandem is very popular in digital circuits. In digital circuits, a transistor is either on or off, or having only two states. This is unlike analogue circuits where a transistor can be working continuously in different states. Because of the simplicity of transistor states in digital circuits, by clever engineering, CMOS can consume a lot less power than NMOS or PMOS alone. The structure of a CMOS transisfor is shown in Figure 10. A complementary structure to the above is also possible where NMOS is immersed in a p well instead.

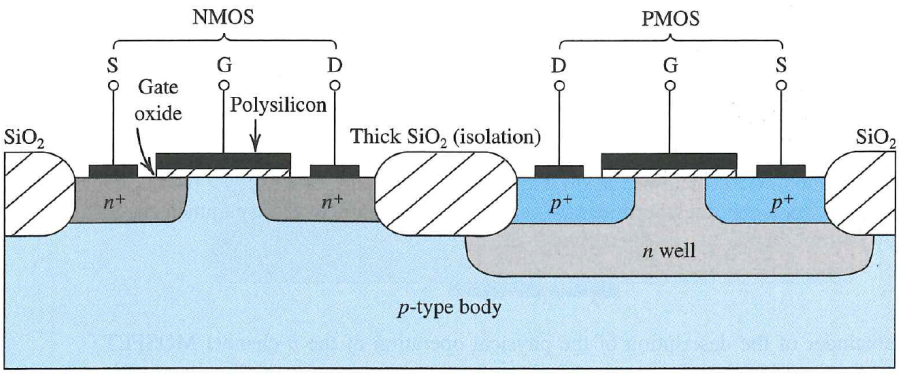

Figure 10: The physical structure of a CMOS integrated circuit (IC). The PMOS is immersed in a n well, and it is isolated from the NMOS by a thick oxide barrier (Courtesy of Sedra and Smith).

Appendix A Derivation if i_D for Strong Signals

If MOSFET is being biased with a strong drain to source voltage v_{DS}, then the voltage from the gate to the substrate will not be uniform. The voltage is highest near to the source, and gradually diminishes toward the drain. Hence, the induced conduction channel or the inversion layer will not be uniform. In this case, the channel is tapered in shape. The development here will parallel those in Section 2, but with adjustment of the equations for accommodate large v_{DS}.

If the channel is tapered, then the overdrive voltage v_{OV} is a function of x. The local electric field at position x is

$$|E| = -\frac{dv_{OV}(x)}{dx} \tag{A.1}$$

Then, instead of (2.5),

$$v_{\text{drift}} = \mu_n |E| = -\mu_n \frac{dv_{OV}(x)}{dx} \tag{A.2}$$

where μ_n is the electron mobility. Thus instead of (2.6), one gets

$$i_D = Q_n v_{\text{drift}} = -C_{ox} W v_{OV} \mu_n \frac{dv_{OV}}{dx} \tag{A.3}$$

The above can be written as

$$i_D dx = -\mu_n C_{ox} W v_{OV} dv_{OV} == -k'_n W v_{OV} dv_{OV} \tag{A.4}$$

In the above, i_D has to be a constant and independent of x due to charge or current conservation. Moreover, $v_{OV} = V_{OV}$ at the source end of the channel,

while $v_{OV} = V_{OV} - v_{DS}$ at the drain end of the channel. Therefore, integrating the above along the channel yields

$$i_D L = -k'_n W \left.\frac{v_{DS}^2}{2}\right|_{V_{OV}}^{V_{OV}-v_{DS}} = -k'_n W \frac{1}{2}\left[(V_{OV} - v_{DS})^2 - V_{OV}^2\right] \tag{A.5}$$

Simplifying the above, it reduces to

$$i_D = k'_n \frac{W}{L}\left(V_{OV} - \frac{1}{2}v_{DS}\right) v_{DS} \tag{A.6}$$

which is the same as (3.5).

Appendix B Tapering of the Channel

From the above, in order for the current i_D to remain constant, v_{OV} can be solved as a function of x. In other words, equating (A.3) and (A.6), one gets

$$v_{OV}\frac{dv_{OV}}{dx} = -C_0 \tag{B.1}$$

where the constant $C_0 = \left(V_{OV} - \frac{1}{2}v_{DS}\right) v_{DS}/L$. Solving the above, one gets

$$v_{OV}(x) = \sqrt{V_{OV}^2 - 2\left(V_{OV} - \frac{1}{2}v_{DS}\right) v_{DS}\frac{x}{L}} \tag{B.2}$$

where the integration constant is chosen such that $v_{OV}(x = 0) = V_{OV}$.

The above indicates that $v_{OV}(x)$ is not linearly tapered at all in order to maintain a constant drift current through the channel. Moreover, it can be shown that at $x = L$, $v_{OV}(x = L) = V_{OV} - v_{DS}$.