UNIVERSITY OF CALIFORNIA AT BERKELEY
College of Engineering
Department of Electrical Engineering and Computer Sciences

Discussion #9

EE 105 **Spring 2008**
Prof. Wu

1 MOSFETs

The standard MOSFET structure is shown in Figure 1. It consists of a metal gate, a layer of insulating oxide, and a silicon substrate (hence the name MOSFET: **m**etal-**o**xide-**s**emiconductor **f**ield-**e**ffect **t**ransistor). Just as we had two types of bipolar junction transistors, we also have two types of MOSFETs: NMOSFET (n-type MOSFET, NMOS, or NFET) and PMOSFET (p-type MOSFET, PMOS, or PFET). Note that in modern devices, we actually use polysilicon as the gate material (for various technological reasons) with few exceptions (one notable one is Intel's latest 45 nm processors, which utilize a metal gate). The oxide material in modern MOSFETs is typically SiO_2 (silicon dioxide, or more precisely, SiON), again with few exceptions (again, of note is Intel's use of HfO_2, hafnium dioxide, it its latest processors).

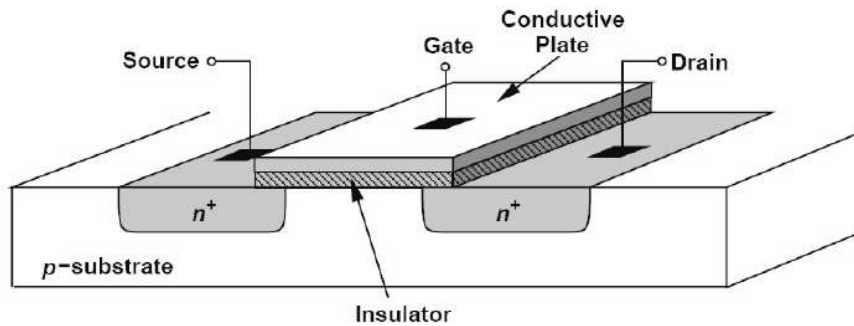


Figure 1: Idealized MOSFET structure

Let's restrict our discussion to NMOS transistors first. We'll generalize to PMOS later on. The structure in Figure 1 is an NMOSFET, since, as we'll see shortly, the charge carriers are electrons. Figure 2 shows a more convenient cross-sectional representation of the MOSFET. A typical NMOS has a n+ source and drain, a p-type substrate, and an n+ gate.

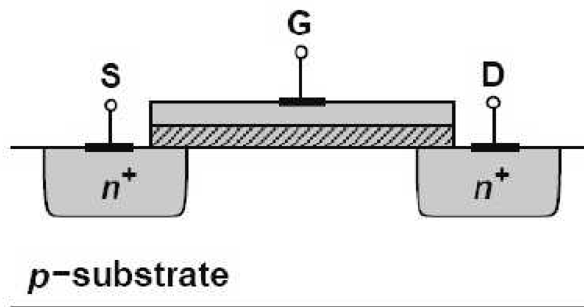


Figure 2: MOSFET cross section

First, consider the potential distribution along the surface of the MOSFET without any biases applied. What you'll see is that there is a large potential barrier (much like in a PN junction in equilibrium) between

the n+ source and the p-type substrate. We can modulate this potential barrier by applying a gate-source voltage V_{GS}. Once V_{GS} exceeds some threshold value V_{TH}, we'll have attracted electrons into the substrate between the source and drain into what we call the channel. The MOSFET then has a conductive path (of mobile electrons) that can conduct current between the source and drain.

Note that we're forming what is essentially an n-type channel between the source and drain, inverting the type of silicon that makes up the substrate (which is p-type). Thus, when the channel has mobile electrons, we often call the channel inverted, or describe it has having an inversion layer.

Figure 3 shows the circuit symbol for a NMOS. Note that for an NMOS, I_D is defined as flowing from drain to source. Like a BJT, a MOSFET operates differently depending on how it is biased, and next we'll discuss the regions of operation of a MOSFET. Note that we'll typically use the notation shown in Figure 4, which is a simplified NMOS symbol.

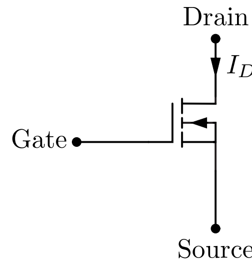


Figure 3: Circuit symbol for an NMOSFET

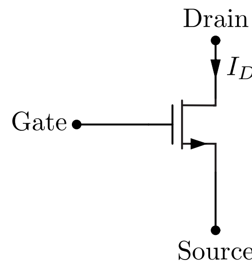


Figure 4: Simplified circuit symbol for an NMOSFET

1.1 Cutoff

When $V_{GS} < V_{TH}$, we have no mobile carriers in the channel. Without mobile carriers, no current can be conducted, so $I_D = 0$.

1.2 Triode/Linear

In triode (also known as linear), we have $V_{GS} > V_{TH}$ and the channel stretches from the source to the drain. The voltage condition where this is true is when $V_{DS} < V_{DSat}$, where $V_{DSat} \triangleq V_{GS} - V_{TH}$ (that is, when $V_{DS} < V_{DSat}$, the channel stretches from the source to the drain, while when $V_{DS} > V_{DSat}$, the channel stops short of the drain).

Note that the primary mode of current flow is by drift. Once we have a conductive channel, applying V_{DS} will set up an electric field that points from drain to source, causing electrons to flow from source to drain. Thus, we have current flowing from drain to source.

When we have a channel that goes all the way to the drain, the voltage V_{DS} must be dropped across the channel. That means the electric field will depend on V_{DS}, meaning the current will depend on V_{DS}. Also note that V_{GS} will affect the current since it controls how much charge is in the channel. The resulting equation for I_D is as follows:

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

Note that I_D depends on both V_{GS} and V_{DS}, which is why this region of operation is called triode. Also note that it is linear with V_{GS}, which is why this region is also called linear.

1.3 Saturation

Once $V_{DS} > V_{DSat}$, the channel no longer goes from the source to the drain. The channel actually ends before the drain edge (or right at the drain edge for $V_{DS} = V_{DSat}$). This effect is called pinch-off (since the channel is pinched off from the drain). When this occurs, the voltage V_{DSat} is dropped from the source to the edge of the channel, and the voltage $V_{DS} - V_{DSat}$ is dropped from the edge of the channel to the drain (for a total voltage drop of V_{DS} from the source to the drain). Note that V_{DSat} does not depend on V_{DS}, meaning that the electric field across the channel (and therefore I_D) no longer depend on V_{DS} in saturation. The resulting equation for I_D is as follows:

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

1.3.1 Channel Length Modulation

Although it's true that the voltage drop across the channel doesn't depend on V_{DS} in saturation, we have implicitly assumed that the channel length also does not change (since the electric field is inversely proportional to the channel length). However, this isn't quite correct. As we increase the drain voltage, we're increasing the reverse bias on the PN junction formed between the drain and the substrate. This causes a depletion region to form that pinches off the channel more and more as we increase V_{DS}. Thus, we in fact do have some dependence of I_D on V_{DS} in saturation due to channel length modulation. In order to correct for this, we simply add a correctional term to the above equation, giving us:

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

Note that this equation is different from that given in the textbook. It models channel length modulation in a more physically intuitive manner (since when $V_{DS} - V_{DSat}$, we shouldn't have any channel length modulation) than the textbook's formula.

1.4 Subthreshold Swing

Recall that V_{GS} is actually adjusting the potential barrier between the source and channel. When V_{GS} is significantly larger than V_{TH}, the channel has plenty of mobile electrons and the current is limited drift due to V_{DS} (or V_{DSat} when the device is in saturation). However, when V_{GS} is around V_{TH} or less than V_{TH}, what actually dominates the current is how many electrons are injected over the source-side potential barrier. This is much the same as the current limited in a diode due to the amount of minority carrier injection we can get from the p-type side to the n-type side (and vice versa). It turns out the drain current in this region (what we call the subthreshold region) is exponentially dependent on V_{GS}:

$$I_D \propto e^{qV_{GS}/\eta kT}$$

Note that η is simply a constant that is greater than 1. In this subthreshold region, one parameter we care a lot about is the subthreshold swing S, which is defined to be the amount of change in V_{GS} required to produce a $10\times$ change in I_D. S is defined in terms of mV per decade (i.e., how many millivolts change in V_{GS} is required to cause a one decade change in I_D?). We can calculate S as follows:

$$
\begin{aligned}
\frac{10I_D}{I_D} &= \frac{e^{qV'_{GS}/\eta kT}}{e^{qV_{GS}/\eta kT}} \\
10 &= e^{q\left(V'_{GS}-V_{GS}\right)/\eta kT} \\
&= e^{q\Delta V_{GS}/\eta kT} \\
\Delta V_{GS} &= \eta \frac{kT}{q} \ln 10 \\
S &= \eta \times 60 \text{ mV per decade}
\end{aligned}
$$

Since $\eta \geq 1$, we can see that the subthreshold swing must be at least 60 mV/dec (typical values are 80–100 mV/dec). Note that a smaller subthreshold swing is better, since we want the device to turn off sharply once V_{GS} drops below V_{TH}.

Note that we have a fundamental trade-off between the on-state current and the off-state current. If we define V_{TH} to be larger, that will give us a smaller off-state current (since V_{GS} will be farther below V_{TH} in the off-state), which reduces static power consumption (i.e., the power consumed when the device is off). However, a larger V_{TH} means a smaller V_{DSat} (since $V_{DSat} = V_{GS} - V_{TH}$), meaning a smaller on-state current (since $I_{D,on} \propto V_{DSat}^2$). Similarly, picking a smaller value of V_{TH} will result in more off-state leakage but more on-state current as well.

1.5 Summary: Large Signal Behavior

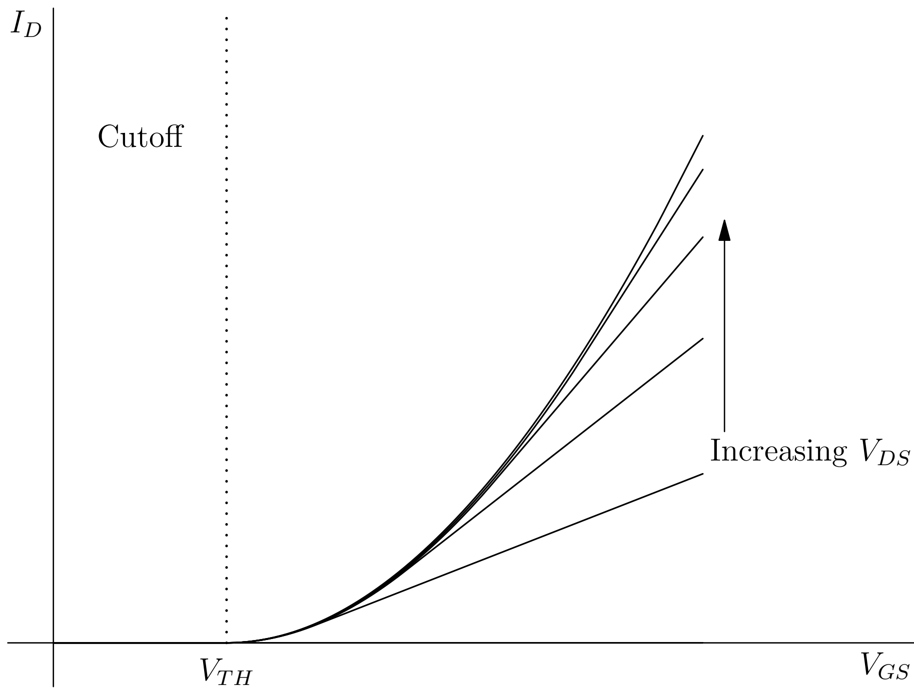
In this class, we'll focus on cutoff, triode/linear, and saturation. The behavior can be summarized as follows:

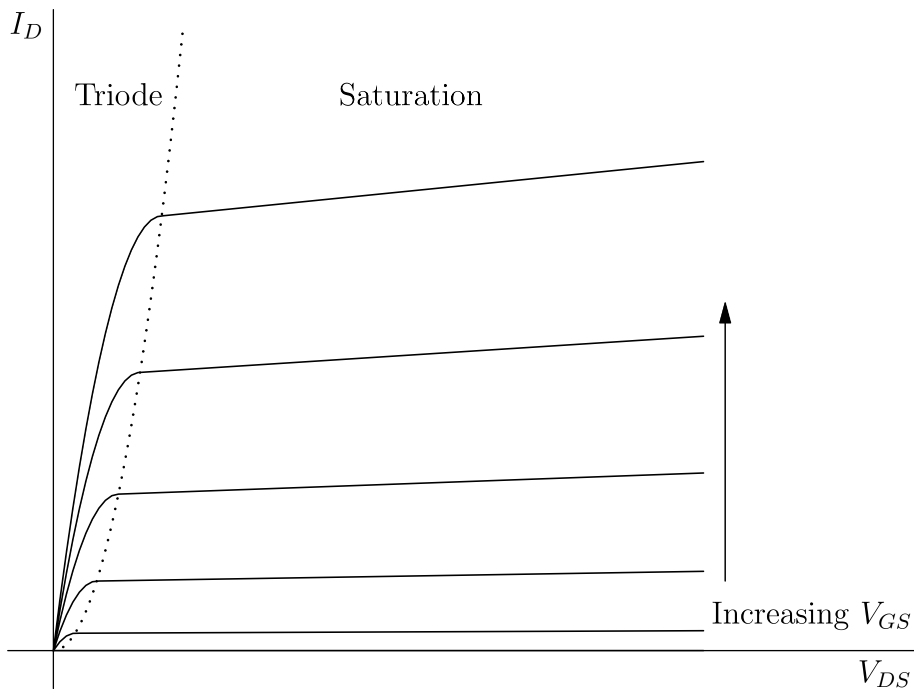
$$
I_D = \begin{cases} 0 & V_{GS} < V_{TH} \text{ (Cutoff)} \\ \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left[2\left(V_{GS}-V_{TH}\right)V_{DS} - V_{DS}^2\right] & V_{GS} > V_{TH}, V_{DS} < V_{DSat} \text{ (Linear/Triode)} \\ \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left(V_{GS}-V_{TH}\right)^2\left[1+\lambda\left(V_{DS}-V_{DSat}\right)\right] & V_{GS} > V_{TH}, V_{DS} > V_{DSat} \text{ (Saturation)} \end{cases}
$$

One other useful equation to remember is what V_{GS} is given I_D (assuming $\lambda = 0$):

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

The following I-V curves show the behavior of the MOSFET graphically.





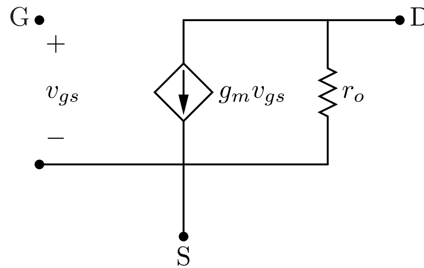
1.6 Small-Signal Model

We can define small-signal quantities analogous to those used in our study of the BJT. Since we will always operate our transistors in saturation, we'll focus on developing a model for the transistor in saturation. We can define the following two small-signal parameters:

$$
\begin{aligned}
g_m &\triangleq \frac{\partial I_D}{\partial V_{GS}} \\
&\approx \mu_n C_{ox} \frac{W}{L} \left(V_{GS} - V_{TH}\right) \\
&= \sqrt{2\mu_n C_{ox} \frac{W}{L} I_D} \\
&= \frac{2 I_D}{V_{GS} - V_{TH}} \\
r_o &\triangleq \left(\frac{\partial I_D}{\partial V_{DS}}\right)^{-1} \\
&= \frac{1}{\lambda \frac{1}{2} \mu_n C_{ox} \left(V_{GS} - V_{TH}\right)^2} \\
&\approx \frac{1}{\lambda I_D}
\end{aligned}
$$

Note that the approximations shown above are commonly used to simplify the expressions for g_m and r_o. Also note that the three expressions given for g_m are all useful in some cases, depending on what information you've been given.

Note that since the gate is insulated, $I_G = 0$, meaning there is no resistor between the gate and source (unlike in a BJT, where we had r_π from the base to the emitter). Thus, our small-signal model is as follows:



This small-signal model is identical to the small-signal model of a BJT except with $r_\pi = \infty$. Thus, we can re-use all of our small-signal analysis on BJTs in analyzing MOSFETs simply by taking the same expressions and letting $r_\pi \to \infty$ (and likewise $\beta \to \infty$).

1.7 PMOSFETs

Figure 5 shows the circuit symbol for a PMOS transistor. Note that the convention we'll use in this class is that I_D flows from the source to the drain, which deviates from the convention introduced in the textbook (our convention is more convenient for circuit analysis). Again, in this class we'll typically use the simplified circuit symbol in Figure 6.

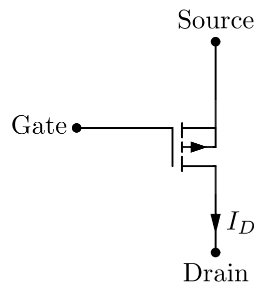


Figure 5: Circuit symbol for a PMOSFET

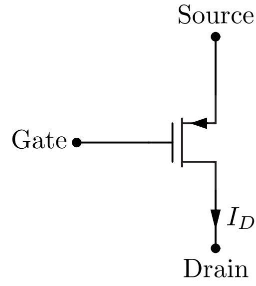


Figure 6: Circuit symbol for a PMOSFET

PMOSFETs are analogous to PNP transistors: they use the opposite type of charge carrier (holes), but otherwise behave very similar to their n-type counterparts (with a few sign changes). Although there are many ways to formulate the large-signal equations for PMOS transistors, I will only present one form here. Another document on the website has other forms that you may find more convenient to work with, so if you do not like the convention introduced here, I encourage you to check out that document.

$$I_D = \begin{cases} 0 & |V_{GS}| < |V_{TH}| \text{ (Cutoff)} \\ \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left[2\left(|V_{GS}| - |V_{TH}|\right)|V_{DS}| - |V_{DS}|^2\right] & |V_{GS}| > |V_{TH}|, |V_{DS}| < |V_{DSat}| \text{ (Linear/Triode)} \\ \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left(|V_{GS}| - |V_{TH}|\right)^2\left[1 + \lambda\left(|V_{DS}| - |V_{DSat}|\right)\right] & |V_{GS}| > |V_{TH}|, |V_{DS}| > |V_{DSat}| \text{ (Saturation)} \end{cases}$$