Metal-Oxide Semiconductor Field-Effect Transistor (MOSFET)

The **metal-oxide semiconductor field-effect transistor (MOSFET)** is actually a four-terminal device. In addition to the drain, gate and source, there is a **substrate**, or **body**, contact. Generally, for practical applications, the substrate is connected to the source terminal. If this is the case (and it usually is), the MOSFET may be considered a standard three-terminal device, with the drain, gate and source the terminals of interest.

Like all FET structures, the MOSFET uses the field effect to operate – the attraction or repulsion of charge carriers through an applied voltage – but this device has a twist that has allowed it to become the predominant technology for silicon based FETs. The MOSFET structure has dominated primarily due to the availability of a high quality oxide (SiO_2, or silicon dioxide) for the silicon system. As we will see, this oxide acts as an insulator and provides electrical isolation between the gate and an active (conduction) channel between the source and drain, thus providing the required input/output isolation.

Now, we get into an interesting part… as you go through this next segment, keep in mind that it's okay to share your head in bewilderment initially and that we will be going into detail about all this!

Recall back to our discussion of BJTs and the definition of the two possible "flavors" that depended on device construction and bipolar operation:

- the npn BJT, where electrons are the majority carrier and holes are the minority carrier; and
- the pnp BJT, where holes are the majority carrier and electrons are the minority carrier.

As mentioned above in our discussion of JFETs, we have the same sort of situation with FETs, but now we only have to define the device in terms of a majority carrier type (unipolar operation). Specifically for MOSFETS we have:

- the n-channel MOSFET, called the **NMOS**, where the majority carrier type is electrons; and
- the p-channel MOSFET, or **PMOS**, where the majority carrier type is holes.

Okay, that's pretty reasonable. *But…* for MOSFETs (and some other FET structures) there is another twist. Due to the insulator layer that exists

between the gate and the substrate, we can modify the device structure to allow two modes. This will become clearer below when device schematics are presented and in subsequent discussions when we investigate individual configurations – right now just hang on, we're in *"definition mode."*

- In the **depletion mode**, a channel region of a material type corresponding to the majority carrier (i.e., n-type material for NMOS and p-type material for PMOS) is physically implanted at the time of fabrication to connect the source and drain. Recall that an active (a.k.a. conduction or inversion) channel is required for current to flow between the source and drain, so the depletion MOSFET is in a normally on condition until the field effect is used to turn it off. To turn a depletion type MOS transistor off, or interrupt current flow between the source and drain, appropriate bias conditions must be generated in order to deplete the channel region of charge carriers, thereby removing the conduction path between source and drain.
- Conversely, in the **enhancement mode**, the conduction channel is created through the field effect. Enhancement mode operation is characterized by the requirement that appropriate biasing conditions be applied to create a conduction channel between the source and drain diffusions. In this mode of operation, the transistor exists in a normally off state until the conduction channel is created and current can flow between the source and drain.

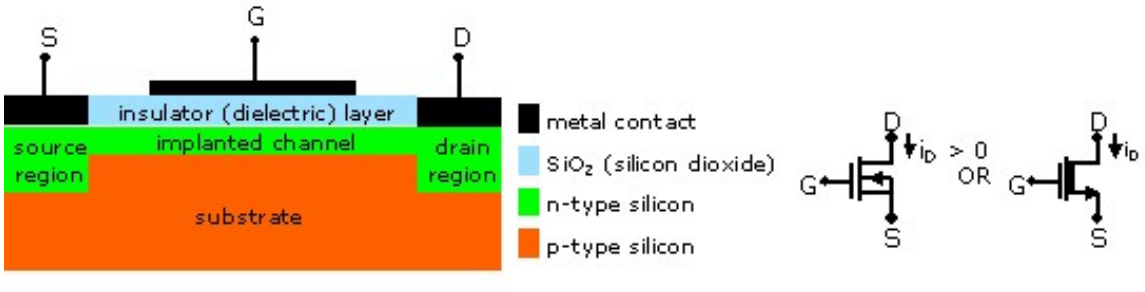
Can you see the handwriting on the wall?

Having two possible majority carrier types and two possible modes of operation as defined above, we come up with four possible MOSFET device types:

depletion NMOS **depletion PMOS**
enhancement NMOS **enhancement PMOS**

The schematic for each of these devices, as well as a general operational discussion, is presented below. Note that these are slightly different representations than what is presented in your text, but all physical properties are unchanged. Each figure illustrates the particular device regions, as well as materials and doping types for the silicon regions. Two symbols are also provided for each figure – the first is used by your author (and the one we will generally use), but don't get upset if the other version shows up in other texts or references. Also, please keep in mind that these schematics are for illustrative purposes only and are not intended to be a true representation of the actual fabrication techniques, relative geometries of the device, or absolute placement of the different structural regions.

The n-channel depletion MOSFET (depletion NMOS)



n-channel depletion MOSFET (depletion NMOS)

The **depletion NMOS device** is formed from a p-type substrate with physically implanted n-type source, drain and channel regions. The dielectric material covers the area between the source and drain to provide electrical isolation as mentioned earlier and allows the field-effect operation to occur. For a gate-to-source voltage (v_{GS}) greater than or equal to zero, the channel is active and, if a sufficiently large drain-to-source voltage (v_{DS}) is applied, electrons will move through the channel from the source to the drain, for a positive drain current ($i_D > 0$).

OK, this is totally goofed up, right? Let's back up for a minute and see what's going on…

- $v_{GS} \geq 0$ means that the implanted channel will remain unchanged ($v_{GS}=0$) or that even more majority carriers will be attracted to the channel at the Si/SiO_2 interface. Charges cannot move through the insulator (ideally), so a positive applied gate bias has the "effect" of attracting more negative charges from the substrate to the channel. Note that since the gate is isolated from the source (and the channel) by the oxide layer, the gate current is negligibly small and may be considered to be zero.
- Now, if we have a conduction channel and apply a large enough positive v_{DS} (remember that this means the potential at the drain is higher than that of the source), electrons in the channel (and source region) will be attracted to the drain. The channel size is increased, resulting in an increased drain current (within limits, of course). Physically what's happening is that the extra majority carriers increase the channel conductivity or, equivalently, reduce the channel resistivity (remember $\sigma=1/\rho$?).
- *But…* since conventional current directions are defined with respect to the movement of positive charges, we have a net positive current flow

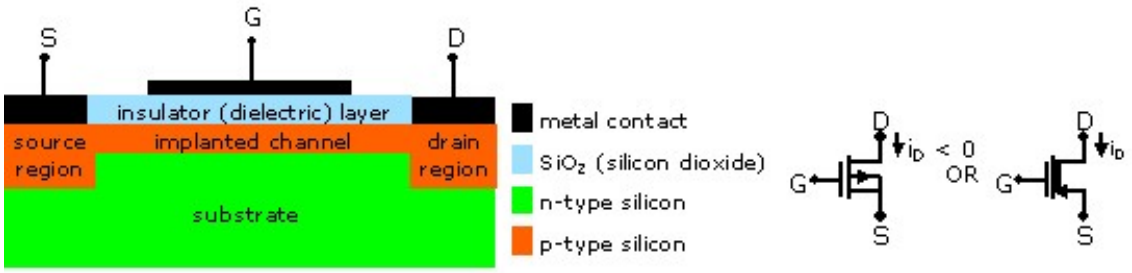
(negative charges moving opposite positive current has the same effect as positive charges moving with positive current).

Clear as mud?!?!? Breathe deeply and go through it again…

Now, lets look at what happens when v_{GS} goes negative. For a negative v_{GS}, the potential at the gate is less positive than that of the source. Usually the source is grounded, so this means that the gate potential is negative, or that negative charges are "piled up" along the gate contact. Since like charges repel, the negative charges on the gate push electrons out of the channel region and into the substrate, thereby depleting the majority carriers in the channel. When v_{GS} reaches a specific magnitude known as the **threshold voltage** V_T (also referred to as pinch-off in your text). At this threshold value, the channel is considered to be completely depleted of majority carriers and the drain current magnitude is reduced to zero for any applied v_{DS}.

A comment here before we go any further. Please do not confuse the threshold voltage of the MOSFET with the thermal voltage (V_T=26mV at room temperature if n=1, remember?) Your text uses the same notation for both parameters, but they have nothing to do with each other!

The p-channel depletion MOSFET (depletion PMOS)



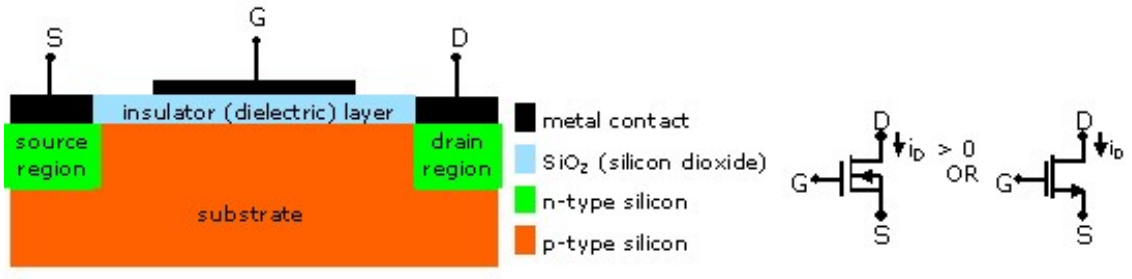
p-channel depletion MOSFET (depletion PMOS)

The **depletion PMOS device** is complementary to the depletion NMOS except the n-type and p-type silicon designations are interchanged. This device operates exactly as discussed above, with the following modifications:

- The conduction channel exists for $v_{GS} \leq 0$, with the conductivity increasing for negative v_{GS} (more holes are attracted to the channel region. For a negative v_{DS} of sufficient magnitude, current flows through the channel and i_D is nonzero.

- **<u>Please note:</u>** there are two ways of looking at the drain current i_D (remember we're dealing with holes as the majority carrier now). For a negative v_{DS}, the drain is at a lower potential than the source and holes will move through the channel from the source to the drain. In the graphic above and Figure 6.2(b) of your text, the direction of the current is shown as out of the drain contact with the notation for the current shown as "$i_D < 0$." (you'll see why in a little while). This is exactly the same thing as a positive current into the drain contact. ***Be sure you are comfortable that these two representations mean exactly the same thing!***
- To deplete the channel in a PMOS device, a positive v_{GS} is applied since the positive gate will repel the holes in the channel. The channel is again pinched off at a threshold voltage V_T, but instead of being a negative as for the depletion NMOS, it is now positive with a magnitude that depends on specific material and structural parameters. Above the threshold voltage magnitude, i_D is once again zero regardless of the v_{DS} applied.

<u>The n-channel enhancement MOSFET (enhancement NMOS)</u>

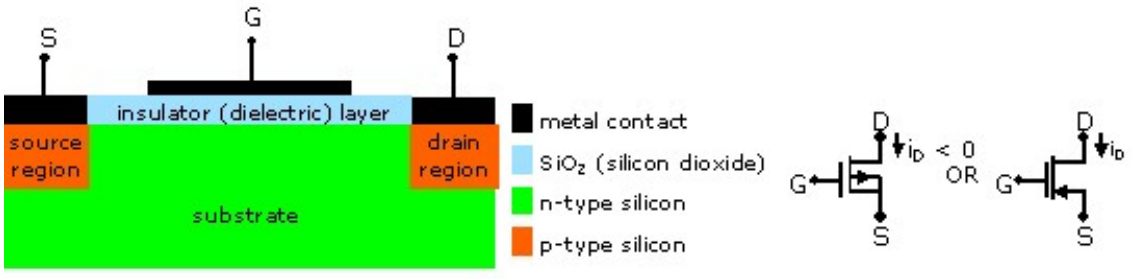


n-channel enhancement MOSFET (enhancement NMOS)

As shown above, the **enhancement NMOS device** differs from the depletion NMOS by the absence of the implanted channel. As stated earlier, this type of device is normally off and requires the creation of a conduction channel to allow current to flow. For the NMOS structure, this is achieved by applying a positive v_{GS} of sufficient magnitude ($v_{GS} > V_T$). The positive gate potential attracts electrons from the substrate to the Si/SiO_2 interface under the oxide layer. This process continues until enough electrons have been accumulated between the source and drain to increase the conductivity to a point that a conduction channel is formed. Note that no appreciable drain current can flow until the v_{GS} magnitude exceeds the threshold voltage V_T.

The remainder of device operation is as discussed for the depletion NMOS. Once the channel has been created, a positive v_{DS} will result in a drain current i_D.

The p-channel enhancement MOSFET (enhancement PMOS)



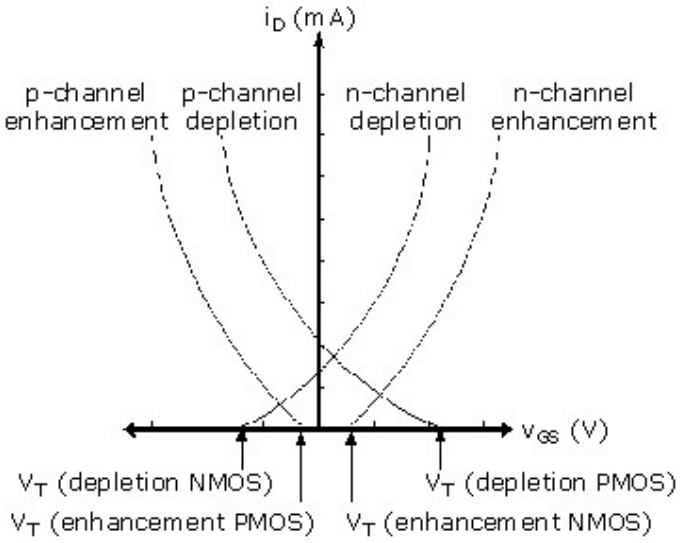
p-channel enhancement MOSFET (enhancement PMOS)

By this time, the functionality of the **enhancement PMOS device** should come as no surprise. Because it does not have a built-in channel and holes are the majority carriers, a negative v_{GS} with a magnitude greater than V_T must be applied to create the conduction channel. Once the channel is created and drain current can flow, a negative v_{DS} of appropriate magnitude results in a negative current with a direction as shown above (or a positive current with an opposite direction).

These discussions get a little shorter each time don't they?

To summarize: The operation of a MOSFET device is based on the variation of charge carriers in a channel that exists at the oxide-semiconductor interface between the source and drain diffusion regions. The control of charge carriers in the channel is not a function of dopants, as for the BJT, but instead is achieved through modulation of an applied gate bias. In enhancement mode, the MOSFET is on, or operating with an active channel, when a high density of majority carriers is attracted to the oxide-semiconductor interface. Variation in applied gate bias provides the mechanism for modulation of the current flow between the source and drain contacts. An analogous discussion may be developed for the depletion type structure, with the difference between the two modes of operation dependent upon gate bias polarity. Using a PMOS device as an example: where a negative gate bias is required for enhancement mode operation to attract positive charge carriers to the oxide-semiconductor interface and create an active channel, a positive gate bias is used for depletion mode operation to repel positive carriers and remove the connection between source and drain.

Finally, Figure 6.5 (reproduced to the right) illustrates sketches of the i_D-v_{GS} characteristics for enhancement and depletion mode MOSFETs. Note that it is assumed that a v_{DS} of appropriate polarity and magnitude has been applied to generate i_D. The characteristic curves intersect the v_{GS} axis at the threshold voltage V_T. Although it may not be readily apparent, keep in mind the $|V_T|$ is dependent upon several physical factors and is determined during device fabrication.

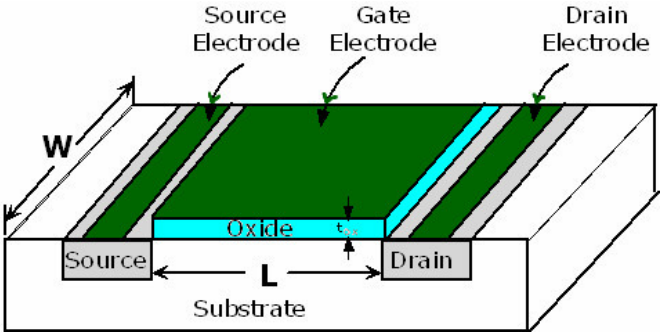


The remainder of this section of our studies is going to have to do with the terminal characteristics of the MOSFETs under consideration. Before we start though, I would like to explicitly state a couple of things that were implicit in the device introductions above:

- Just as we defined the normal flow of current in the BJT as being from the collector to emitter (assuming conventional current direction – please refer to the modified version of Figure 4.2(a) in Section C2 for clarification), the normal current flow in a MOSFET is defined as being from the drain to the source.
- This next one may be a little harder to see. If you look at a horizontal slice of the device that encompasses drain, substrate and source, you can see that we effectively have two back-to-back diodes between the drain and source. This results in a similar situation to our discussion of the BJT and means that appropriate external voltages must be applied to bias the device and allow current to flow.

We're going to be talking about the enhancement mode and depletion mode devices in detail next, but as a general overview…

Current flow in a MOS device is a function of several material, structural, and bias parameters, including (some of which are illustrated in the sketch to the right for an enhancement mode device): the **channel length L**, which is the separation distance



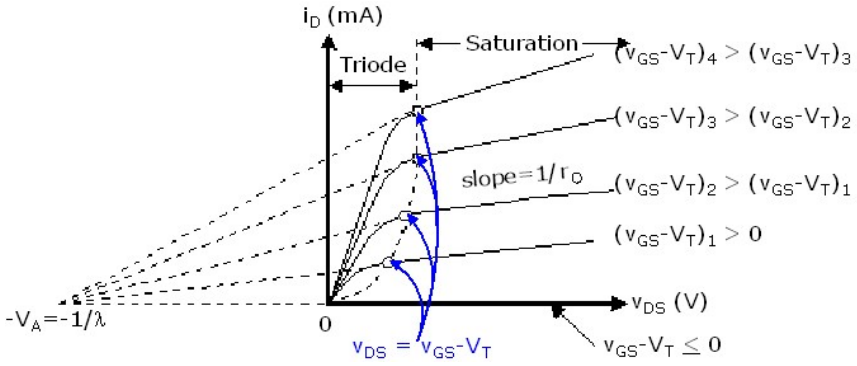
between the source and drain; the **channel width W**, **charge mobilities** within the active channel region (μ_n for electrons or μ_p for holes); the **capacitance of the oxide layer C_{ox}**, which is modeled as a parallel plate capacitor and is in turn a function of oxide layer thickness (t_{ox}) and permittivity (σ); the applied gate potential V_{GS}; drain to source bias V_{DS}; and the threshold voltage V_T. Although the MOSFET structure is inherently less sensitive to doping considerations when compared to a conventional bipolar junction transistor, the threshold voltage is dependent on the impurity concentration of the channel region. Additionally, since several breakdown mechanisms associated with MOS structures are contingent on the maximum electric field in the drain space charge region, doping criteria may become important to reduce these effects (we're not going to be getting into threshold voltage variations too much, but I just wanted to briefly present a little of what's going on).

Enhancement Mode MOSFET Terminal Characteristics

Schematics of the enhancement mode MOSFET devices are given above and are illustrated by Figures 6.3 and 6.4 in your text (note that there is an error in the text narrative when referencing the text's figures). To briefly review:

- Enhancement mode devices have no built in channel (i.e., the device is normally off).
- A threshold voltage, determined by physical and fabrication parameters, exists for the MOSFET ($V_T > 0$ for NMOS, $V_T < 0$ for PMOS).
- Generally, the source is grounded and is common to both gate and drain.
- An active channel is created between drain and source (the transistor is turned on) through the application of a gate voltage of appropriate polarity and magnitude ($v_{GS} – V_T > 0$ for NMOS, $v_{GS} - V_T < 0$ for PMOS).
- Current flow between the drain and source is a function of both v_{GS} and v_{DS}.

A representation of the generic terminal characteristics of an enhancement mode NMOS is shown to the right (note that this is a modified version of Figure 6.8 of your text). Analogous to the BJT forward characteristic curves of Section C4, once the device has been turned on ($v_{GS} > V_T$), current may flow between drain and source with an applied v_{DS},



generating individual curves for different values of v_{GS}-V_T. Also similar to the BJT, the MOSFET has different operational regions, depending upon external biases. The delineation of the two regions for the MOSFET (triode and saturation) is determined by the relationship between the applied drain-to-source voltage, the applied gate-to-source voltage and the threshold voltage, which occurs when $v_{DS}=v_{GS}-V_T$, as illustrated above.

If $v_{DS} < v_{GS}-V_T$, the transistor is operating in the **triode** region. As shown in the expanded view of Figure 6.6 of your text, the relationship between i_D and v_{DS} is approximately linear when operating in the triode region. This allows the MOSFET to be operated as a linear resistor whose resistance is controlled by v_{GS}. In the triode region, the potentials at all three terminals strongly affect the drain current (hence the name), and the drain current obeys the relationship:

$$i_D = K\left[2(v_{GS} - V_T)v_{DS} - v_{DS}^2\right], \text{ where } K = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}, \quad \text{(Equations 6.2 \& 6.3)}$$

with the parameters μ_n, C_{ox}, W and L as defined earlier.

If we now hold v_{GS} constant (let's look at a single curve at first), and increase v_{DS}, things start to change. Bear with me for a few minutes – the following discussion is slightly different from your text, but (I hope) will serve to clarify what's going on. Let's look at the applied v_{DS} as a distributed quantity along the channel – with respect to the grounded source, voltage in the channel varies from zero (when measured at the source) to v_{DS} (when measured at the drain). Now, if we simultaneously consider the effect of v_{GS}, we can see that both sources are trying to do the same thing (attract electrons) and are essentially opposing each other. Since the gate-to-source voltage controls the channel depth, and v_{GS} and v_{DS} are essentially competing for available charges, the channel does not have a uniform depth for any $v_{DS} > 0$. In fact, the effective gate-to-source voltage decreases from the applied v_{GS} at the source, to v_{GS}-v_{DS} at the drain and the channel takes on a tapered shape.

Another way of looking at this effect is to consider the gate-to-drain voltage (v_{GD}). This is just another way of looking at v_{GS}-v_{DS}, since:

$$v_{GD} = v_G - v_D = (v_G - v_S) - (v_D - v_S) = v_{GS} - v_{DS}.$$

However you want to look at it, when the difference between v_{GS} and v_{DS} becomes less than V_T, the depth of the active channel becomes zero and the channel is **constrained** or **pinched off**. This means that further increases in v_{DS} have little effect on i_D (ideally, no effect).

The condition $v_{DS} > v_{GS}-V_T$, defines the **normal active**, or **saturation** region of operation on the i_D-v_{DS} curves, with the boundary between the triode and saturation regions (called the **knee**) defined by $v_{DS}=v_{GS}-V_T$. If we could have an ideal device, the curves in the saturation region would be perfectly horizontal (i.e., absolutely no change in i_D with changes in v_{DS}). Using the saturation condition in Equation 6.2 and simplifying yields the expression for the drain current of an ideal MOSFET in saturation,

$$i_D = K\left[v_{GS} - V_T\right]^2 = KV_T^2\left(\frac{v_{GS}}{V_T} - 1\right)^2. \qquad \text{(Equations 6.4 and 6.6)}$$

However, since we *insist* on dealing with reality, the characteristic curves actually have a nonzero slope. This slope is accounted for by adding a linear factor $(1+\lambda v_{DS})$, where λ is a MOSFET parameter known as the **channel length modulation parameter** (just like for the JFET).

$$i_D = K\left[v_{GS} - V_T\right]^2(1+\lambda v_{DS}) \qquad \text{(Equation 6.5)}$$

Often, the term $|\lambda v_{DS}| << 1$, so the expression of Equation 6.4 (or 6.6) will suffice to define the drain current. However, the parameter λ does serve to define an important device parameter. If the straight-line characteristics of the i_D-v_{DS} curves are extrapolated backwards, they intercept the v_{DS} axis at a common point, labeled $-V_A$ in the figure above. $\mathbf{V_A}$ is a positive voltage for the enhancement NMOS that is similar to the **Early voltage** in the BJT. If the device were ideal, λ would be zero and V_A would be infinite. Practical devices generally have λ values on the order of $10^{-3}V^{-1}$, with corresponding V_A values of several hundred volts.

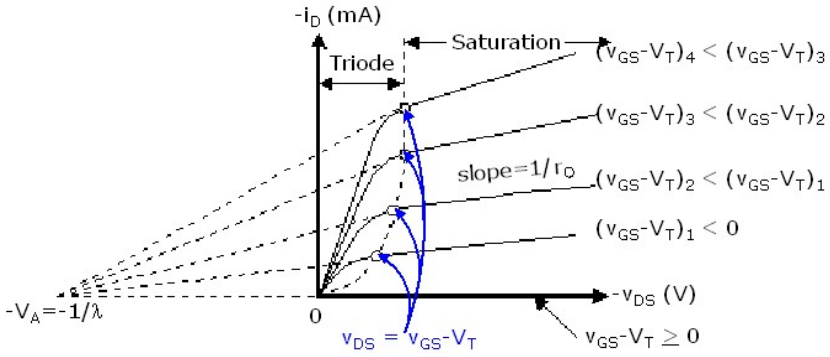
The final term illustrated in the figure above is the incremental **output resistance**, $\mathbf{r_O}$. For a constant v_{GS}, we can operate along one of the parametric curves of the figure and, in the saturation region; the instantaneous relationship between i_d and v_{ds} (ac quantities) defines r_O in terms of a partial derivative

$$r_O = \left[\frac{\partial i_d}{\partial v_{ds}}\right]^{-1}. \qquad \text{(Equation 6.12, Modified)}$$

Using the expanded expression of Equation 6.5, this partial turns out to be

$$r_O = \frac{1}{K(V_{GS}-V_T)^2\lambda} \cong \frac{1}{\lambda I_D} = \frac{V_A}{I_D}, \quad \text{(Equations 6.13 \& 6.15, Modified)}$$

where I_D is the bias current associated with the constant value of v_{GS} (denoted V_{GS}).



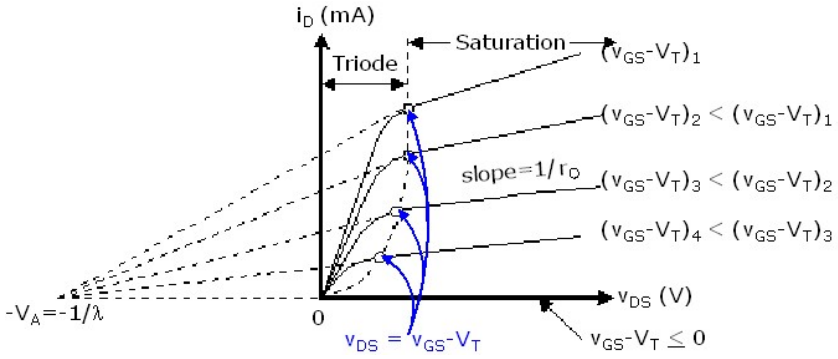
The figure to the right provides a possible representation of the terminal characteristics of an enhancement PMOS. Note that this is complementary to the NMOS in every respect, except the signs of the axes have been switched (v_{DS} is now $-v_{DS}$, i_D is now $-i_D$), the λ parameter is negative ($\partial i_D/\partial v_{DS} < 0$), and the curves for increasing current magnitudes correspond to more negative gate voltages. Since v_{GS} and v_{DS} are both negative, the expressions for currents in the triode and saturation regions are the same as for the enhancement NMOS. Keep in mind that these relationships give you current magnitudes and that, while the current flows into the drain for the NMOS, it flows out of the drain for the PMOS.

Depletion Mode MOSFET Terminal Characteristics

Schematics of the depletion mode MOSFET devices were given earlier and are illustrated by Figures 6.1 and 6.2 in your text. Following the same strategy as our discussion of enhancement mode devices, we will briefly review depletion mode device characteristics:

- Depletion mode devices have a built in channel (i.e., the device is normally on).
- A threshold voltage, determined by physical and fabrication parameters, exists for the MOSFET ($V_T < 0$ for NMOS, $V_T > 0$ for PMOS).
- Generally, the source is grounded and is common to both gate and drain.
- The active channel is depleted between drain and source (the transistor is turned off) through the application of a gate voltage of appropriate polarity and magnitude ($V_{GS} - V_T \leq 0$ for NMOS, $V_{GS} - V_T \geq 0$ for PMOS).
- Current flow between the drain and source is a function of both v_{GS} and v_{DS}. However, in the depletion mode, current can flow for both positive and negative values of v_{GS} for both NMOS and PMOS devices until the cutoff condition is reached.

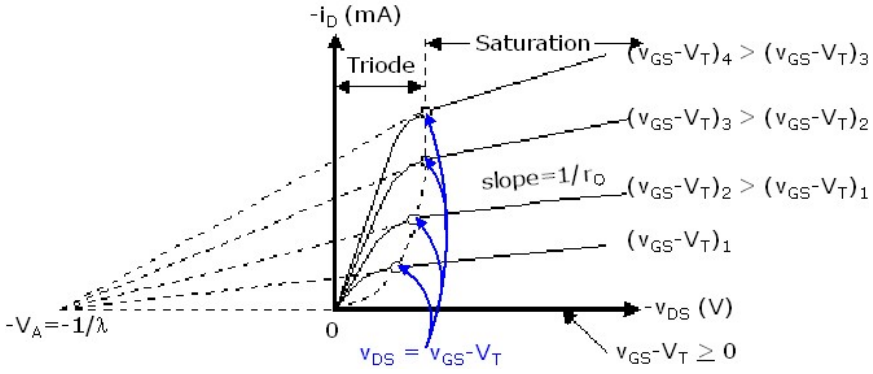
In keeping with our previous presentations of device terminal characteristics, a set of generic curves for the depletion NMOS is presented to the right. This figure is based on Figure 6.9 of your text, but has been modified to allow the same presentation as for the enhancement mode devices.



The equations that define the operation of the depletion mode MOSFET are very similar to those we developed for enhancement mode. However, since we can have a current when v_{GS} = 0, we can define the **drain-source saturation current**, also known as the **zero-gate drain current**, I_{DSS}. Using Equation 6.4 with v_{GS} = 0, we define $I_{DSS}=KV_T^2$. If desired, this parameter may be introduced into Equation 6.4 for the representation

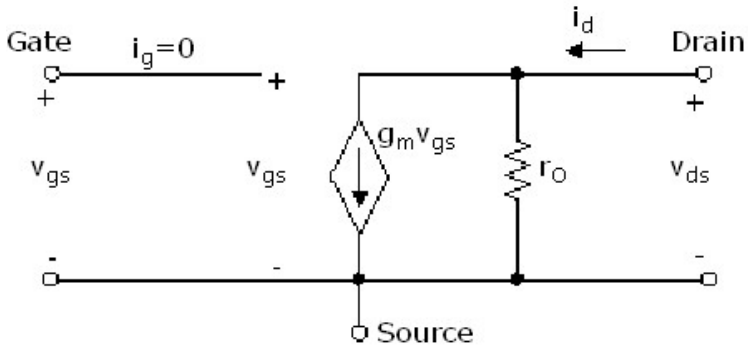
$$i_D = I_{DSS}\left(1-\frac{v_{GS}}{V_T}\right)^2 . \qquad \text{(Equation 6.11)}$$

Finally (literally!), we have the depletion PMOS. Everything we've talked about so far still holds, so we'll just take a quick look at some characteristic curves for the sake of completeness.



Small-Signal Model of the MOSFET

The last thing we're going to do in this section is introduce the small signal model for the MOSFET. The schematic for the small signal model is given in Figure 6.12 of your text and is reproduced to the right. Notice that this model is



directly analogous to the model derived for the BJT (with the appropriate analogies substituted), but that the FET has no gate current (implying an infinite differential input resistance). Remember that lower case symbols with lower case subscripts mean that we're dealing only with ac quantities. Everything in the figure has been defined except for the **transconductance** parameter, g_m. Using the same method as we used for the BJT, we find

$$g_m = \left.\frac{\partial i_d}{\partial v_{gs}}\right|_{v_{gs}=V_{GS}} = 2K(V_{GS} - V_T)(1 + \lambda v_{DS}) \cong 2K(V_{GS} - V_T) \quad \text{(Equation 6.15, Mod.)}$$

where the final approximation was made by assuming that λ is small.

Well, wasn't that just too much fun!

Now that you fought your way through all the words and pictures, the table below provides a summary of MOSFET characteristics:

	NMOS		**PMOS**	
	Enhancement Mode	Depletion Mode	Enhancement Mode	Depletion Mode
V_T	> 0	< 0	< 0	> 0
K	$\frac{1}{2}\mu_n C_{ox}\frac{W}{L}$		$\frac{1}{2}\mu_p C_{ox}\frac{W}{L}$	
Normally	Off	On	Off	On
To turn device on	$v_{GS} > V_T$		$v_{GS} < V_T$	
λ ($= V_A^{-1}$)	> 0		< 0	
To operate in triode region	$v_{DS} \leq v_{GS} - V_T$		$v_{DS} \geq v_{GS} - V_T$	
Drain current in triode region	$i_D = K[2(v_{GS} - V_T)v_{DS} - v_{DS}^2]$			
To operate in saturation region	$v_{DS} \geq v_{GS} - V_T$		$v_{DS} \leq v_{GS} - V_T$	
Drain current in saturation region	$i_D = K(v_{GS} - V_T)^2(1 + \lambda v_{DS})$			
Output resistance	$r_O \cong \frac{1}{\lvert\lambda\rvert I_D} = \frac{\lvert V_A\rvert}{I_D}$			
Transconductance	$g_m = 2K(V_{GS} - V_T)(1 + \lambda v_{DS}) \cong 2K(V_{GS} - V_T)$			