Metal oxide semiconductor field effect transistors (MOSFETs)

A newer form of transistor which has pretty much replaced BJT technology for all digital applications and much of analog.

Basically, an FET is a resistor whose resistance can be controlled through a third terminal. So the transistor mechanism is much different than that of the BJT. A BJT used both electrons and holes injected across the junctions (hence bipolar). An FET uses either electrons or holes (hence unipolar) flowing by drift current between two contacts (source and drain). The amount of flow is controlled by a voltage applied at the gate.

electron-current device: n-channel MOSFET (NMOS)
hole-current device: p-channel MOSFET (PMOS)

We’ll see shortly that using NMOS and PMOS together (known as CMOS) offers some significant advantages in circuit design.
Basic NMOS structure
(n-channel device)

Start with a $p$-type substrate.

Form two $n$-type regions
(source & drain)

Form a thin layer of silicon
dioxide on the $p$-type silicon
between source and drain

Put metal contacts on the
source, drain, and oxide (gate)
for electrical connection

PMOS has $n$- and $p$-type regions reversed. (Duh!)
Critical dimensions

oxide thickness: typical 1 - 10 nm.

gate length (distance from source to drain) – as small as 20 nm.

width: typical $L$ to $10L$ 
($W/L$ ratio is important)
Will current flow?

Apply a voltage between drain and source ($V_{DS}$).

If $V_{DS} > 0$, the diode at the drain end will be reverse-biased, preventing current flow.

Reversing the polarity doesn’t help, because then the other junction will be reverse biased.
The MOS capacitor

The secret to MOSFET operation lies in the MOS capacitor (the central part of the FET).

The structure is essentially a little parallel-plate capacitor formed by the gate metal and the semiconductor with the oxide as the dielectric.

\[
C_{gate} = \frac{\varepsilon_{ox} A}{t_{ox}}
\]

Define oxide capacitance

\[
C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}
\]

units: \( \text{F/m}^2 \) (or \( \text{F/\mu m}^2 \))

\( \varepsilon_{ox} = 3.9\varepsilon_0 = 3.9(8.85\times10^{-12} \text{ F/m}) \)

for \( t_{ox} = 10 \text{ nm} \) (10^{-8} \text{ m})

\[
C_{ox} = 0.00345 \text{ F/m}^2
\]

\( = 3.45\times10^{-15} \text{ F/\mu m}^2 \)
$P$-type semiconductor – the carriers are mostly holes, but there are few electron lurking around.

Try applying a negative voltage between the gate and the body: Holes are attracted to the gate region and electrons are pushed away from the negative voltage.

Hole concentration under the gate is enhanced. So negative voltage causes holes to “accumulate”. Probably not helpful for our NMOS.
So try applying a positive voltage to the gate. Now the holes are pushed away and the electrons are attracted.

The region under the gate becomes “depleted” of holes. So this is probably not helpful either, but the movement of the electrons towards the gate looks promising.
So try increasing the gate voltage further. Holes are pushed away further and electrons are attracted more strongly.

At a sufficiently high voltage, enough electrons will gather together to make the region under the gate “invert” and behave like it was n-type!

The gate voltage required create the electron “inversion layer” is called the threshold voltage, $V_T$. If the voltage is increase beyond $V_T$, even more electrons are gathered together under the gate.
The inversion layer

How many electrons are in the inversion layer? The usual definition of the threshold voltage says that when \( v_{GB} = V_T \), the electron concentration in the inversion layer is equal to the hole concentration in the substrate:

\[
n_{\text{inversion}} = p_{\text{substrate}}.
\]

This gives just enough electrons so that the region under the gate begins to act like it is \( n \)-type. A detailed physical description is beyond our capabilities now. (See EE 332 and beyond.) But we can get a good approximation by using the fact that the MOS structure is basically a capacitor. Once the gate voltage is bigger than the threshold, electrons pile up under gate just like charge on a plate of a capacitor.

\[
Q_{\text{inv}} = C_{\text{gate}} v_{ov} = C_{\text{gate}} (v_{GB} - V_t)
\]

\[
\frac{Q_{\text{inv}}}{A} = C_{\text{ox}} (v_{GB} - V_t) = qn_s
\]

where \( n_s \) is the sheet concentration of electrons in the inversion layer. (electrons per unit area).
The threshold voltage for a MOSFET depends primarily on two things. The doping level of the substrate and the thickness of the insulating gate oxide layer.

The definition of threshold is that the electron concentration under the gate is equal to the hole concentration in the rest of the substrate. If the hole concentration in the substrate is higher (due to higher doping) more voltage is required to induce the higher number of electrons needed for inversion. Increased doping means higher threshold.

It is actually the electric field in the gate that controls what is happening in the inversion layer. Since $\mathcal{E} = \frac{v_{GB}}{t_{ox}}$, making the oxide thicker requires more voltage in order to create the same electric field. So a thicker oxide also leads to higher a threshold voltage.

Integrated circuit fabrication engineers work very hard to control the threshold voltages of the MOSFETs. Not easy when there are millions (or billions) of transistors in a chip!
Summary of MOS capacitor operation

So through the application of the gate voltage, we can control what is happening with carriers under the gate.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{GB} &lt; 0$</td>
<td>hole accumulation</td>
</tr>
<tr>
<td>$0 &lt; v_{GB} &lt; V_T$</td>
<td>carrier depletion</td>
</tr>
<tr>
<td>$v_{GB} &gt; V_{GB}$</td>
<td>inversion – electron layer forms, $q \rho_s = C_{ox}(v_{GB} - V_T)$</td>
</tr>
</tbody>
</table>

So through the application of the gate voltage, we can control what is happening with carriers under the gate.
Now we see a mechanism by which we might get current to flow between drain and source.

Apply a gate voltage to create an electron inversion layer. Then electrons from the source can flow to the drain using the inversion layer as a channel.

For now, we connect the source to the body (call this ground) and apply the controlling voltage between the gate and the source. This is OK for the time being, but we will have to revisit the issue of the body connection later. With the drain also at ground, the inversion layer (channel) is uniform between source and drain.
Drain current

With the inversion layer in place, we can apply a positive voltage between the drain and source. This creates an electric field that will push electrons from the source through the inversion layer channel to the drain. The moving electrons represent a drain current flowing from drain to source.

\[ v_{DS} > 0 \]

\[ v_{GS} > V_T \]
At first inspection, the current flow appears to be “ohmic”. The current is entirely due to electrons and the electrons move due to an applied voltage. Basically, it looks like a resistor.

The total resistance is the sum the resistances of the three regions: $R_{DS} = R_{\text{source}} + R_{\text{inversion}} + R_{\text{drain}}$. Usually, the source and drain are very heavily doped (lots of electrons) and so the resistances there are small. Most of the resistance is associated with the inversion layer: $R_{DS} \approx R_{\text{inversion}}$. 

\[ v_{DS} > 0 \]
\[ v_{GS} > V_T \]
Using the basic definition for a resistor:

\[
R_{DS} = \rho_{inv} \frac{L}{Wt_{inv}} = \frac{1}{q\mu_n n_{inv}} \frac{L}{Wt_{inv}}
\]

where \( n_{inv} \) is the electron concentration \((m^{-3})\) in the inversion layer, \( t_{inv} \) is the “thickness” of the inversion layer, \( \mu_n \) is the electron mobility, \( q \) is the charge on one electron and \( L \) and \( W \) are the gate length and width.

\[
i_D = \frac{v_{DS}}{R_{DS}} = \left( q\mu_n n_{inv} t_{inv} \right) \left( \frac{W}{L} \right) v_{DS}
\]

The electron concentration and inversion layer thickness are difficult to quantify because of the sheet-like nature of the inversion layer. However, the product \( n_{inv} \cdot t_{inv} \) can be re-expressed more easily – it represents the “sheet concentration” \((m^{-2})\) of electrons in the inversion layer, which we denote as \( n_s \).

We saw earlier that \( qn_s = C_{ox} (v_{GS} - V_T) \)
The resistance of the NMOS can be altered by changing the gate voltage. If $v_{GS} < V_T$, there is no inversion layer: $R_{DS} \rightarrow \infty$ and $i_D = 0$. (green curve)

For $v_{GS} \geq V_T$, the inversion layer forms and the electron concentration in it increases with increasing gate voltage. The channel resistance decreases, which in turn causes the current to increase. (Stated another way, channel conductance increases as $v_{GS}$ increases.) (blue, black, red curves.)

Inserting sheet concentration relationship into the drain current equation gives:

$$i_D = \left[ \left( \mu_n C_{ox} \frac{W}{L} \right) (v_{GS} - V_T) \right] v_{DS}$$

$\mu_n C_{ox} W/L = 1 \text{ mA/V}^2$ and $V_T = 2 \text{ V}$. 
However, the picture gets more complicated as $v_{DS}$ increases.

The electron sheet concentration depends on the local voltage of the semiconductor underneath the gate. In the MOS capacitor, the voltage of the semiconductor was the same everywhere, so the electron concentration was uniform. In the case of the MOSFET, the local voltage under the gate can vary from one end to the other. At the source end, the channel voltage is 0, (because we have tied the source to ground). At the drain end, the channel voltage is $v_D$, because we have applied a voltage there in order to make current flow. In between, the channel voltage ranges between 0 and $v_D$. 
Since $v_{GD} = v_{GS} - v_{DS}$, an increase in $v_{DS}$ decreases $v_{GD}$ and so decreases the electron concentration at the drain end. This will cause a slight increase in the total drain resistance.

The result of all this is that the inversion layer (channel) resistance is non-linear – the resistance changes as the drain voltage changes.
The non-uniform electron concentration along the channel implies that the channel resistance is not uniform — it is greater at the drain end, where there are fewer electrons.

So, as the drain voltage increases, the channel resistance increases. The effect is that the I-V curve bends over at higher voltage.
So a more correct analysis of the current flow shows that the $i_D-v_{DS}$ relationship is not linear, but parabolic. (See EE 332.)

$$i_D = \left( \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \right) \left[ 2 (v_{GS} - V_T) v_{DS} - v_{DS}^2 \right]$$

Study exercise: Show that this reduces to the linear equation for small values of $v_{DS}$.

Notation: To save on ink and time, define the constant in front as:

$$K = \frac{1}{2} \mu_n C_{ox} \frac{W}{L}$$

The units of this “MOS current parameter” are $A/V^2$. Or more typically $mA/V^2$. (Check it.)

Note: This use of $K$ is not standard. Different textbooks will handle the current parameter differently.
current saturation

Now take the final step: If $v_{DS}$ is increased sufficiently, we reach a point where $v_{GD} < V_T$. The electron concentration should disappear at the drain end!

The usual terminology is to say that the channel is “pinched off” at the drain end, but this is slightly misleading. If the channel were truly pinched off, then the electron concentration would go to zero there and the current would necessarily have to go to zero, also.

Instead, it might be better to say that the channel is “pinched down”. The electron concentration goes to some minimum value, but never really shuts off. In effect, the flowing current works to hold the channel open, even though the drain voltage seems high enough to truly pinch off the channel.
The mathematical analysis of what happens in a FET at pinch-off is a fairly complicated problem in electromagnetics – definitely beyond our current capabilities. (Again, see EE 332.)

However, the end result on the device behavior is easy to grasp – the current saturates at the pinch down value. For higher values of $v_{DS}$, the current stays constant at a constant value.

The condition for the NMOS going reaching the “pinch-down” condition is $v_{GD} \leq V_T$. This can be re-expressed in terms of $v_{DS}$ and $v_{GS}$.

$$v_{DS} \geq v_{GS} - V_T$$

To determine the saturation current, we can insert the pinch-down condition given above into the $i_D-v_{DS}$ equation from the ohmic region of operation. Then the saturated current is

$$i_D(sat) = \left( \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \right) [v_{GS} - V_T]^2 = K [v_{GS} - V_T]^2$$

(Work out these details for yourself.)
NMOS $i_D$-$V_{DS}$ curves

Fix $v_{GS}$ and vary $v_{DS}$. Plot the drain current.

For small $v_{DS}$ ($v_{DS} < v_{GS} - V_T$), the NMOS will be in the ohmic region.

At higher $v_{DS}$ ($v_{DS} \geq v_{GS} - V_T$), the channel pinches down and the current saturates.

NMOS with $K = 0.5$ mA/V$^2$ and $V_T = 2$ V.

If $v_{GS} < V_T$, there is no inversion layer and the NMOS is off ($i_D = 0$).
Another useful plot is $i_D$ vs. $v_{GS}$, with $v_{DS}$ fixed. If $v_{DS}$ is kept large so that the NMOS does not go into the ohmic region, the curve is essentially a plot of the saturation equation.

In a slight modification, we can plot $i_D^{1/2} = K^{1/2}(v_{GS} - V_T)$. From this graph, we can immediately pick out the two important NMOS parameters: $V_T$ and $K$. 

![Graph of NMOS characteristics](image)
Summary of NMOS equations

\[ V_T, \quad K = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \]

\[ i_D = 0 \] (at least at DC)

\[ v_{GS} < V_T \quad \text{off} \quad i_D = 0 \]

\[ v_{GS} \geq V_T \quad \text{on} \quad v_{DS} < v_{GS} - V_T \quad \text{ohmic or linear} \]

\[ i_D = \left( \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \right) \left[ 2 (v_{GS} - V_T) v_{DS} - v_{DS}^2 \right] \]

\[ v_{DS} \geq v_{GS} - V_T \quad \text{saturation} \]

\[ i_D = K [v_{GS} - V_T]^2 \]