The MOS Transistor in Weak Inversion

In this section we will explore the behavior of the MOS transistor in the subthreshold regime where the channel is weakly inverted. This will allow us to model transistors operating with small gate voltages, where the strong inversion model erroneously predicts zero current.

Remember, the strong inversion MOSFET model makes the assumption that the inversion charge $Q_I$ goes to zero when the gate voltage drops below the threshold voltage. We saw that this is not quite true. Below threshold, the channel charge drops exponentially with decreasing gate voltage.
This exponential relationship becomes clearer if we redraw the above figure with a logarithmic $y$ axis:

We can think of weak inversion as the region where $Q_I$ is an exponential function of gate voltage, strong inversion as the region where $Q_I$ is a linear function of gate voltage, and moderate inversion as a transition region between the two.

In weak inversion, the inversion layer charge is much less than the depletion region charge:

$$Q_I \ll Q_B \text{ in weak inversion}$$

Since the substrate is weakly doped, $Q_B$ is small, and there is not enough charge in the channel to generate a significant electric field to pull electrons from the source to the drain. Current flows by diffusion, not drift.

We shall once again consider our fluid model of a transistor. Let’s take another look at a transistor operating in the subthreshold regime:
Clearly, this model is insufficient to account for weak inversion since the channel charge is zero. We have to add another degree of realism to our model to account for subthreshold current flow: water vapor.

Electrons, like water molecules, can be excited by thermal energy into higher energy levels. Although most water molecules have a potential energy at or below the level of the liquid in the source and drain tanks, some water molecules have gained enough energy (thermally) to rise above this level – as vapor. Similarly, a small fraction of electrons in the source and drain acquire significantly more energy than the majority of charge carriers in the conduction band.

The density of water vapor above a liquid follows a decaying exponential with height (i.e., potential energy). Similarly, air pressure drops exponentially with altitude. Electrons in a solid obey Fermi-Dirac statistics which leads to this exponential distribution according to energy. This is similar to the Maxwell-Boltzmann statistics obeyed by atoms in a gas. Let’s add water vapor to our fluid model:
Now it’s clear that the inversion charge in the channel, while small, is an exponential function of the barrier height. The barrier height represents the surface potential $\psi_s$. As we mentioned before, in weak inversion the surface potential is flat - it does not change over the length of the channel. The surface potential can be modeled fairly accurately by considering the capacitive divider between the oxide capacitance $C_{ox}$ and the depletion capacitance $C_{dep}$:

$$G_s V_\kappa \psi_s =$$

where kappa - the gate coupling coefficient - represents the coupling of the gate to the surface potential:

$$\kappa = \frac{C_{ox}}{C_{ox} + C_{dep}}$$
The depletion capacitance stays fairly constant over the subthreshold region, and kappa is usually considered to be constant, although it increases slightly with gate voltage. In modern CMOS processes, kappa ranges between 0.6 and 0.8. It can have slightly different values for pMOS and nMOS devices. A good, all-around approximation for kappa (unless another value is given) is:

\[ \kappa \approx 0.7 \]

- Some texts use \( n \) or \( \zeta \) (zeta) instead of \( \kappa \), where \( n = \zeta = (1/\kappa) \approx 1.4 \).

Now let’s return to the fluid model for \( V_{DS} > 0 \):

The important parameter is the concentration of carriers at channel level (and above). Since the source tank is higher than the drain tank, the carrier concentration is higher where the source meets the channel than where the drain meets the channel. Electrons diffuse from the source to the drain.

The charge concentration (at channel level) in the source \((x = 0)\) and the drain \((x = L)\) is given by:

\[ |Q'_{T0}| \propto \exp\left(\frac{V_s - \kappa V_G}{U_T}\right) \]

\[ |Q'_{Tl}| \propto \exp\left(\frac{V_D - \kappa V_G}{U_T}\right) \]

Where \( U_T \) is the thermal voltage:

\[ U_T = \frac{kT}{q} \approx 26 \text{mV at room temperature} \]

(We use “U” instead of “V” for voltage to avoid confusion with the threshold voltage \( V_T \).)
From our previous discussion of diffusion, we know that particle motion is proportional to the concentration gradient. The concentration of electrons decreases linearly from the source to the drain (i.e., concentration gradient is constant), so we can write an expression for the drain current:

$$I_D = -WD_n \frac{(Q'_{t_0} - Q'_{t_L})}{L} = -\frac{W}{L} \mu_n U_T (Q'_{t_0} - Q'_{t_L})$$

This leads us (details omitted) to the expression for drain current in a subthreshold MOSFET:

$$I_D = I_0 \frac{W}{L} e^{\frac{-V_S}{U_T}} \left( \frac{V_S}{U_T} - e^{-\frac{V_D}{U_T}} \right)$$

or, writing it larger:

$$I_D = I_0 \frac{W}{L} e^{\frac{-\kappa V_G}{U_T}} \left[ \exp\left(\frac{-V_S}{U_T}\right) - \exp\left(-\frac{V_D}{U_T}\right) \right]$$

where $I_0$ is a process-dependant constant. For nFETs,

$$I_{0n} = \frac{2\mu_n C_{ox} U_T^2}{\kappa} \exp\left(-\frac{-\kappa V_{T_{0n}}}{U_T}\right)$$

Typical values of $I_{0n}$ range from $10^{-15} \text{A}$ to $10^{-12} \text{A}$.

We can rearrange terms and rewrite the expression for drain current as:

$$I_D = I_0 \frac{W}{L} e^{\frac{-\kappa V_G - V_S}{U_T}} \left[ 1 - \exp\left(-\frac{-V_{DS}}{U_T}\right) \right]$$

Notice that when $\exp(-V_{DS}/U_T) << 1$, the last term is approximately equal to one, and can be ignored. This occurs (to within 2%) for $V_{DS} > 4U_T$, since $e^{-4} \approx 0.018$. The expression for drain current then simplifies to:

$$I_D = I_0 \frac{W}{L} e^{\frac{-\kappa V_G - V_S}{U_T}} \text{ for } V_{DS} > 4U_T \text{ (saturation)}$$

At room temperature, $4U_T \approx 100 \text{mV}$, an easy value to remember. It is quite easy to keep a subthreshold MOSFET in saturation, and the $V_{DS}$ required to do so does not depend on $V_{GS}$ as is the case above threshold. This is very advantageous for low-voltage designs.
This saturation behavior is easy to understand if we look at our fluid model. If the drain tank is significantly lower than the source tank, the concentration of carriers in the drain at channel level will be much, much small than the concentration of carriers in the source at channel level, so the exact level doesn’t matter.

Another difference between subthreshold and above threshold operation is the way $I_D$ changes as we increase $V_{GS}$. In a weakly-inverted FET, the current increases exponentially. In a strongly-inverted FET, the current increases quadratically (square law). This can be understood by looking at a plot of $I_D$ vs. $V_{GS}$ in two ways: with a linear $I_D$ axis and with a logarithmic $I_D$ axis:
For a pFET, we have to consider the gate, drain, and source potentials with respect with the well potential \( V_W \). Unlike the substrate, the well will not be at ground, so we need to write it in explicitly:

\[
I_D = I_{0p} \frac{W}{L} \exp\left(\frac{\kappa(V_W - V_G)}{U_T}\right) \cdot \left[\exp\left(-\frac{(V_W - V_S)}{U_T}\right) - \exp\left(-\frac{(V_W - V_D)}{U_T}\right)\right]
\]

\[
I_{0p} = \frac{2\mu_p C_{ox}^e U_T^2}{\kappa} \cdot \exp\left(-\frac{\kappa V_{T0P}}{U_T}\right)
\]

Typical values of \( I_{0p} \) range from \( 10^{-19} \) A to \( 10^{-14} \) A.
In saturation:

\[
I_D = I_0 \frac{W}{L} \exp\left(\frac{\kappa \left(V_{w} - V_G\right) - \left(V_{w} - V_S\right)}{U_T}\right) \quad \text{for } |V_{DS}| > 4U_T
\]

This can be rewritten as:

\[
I_D = I_0 \frac{W}{L} \exp\left(-\frac{\kappa V_{GS} + (1 - \kappa)V_{WS}}{U_T}\right)
\]

Where \((1 - \kappa)\) is the back-gate coefficient. Notice that the body effect is explicit in the weak-inversion model without having to add a "fudge factor" like the variable threshold voltage in the strong inversion model.

The transconductance of a subthreshold MOSFET is easily derived:

**In weak inversion:**

\[
g_m = \frac{\kappa I_D}{U_T}
\]

Subthreshold MOSFETs behave similarly to bipolar junction transistors (BJTs). The collector current of an \(n\!p\!n\) bipolar transistor exhibits an exponential dependence on base-to-emitter voltage:

\[
I_C = I_s \exp\left(\frac{V_{BE}}{U_T}\right)
\]

A bipolar transistor has a transconductance of \(g_m = I_C/U_T\), which is equivalent to the expression for a subthreshold MOSFET if we set \(\kappa = 1\).

Of course, a MOSFET doesn’t pull any current through its gate like a bipolar transistor pulls through its base. This can make circuit design much easier.
MOSFET Operation in Weak and Moderate Inversion

R.R. Harrison

Moderate Inversion

A transistor does not switch immediately from an exponential, weak-inversion behavior to a quadratic, strong-inversion behavior. There is a smooth transition between the two extremes where drift and diffusion generate the current with neither effect dominating. Modeling this area is extremely difficult, but the behavior is easily understood as a hybrid between weak- and strong-inversion behavior.

The boundaries between weak, moderate, and strong inversion can be approximated in terms of voltages or currents. Of course, the boundaries between the areas are fuzzy, with no clean breaks. Here are two approximations sometimes used:

\[
\begin{array}{ll}
V_{GS} > V_T + 100\text{mV} & \text{strong inversion} \\
V_T + 100\text{mV} > V_{GS} > V_T - 100\text{mV} & \text{moderate inversion} \\
V_{GS} < V_T - 100\text{mV} & \text{weak inversion}
\end{array}
\]

(This assumes that the threshold voltage \( V_T \) has been adjusted for the body effect, if necessary.)
It is more common to design circuits with bias currents in mind. Following is a good estimate of the inversion boundaries in terms of drain current:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_D &gt; 10I_S$</td>
<td>strong inversion</td>
</tr>
<tr>
<td>$10I_S &gt; I_D &gt; 0.1I_S$</td>
<td>moderate inversion</td>
</tr>
<tr>
<td>$I_D &lt; 0.1I_S$</td>
<td>weak inversion</td>
</tr>
</tbody>
</table>

where $I_S$ is called the moderate inversion characteristic current:

$$I_S = \frac{2\mu'C_{ox}U_T^2 W}{\kappa} \frac{W}{L}$$

Typical values of $I_S$ range from 100nA to 500nA for nFETs with $W/L = 1$, and 40nA to 120nA for pFETs with $W/L = 1$. Of course, for large $W/L$ ratios, the weak inversion region can extend well into the microamp range.

**Example:** In AMI’s 0.5µm CMOS process, $\mu_nC_{ox} = 116\mu A/V^2$ and $\mu_pC_{ox} = 38\mu A/V^2$. Estimate the boundaries between weak, moderate, and strong inversion (in terms of drain currents) for nMOS and pMOS transistors with $W/L = 1$ and $W/L = 100$. Assume $\kappa = 0.7$.

*nFETs:*

$$I_{Sn} = \frac{2\mu_nC_{ox}'U_T^2 W}{\kappa} \frac{W}{L} = 220nA \text{ for } W/L = 1$$

- So for $W/L = 1$, the boundary between weak and moderate inversion is around 22nA, and the boundary between moderate and strong inversion is around 2.2µA.
- For $W/L = 100$, the boundary between weak and moderate inversion is around 2.2µA, and the boundary between moderate and strong inversion is around 220µA.

*pFETs:*

$$I_{Sp} = \frac{2\mu_pC_{ox}'U_T^2 W}{\kappa} \frac{W}{L} = 73nA \text{ for } W/L = 1$$

- So for $W/L = 1$, the boundary between weak and moderate inversion is around 7.3nA, and the boundary between moderate and strong inversion is around 0.73µA.
- For $W/L = 100$, the boundary between weak and moderate inversion is around 0.73µA, and the boundary between moderate and strong inversion is around 73µA.
A More Accurate Model of Kappa

The subthreshold gate coupling coefficient $\kappa$ is rarely reported in process data. Sometimes, people will report the "subthreshold slope" $U_T/\kappa$. If it is not given, $\kappa$ can be calculated by:

$$\kappa = \left( 1 + \frac{\gamma}{2\sqrt{(1 + \alpha)\Phi_F}} \right)^{-1}$$

where

$$\gamma = \sqrt{\frac{2q\varepsilon N_{sub}}{C_{ox}'}}$$

$$\Phi_F = \frac{kT}{q} \ln \frac{N_{sub}}{n_i}$$

$$C_{ax}' = \frac{\varepsilon_{ax}}{t_{ax}}$$

The $\alpha$ parameter should be set between zero - for extreme weak inversion (near depletion mode) - and one - for the boundary between weak and moderate inversion - to account for the slight change in depletion capacitance. Usually, $\alpha = 0.5$ is a good value to use for general-purpose use.

So all we really need to know is the oxide thickness (to determine $C_{ax}'$) and the channel doping ($N_{sub}$) to estimate $\kappa$. In SPICE models, oxide thickness is called $TOX$ (units = m) and the channel doping is called $NSUB$ or $NCH$ (units = cm$^{-3}$).

Example: An nFET has a substrate doping level of $1.7 \times 10^{17}$ cm$^{-3}$ and an oxide thickness of 139Å. Compute $\kappa$, using $\alpha = 0.5$.

$t_{ax} = 1.39 \times 10^{-6}$ cm

$C_{ax}' = 0.248 \mu$F/cm$^2$

$\gamma = 0.960V^{1/2}$

$\Phi_F = 0.422V$

$\kappa = 0.62$
The EKV Model

In 1995, Enz, Krummenacher and Vittoz proposed a relatively simple MOSFET model valid in all regions of operation: weak, moderate, and strong inversion. This has come to be known as the EKV model. Their basic equation for drain current (in saturation) is given by

$$I_D = I_S \cdot \left[ \ln \left( 1 + \exp \left[ \frac{\kappa (V_G - V_{T0}) - V_S}{2U_T} \right] \right) \right]^2$$

where

$$I_S = \frac{2 \mu C' U_T^2}{\kappa} \cdot \frac{W}{L}$$

An expression for transconductance $g_m$ valid in all regions of operation is given by

$$g_m = \frac{k I_D}{U_T} \cdot G(I_D)$$

where

$$G(I_D) = \frac{1 - e^{-\sqrt{I_D / I_S}}}{\sqrt{I_D / I_S}}.$$

Note that $I_D / I_S$ is the inversion coefficient and that $G(I_D)$ approaches unity in weak inversion.
Alternate expression for $g_m$

The expression used in the previous section for computing transconductance has the significant disadvantage that it cannot be solved to find the inversion coefficient $I_D/I_S$ as a function of the required transconductance and drain current $I_D$. A simpler (and, it is claimed, more accurate) interpolation function for transconductance is given by

$$g_m = \frac{\kappa I_D}{U_T} \cdot G'(I_D)$$

where

$$G'(I_D) = \frac{1}{\sqrt{(I_D/I_S) + \frac{1}{2} I_D/I_S + 1}}$$

or

$$G'(I_D) = \frac{2}{1 + \sqrt{1 + 4(I_D/I_S)}}.$$

Either function seems to work well. The latter expression is a bit easier to calculate.
Circles represent measured data from a real transistor. Note that in the moderate inversion region, both strong and weak inversion models overestimate the true transconductance.