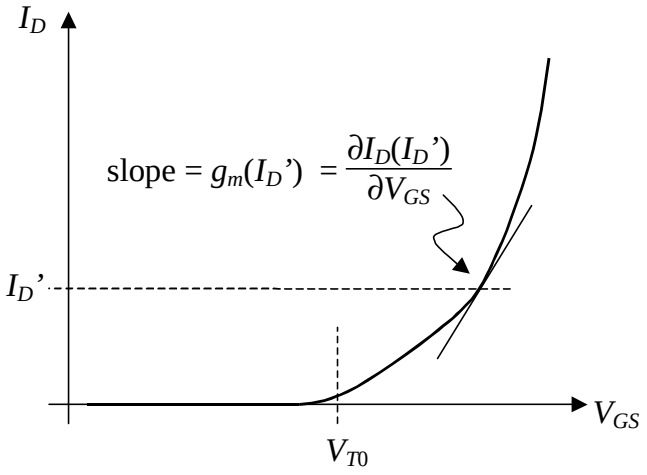
Small-Signal Parameters

The **large-signal equations** that describe MOSFET behavior are nonlinear. A circuit represents a system of nonlinear equations. Many times we are interested in building linear circuits, like filters that select certain frequencies but do not distort signals. How do we get linear behavior out of nonlinear devices?

If we zoom in on a nonlinear curve, it will eventually begin to look linear over a small range. This is similar to performing a Taylor expansion around a certain point on the curve, and then only keeping the first-order term. If we **bias** a transistor around an certain **operating point** and make only small perturbations about this point, we can approximate the device as a linear system. This is called **small-signal analysis**, because the amplitude of our signals must remain small for this analysis to remain valid.

The most important **small-signal parameter** of a MOSFET is its **transconductance**, g_m. If we make a *small* change in gate voltage, the transconductance tells us how much that affects the drain current:

$$g_m \equiv \frac{\partial I_D}{\partial V_{GS}}$$



We can express the transconductance of an above-threshold MOSFET in saturation in three equivalent ways:

In strong inversion:

$$g_m = \mu C'_{ox} \frac{W}{L} V_{eff}$$

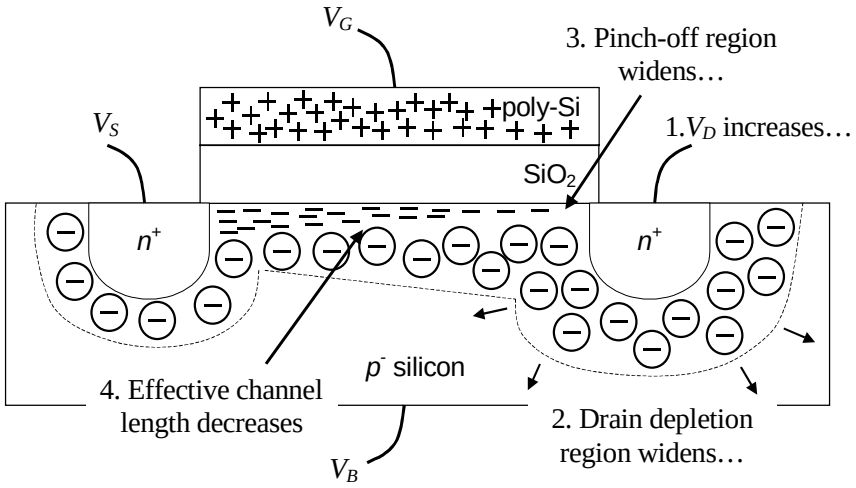
$$g_m = \sqrt{2\mu C'_{ox} \frac{W}{L} I_D}$$

$$g_m = \frac{2I_D}{V_{eff}}$$

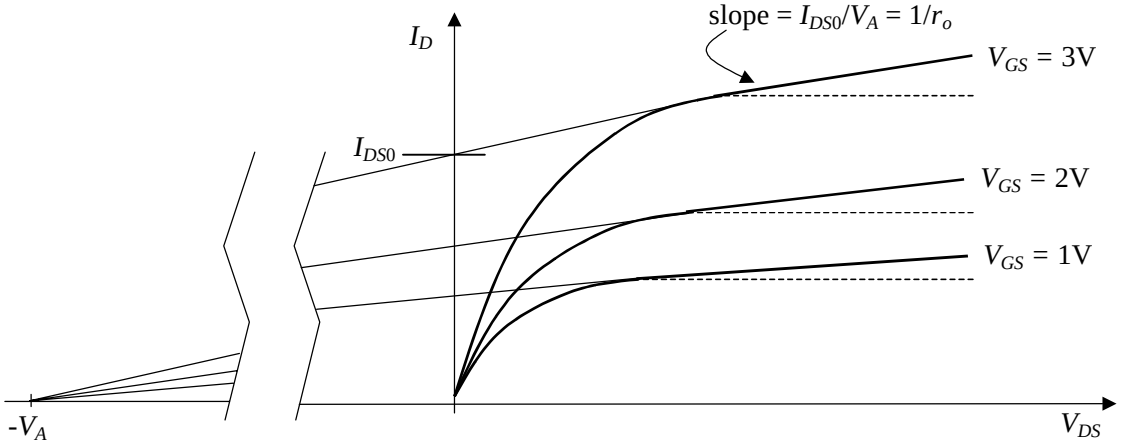
Channel-Length Modulation

When we discussed pinch-off and the saturation region, we said that after pinch-off occurred, the drain current was constant. (Remember the waterfall in the fluid model?) Well, that's not quite true due to an effect called **channel-length modulation**.

As we increase the drain voltage in the saturation region, the drain depletion region widens. This has the effect of widening the pinch-off region and thus shrinking the channel by a small amount.



Since the current in a MOSFET is proportional to W/L, a shrinking channel length increases the current through the device. In "long-channel" devices (which usually means at least 3 times the minimum length allowed by the process), the slight increase in current as the drain voltage increases is nearly linear. If we plot I_D vs. V_{DS} for several values of V_{GS} and extrapolate all the current traces backwards, they tend to converge at the same point on the V_{DS} axis. This voltage is called the **Early voltage** (V_A), and is the characteristic voltage in the first-order model of channel length modulation.



We can thus modify our expression for current in the saturation region by adding an extra term:

$$I_D = \frac{1}{2}\mu_n C'_{ox}\frac{W}{L}(V_{GS}-V_T)^2 \cdot \left(1+\frac{V_{DS}}{V_A}\right)$$

We can define a small-signal parameter called **output resistance** to describe this effect:

$$r_o \equiv \frac{\partial V_{DS}}{\partial I_D} = \frac{V_A}{I_D}$$

Since longer channels are affected less severely by drain depletion region widening, they show behavior closer to that of an ideal current source:

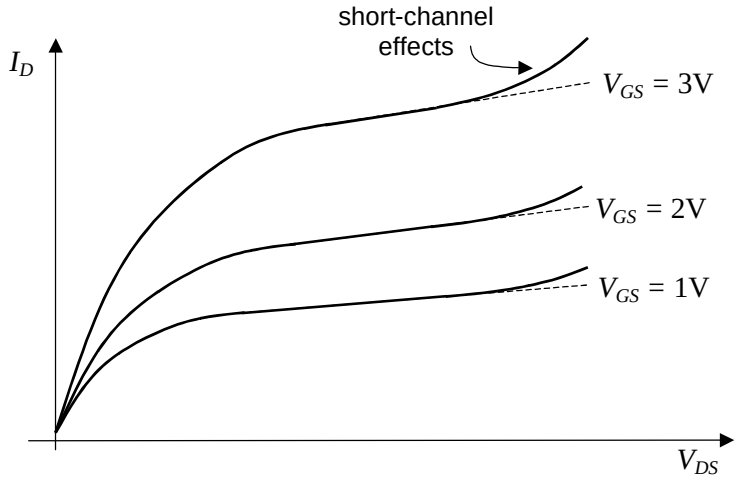
$$V_A, r_o \propto L \text{ for long-channel devices}$$

Some people use the **channel length modulation coefficient** λ instead of V_A:

$$I_D = \frac{1}{2}\mu_n C'_{ox}\frac{W}{L}(V_{GS}-V_T)^2 \cdot (1+\lambda V_{DS})$$

where $\lambda = 1/V_A$.

In modern **short-channel devices** (lengths near the minimum length allowed by the process), the effect is more severe. Several additional effects contribute to drain current increasing faster than linearly. One of these effects is called **drain-induced barrier lowering**, or **DIBL** ("dibble"). In very short devices, the drain can act as an additional gate ("side gate?"), coupling to the channel through the drain depletion capacitance. This can lead to dramatic increases in drain current (and dramatic decreases in r_o) as the drain voltage is increased.



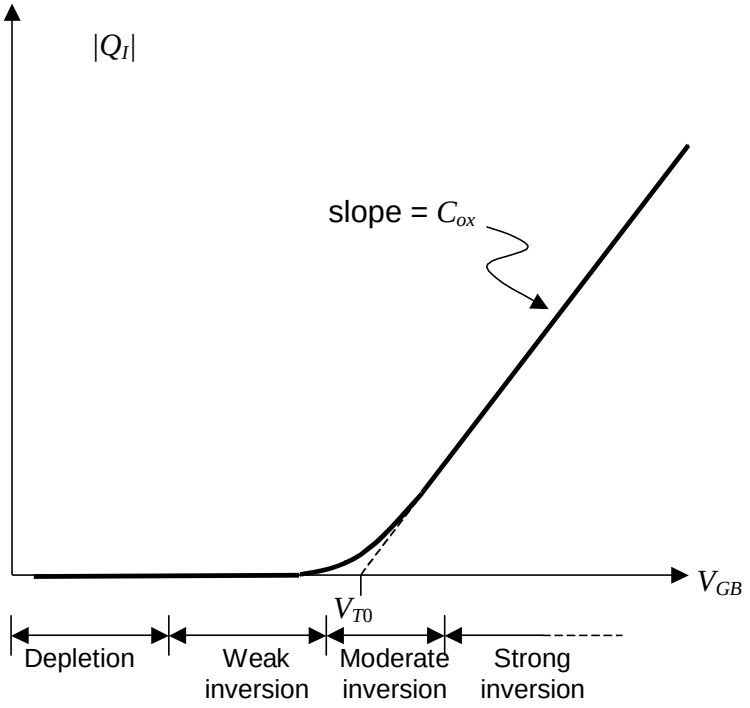
Most models of channel-length modulation are **empirical models**. They basically fit the observed behavior with first-order or higher-order models. The physical details – even in long-channel devices – are complex and difficult to model reliably.

In other words, the Early voltage is mostly a hack to fit the data.

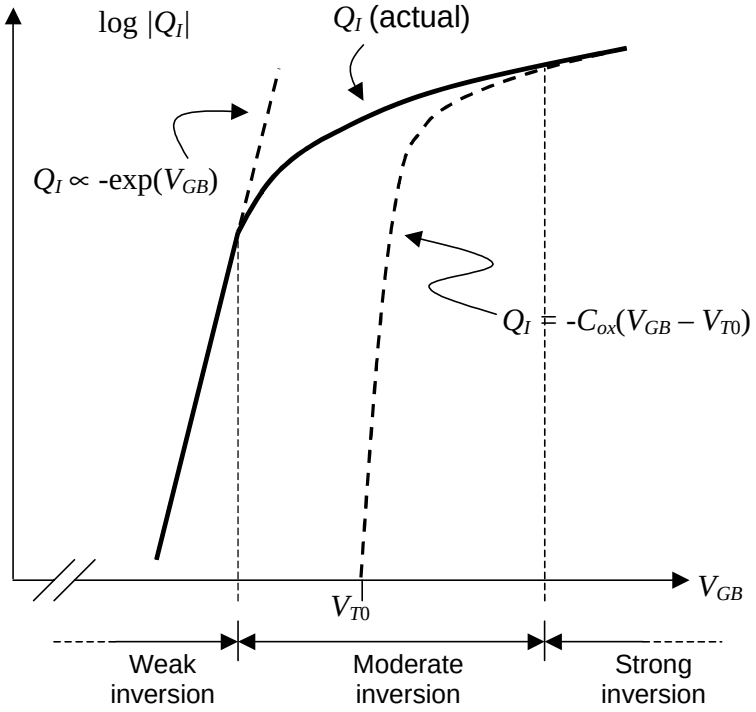
The MOS Transistor in Weak Inversion

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 drain currents of less than 1μA, 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:



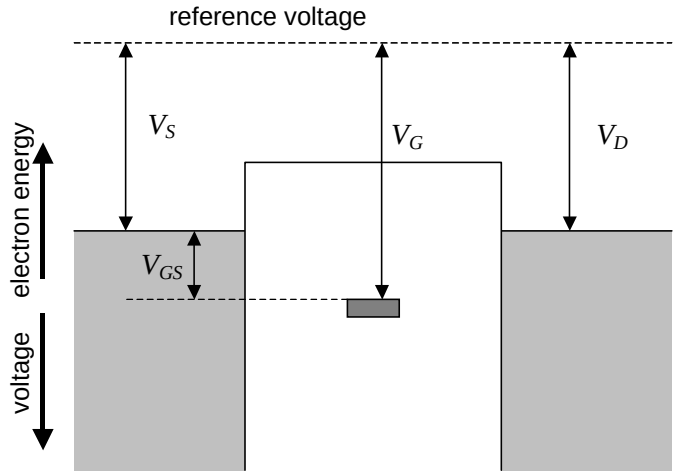
We can think of weak inversion as the region where Q_I is an exponential function of gate voltage, strong inversion as the region where Q_I is a linear function of gate voltage, and moderate inversion as a transition region between the two.

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

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

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.

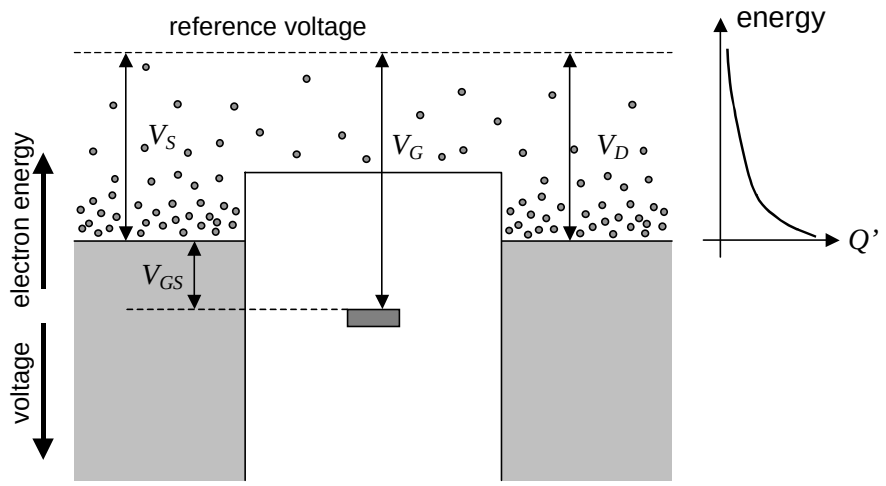
We shall once again consider our fluid model of a transistor. Let's take another look at a transistor operating in the subthreshold regime:



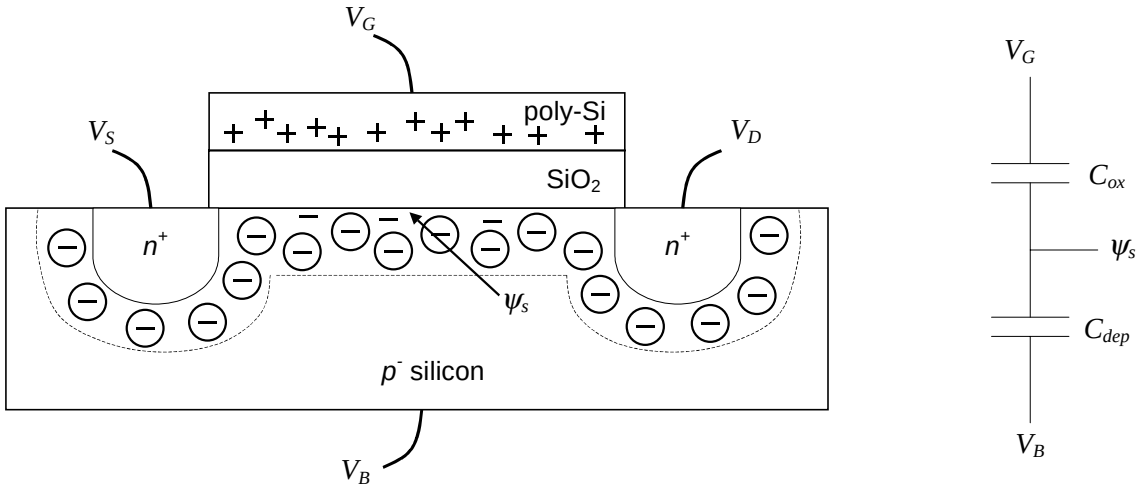
Clearly, this model is insufficient to account for weak inversion since the channel charge is zero. We have to add another degree of realism to our model to account for subthreshold current flow: water vapor.

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 ψ_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}:



Using the equation for a capacitive divider and assuming that $V_B = 0$, we find that:

$$\psi_s = \kappa V_G$$

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

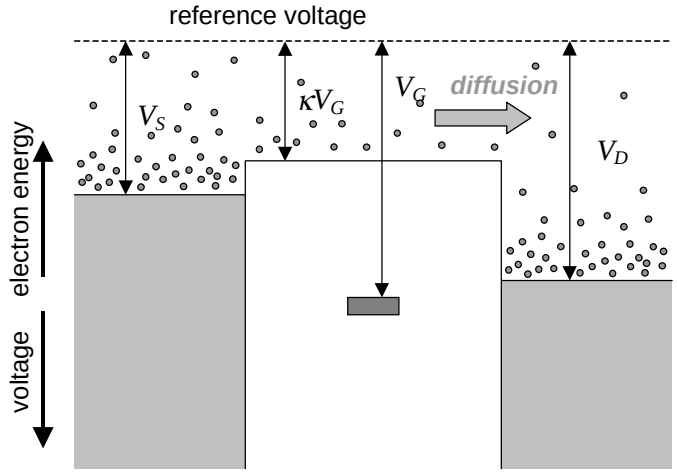
$$\kappa = \frac{C_{ox}}{C_{ox} + C_{dep}}$$

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:

$$\kappa \cong 0.7$$

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

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:

$$\left|Q'_{I0}\right| \propto \exp\left(\frac{V_S - \kappa V_G}{U_T}\right)$$

$$\left|Q'_{IL}\right| \propto \exp\left(\frac{V_D - \kappa V_G}{U_T}\right)$$

Where U_T is the thermal voltage:

$$U_T \equiv \frac{kT}{q} \cong 26\text{mV at room temperature}$$

(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:

$$I_D = -WD_n \frac{(Q'_{I0} - Q'_{IL})}{L} = -\frac{W}{L}\mu_n U_T (Q'_{I0} - Q'_{IL})$$

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

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

or, writing it larger:

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

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

$$I_{0n} \equiv \frac{2\mu_n C'_{ox} U_T^2}{\kappa} \cdot \exp\left(\frac{-\kappa V_{T0n}}{U_T}\right)$$

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

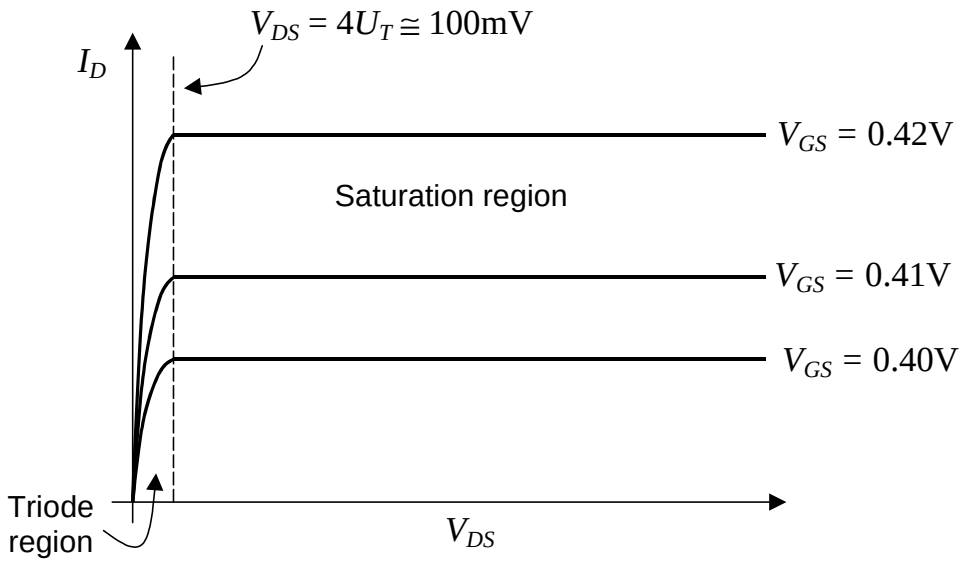
We can rearrange terms and rewrite the expression for drain current as:

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

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 $e^{-4} \cong 0.018$. The expression for drain current then simplifies to:

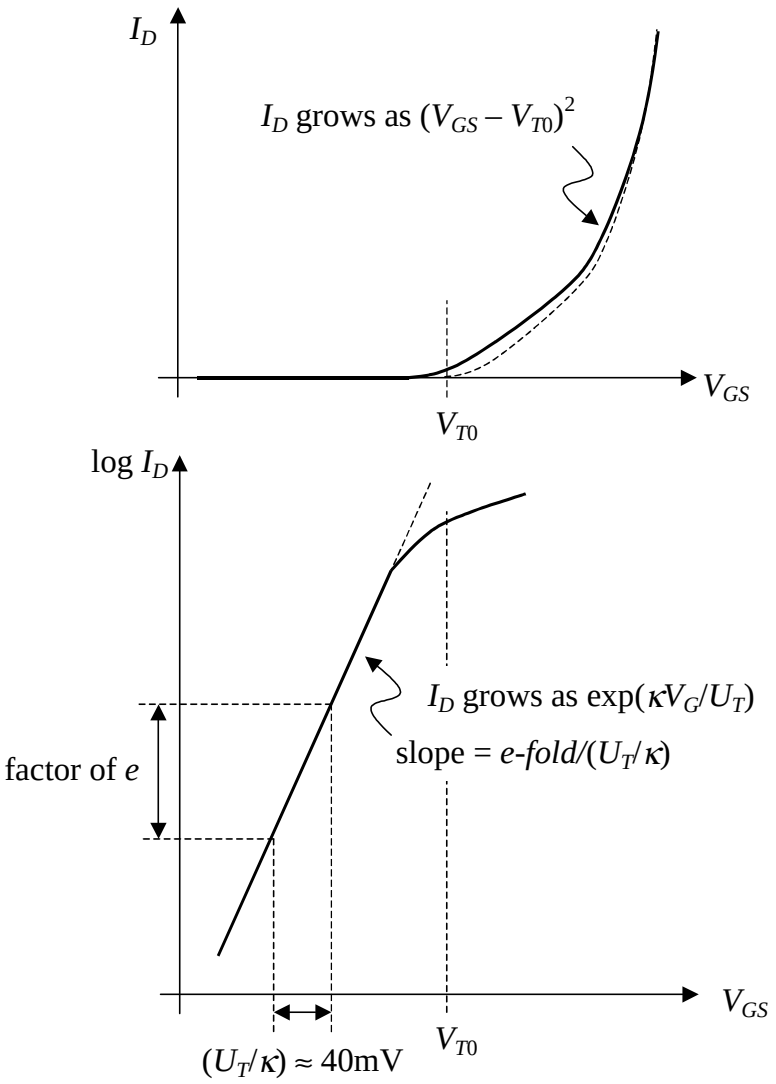
$$I_D = I_0 \frac{W}{L} \exp\left(\frac{\kappa V_G - V_S}{U_T}\right) \quad \text{for } V_{DS} > 4U_T \text{ (saturation)}$$

At room temperature, $4U_T \cong 100\text{mV}$, 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:

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

$$I_{0p} \equiv \frac{2\mu_p C'_{ox} U_T^2}{\kappa}\cdot\exp\left(\frac{-\kappa|V_{T0p}|}{U_T}\right)$$

Typical values of I_{0p} range from 10^{-19}A to 10^{-14}A.

In saturation:

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

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 *npn* bipolar transistor exhibits an exponential dependence on base-to-emitter voltage:

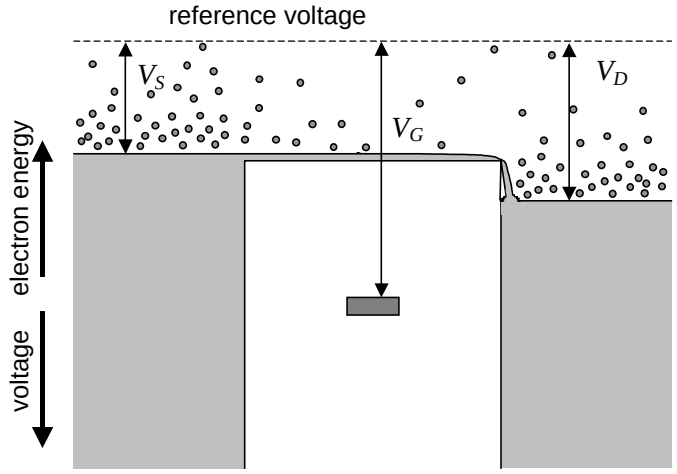
$$I_C = I_S \exp\left(\frac{V_{BE}}{U_T} \right)$$

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.

Moderate Inversion

A transistor does 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:

$V_{GS} > V_T + 100\text{mV}$	strong inversion
$V_T + 100\text{mV} > V_{GS} > V_T - 100\text{mV}$	moderate inversion
$V_{GS} < V_T - 100\text{mV}$	weak inversion

(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:

$I_D > 10I_S$	strong inversion
$10I_S > I_D > 0.1I_S$	moderate inversion
$I_D < 0.1I_S$	weak inversion

where I_S is called the moderate inversion characteristic current:

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

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, $\mu_n C_{ox}$' = 116μA/V^2 and $\mu_p C_{ox}$' = 38μA/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 κ = 0.7.

***n*FETs:**

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

- 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.

***p*FETs:**

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

- 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 κ is rarely reported in process data. Sometimes, people will report the "subthreshold slope" U_T/κ. If it is not given, κ can be calculated by:

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

where

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

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

$$C'_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$$

The α 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_{ox}') 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 x $10^{17}cm^{-3}$ and an oxide thickness of 139Å. Compute κ, using $\alpha = 0.5$.

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

C_{ox}' = 0.248μF/cm^2

γ = 0.960$V^{1/2}$

Φ_F = 0.422V

κ = 0.62