MOS Transistors

We can extend the analysis we did of the MOS capacitor system to the MOS transistor by considering adding source and drain regions doped with the opposite impurity type from the substrate.

![Diagram of MOS Transistor]

As before $V_G$ controls the charges in the channel region.

$V_G < V_{th}$:

The structure consists of two back-to-back diodes and only reverse leakage current flows between the source and drain.

$V_G > V_{th}$:

Inversion layer exists. This inversion layer forms a conducting channel between the source and drain allowing current to flow.

$V_{th}$ is unchanged from that calculated for the MOS capacitor.

$$V_{th} = \Phi_{MS} - \frac{Q_{ss}}{C_{ox}^i} - \frac{1}{C_{ox}^i} \int_0^{x_{ox}} \frac{x}{x_{ox}} \rho(x) \, dx - 2\phi_F \pm \frac{\sqrt{2q\epsilon_s N_B |2\phi_F|}}{C_{ox}^i}$$

(1)

where $N_B$ is the substrate doping and the term for depletion charge is $+$ for n-channel (p-substrate) devices and $-$ for p-channel (n-substrate) devices.
Current - Voltage Characteristics

1. The depletion region around the drain is larger than around the channel because the voltage \( V_D \) applied to the drain reverse biases the drain to substrate junction.

2. The potential along the channel varies from \( V_S = 0 \) to \( V_D \) between the source and drain.

3. The channel inversion charge \( Q'_i \) and the bulk depletion charge \( Q'_B \) are functions of distance along the channel since the surface potential varies along the channel.

The voltage drop across any elemental length of the channel is given by

\[
dV = I_D dy = \frac{I_D dy}{W \mu_n^* Q'_i(y)}
\]

(2)

directly from the expression for the channel conductance of the MOS capacitor.

At any point the total charge in the silicon is

\[
Q'_i(y) = Q'_i(y) + Q'_B(y)
\]

(3)

Recall that

\[
V_G = V_{FB} - \frac{Q'_i}{C_{ox}} + (\phi_s - \phi_F)
\]

(4)

where \( V_{FB} \) includes \( \Phi_{MS} \) and oxide charges.
Combining (3) and (4),

\[ Q'_i(y) = - (V_G - V_{FB} - (\phi_s(y) - \phi_F)) C_{ox}' - Q'_B(y) \tag{5} \]

Since the surface is inverted, the surface potential will differ from the bulk potential by approximately \(2\phi_F\) plus any reverse bias which exists between the channel and the substrate (due to \(V_D\) or substrate bias)

\[ \phi_s(y) - \phi_F \equiv V_{CB}(y) - 2\phi_F \tag{6} \]

We also know from our analysis of the MOS capacitor that

\[ Q'_B(y) = \pm q N_B x_{d_{max}}(y) = \pm \sqrt{2K_s \varepsilon_0 q N_B |V_{CB}(y) - 2\phi_F|} \tag{7} \]

For p-substrate (n-channel), as we move from \(S \rightarrow D\), \(V_{CB}(y)\) increases due to \(IR\) drop in channel.

- \(x_{d_{max}}\) increases as we move towards the drain.
- \(|Q'_B|\) increases as we move towards the drain.

Substituting (6) and (7) into (5),

\[ Q'_i(y) = - (V_G - V_{FB} - V_{CB}(y) + 2\phi_F) C_{ox}' \pm \sqrt{2K_s \varepsilon_0 q N_B |V_{CB}(y) - 2\phi_F|} \tag{8} \]

Note that

\( (V_G - V_{FB}) > 0, \ V_{CB}(y) > 0, \ \phi_F < 0, \ Q'_B < 0 \) for n-channel device

\( (V_G - V_{FB}) < 0, \ V_{CB}(y) < 0, \ \phi_F > 0, \ Q'_B > 0 \) for p-channel device

Since \(Q'_i\) is a function of channel voltage, Equation (2) can be rearranged and integrated from the source \((y = 0, V_{CH} = 0)\) to the drain \((y = L, V_{CB} = V_D)\).

\[ \int_0^{V_D} W \mu_n' Q'_i(V) dV = \int_0^L I_D dy \]
If equation (8) is used for $Q'_t$, we obtain

$$I_D = \frac{W}{L} \mu_n' C_{ox}' \left\{ \left[ V_G - V_{FB} + 2\phi_F - \frac{V_D}{2} \right] V_D \pm \frac{2}{3} \sqrt{2\varepsilon_s q N_B} \left[ \left[ V_D - 2\phi_F \right]^3 - \left[ 2\phi_F \right]^3 \right] \right\}$$

(9)

If we had assumed $Q'_B(y)$ was just a constant given by its value without an applied drain voltage (ignoring the influence of channel voltage on $Q_B$), we would have found

$$I_D \approx \frac{W}{L} \mu_n' C_{ox}' \left[ V_G - V_{th} - \frac{V_D}{2} \right] V_D$$

(10)

where $V_{th}$ includes the effects of $\phi_F$, $\phi_{MS}$, oxide charges and $Q'_B(y = 0)$.

Note that for a typical device, the approximate expression (10) works well for low $V_D$, but predicts too high a current at larger $V_D$. This makes sense since some of the gate charge must support the extra depletion charge and therefore the inversion layer charge is correspondingly reduced, an effect which was ignored in the approximation.

- Both equations (9) and (10) are only valid when an inversion layer exists all the way across the channel from the source to the drain.
- As $V_D$ increases, the voltage between the gate and the channel near the drain decreases until it is less than $V_{th}$, pinching off the channel. This occurs when

$$V_D > V_G - V_{th} = V_{Dsat}$$

(11)

When $V_D$ exceeds the saturation voltage, a channel no longer exist all the way to the drain. This is known as the saturation or "pinch-off" region.
$V_D > V_{D\text{sat}}$

- Electrons will travel along the inversion layer and then become injected into the depletion region where the high $E$ field pulls them into the drain.

- Further increase of $V_D$ does not change $I$ significantly.

$I_D \approx \text{constant for } V_D > V_{D\text{sat}}$

The boundary between the linear and saturation regions is described by

$$V_D = V_G - V_{th}$$

More accurately, we can solve equation (8) for $Q'_1(y = L) = 0$ (no inversion layer),

$$V_{D\text{sat}} = V_G - V_{FB} + 2\phi_F \pm \frac{K_s \varepsilon_0 q N_B}{C_{\varepsilon z}^2} \left[ 1 - \sqrt{1 + \frac{2C_{\varepsilon z}^2 |V_G - V_{FB}|}{K_s \varepsilon_0 q N_B}} \right]$$

(12)
MOSFET Current versus Voltage with Linearized Depletion Charge

The inversion charge is given by:

$$Q'_I(y) = -[V_{GB} - V_{FB} - (V_{CB}(y) - 2\phi_F)]C'_{ox} - Q'_B(y).$$  \hspace{1cm} (1)

Rather than using

$$Q'_B(y) = \mp qN_B x_{d_{max}}(y) = \mp \sqrt{2K_s\epsilon_0 q N_B |V_{CB}(y) - 2\phi_F|},$$  \hspace{1cm} (2)

we can instead approximate $Q'_B(y)$ by its first order Taylor series around $V_{CB} = V_{SB}$:

$$Q'_B(y) \approx Q'_B(0) + V_{CS} (dQ'_B(y)/dV_{CB})_{V_{SB}}$$

$$= \mp \sqrt{2K_s\epsilon_0 q N_B |V_{SB}(y) - 2\phi_F|} - V_{CS} \sqrt{\frac{K_s\epsilon_0 q N_B}{2|V_{SB}(y) - 2\phi_F|}}$$  \hspace{1cm} (3)

Thus,

$$Q'_I(y) = -[V_{GS} - V_{th} - (1 + \alpha)V_{CS}(y)]C'_{ox},$$  \hspace{1cm} (4)

where

$$\alpha = \frac{1}{C'_{ox}} \sqrt{\frac{K_s\epsilon_0 q N_B}{2|V_{SB} - 2\phi_F|}} = \frac{C'_d}{C'_{ox}}$$ \hspace{1cm} (5)

Integrating from $y = 0$ to $L$ and $V_{GS} = 0$ to $V_{DS}$,

$$I_{DS} = \pm \frac{W}{L} \mu' C'_{ox} \left[ V_{GS} - V_{th} - (1 + \alpha) \frac{V_{DS}}{2} \right] V_{DS}$$  \hspace{1cm} (6)

where $V_{th}$ includes the effects of $\phi_F$, $\phi_{MS}$, oxide charges and $Q'_B(y = 0)$.

For a typical device, the approximate expression (6) works well in capturing the effect of changing depletion charge. For large $V_{DS}$, a slightly smaller value of $\alpha$ based on expanding $V_{CB}$ around a value between $V_{DB}$ and $V_{SB}$ may be better.

Equation (6) is only valid when an inversion layer exists all the way across the channel from the source to the drain. As $V_{DS}$ increases, the voltage between the gate and the channel near the drain decreases until it is less than $V_{th}$, pinching off the channel. This occurs when

$$V_{DS} > \frac{V_{GS} - V_{th}}{1 + \alpha} = V_{DS_{sat}}.$$ \hspace{1cm} (7)

When $V_{DS}$ exceeds the saturation voltage, a channel no longer exist all the way to the drain. This is known as the saturation or “pinch-off” region. Additional voltage is dropped across a narrow depleted (not inverted) region near the drain and the drain current remains near its maximum value:

$$I_{DS_{sat}} = \pm \frac{W}{L} \mu' C'_{ox} \left[ \frac{(V_{GS} - V_{th})^2}{2(1 + \alpha)} \right]$$  \hspace{1cm} (8)
Linear Region Conductance

If $V_D \ll (V_G - V_{th})$, then (10) reduces to

$$I_D \approx \frac{W}{L} \mu_n' C_{ox}' (V_G - V_{th}) V_D$$

(13)

The device therefore looks like a voltage controlled resistor with a conductance given by

$$g_D = \frac{\partial I_D}{\partial V_D} \bigg|_{V_G} = \frac{1}{R_{on}} = \frac{W}{L} \mu_n' C_{ox}' (V_G - V_{th})$$

(14)

This is the same relationship which was derived for the MOS capacitor.
Saturation Region Transconductance

The transconductance or gain of the device is defined as

\[ g_m = \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D} \]  \hspace{1cm} (15) \]

Differentiating (9) with respect to \( V_G \) yields

\[ g_m = \frac{W}{L} \mu_n' C'_{oz} V_D \quad (V_D < V_{D_{sat}}) \]  \hspace{1cm} (16) \]

In the saturation region, \( I \) is approximately constant, so we can evaluate the transconductance at the edge of the saturation region and that will hold throughout the saturation region. We can therefore substitute \( V_{D_{sat}} \) as given in (12) into (15).

\[ g_{m_{sat}} = \frac{W}{L} \mu_n' C'_{oz} \left\{ V_G - V_{FB} + 2\phi_F \pm \frac{K_s \epsilon_0 q N_B}{C'_{oz}} \left[ 1 - \sqrt{1 + \frac{2C'_{oz} |V_G - V_{FB}|}{K_s \epsilon_0 q N_B}} \right] \right\} \]  \hspace{1cm} (17) \]

Making the same approximation we made in going from (9) to (10) (ignore change in \( Q_B' \) with \( y \))

\[ g_{m_{sat}} \approx \frac{W}{L} \mu_n' C'_{oz} (V_G - V_{th}) \]  \hspace{1cm} (18) \]

\[ R_{on} \approx \frac{1}{\frac{W}{L} \mu_n' C'_{oz} (V_G - V_{th})} \]

\[ g_m \approx \frac{1}{R_{on}} \]

The on resistance and the saturation transconductance (gain) are inversely related.
Threshold Control

It is usually desirable to make enhancement mode devices. \((V_{th}(n\text{-channel}) > 0, V_{th}(p\text{-channel}) < 0)\)

\[
p\text{-channel: } V_{th} = \Phi_{MS} - 2\phi_p - \frac{Q'_B}{C'_{oz}} - \frac{Q'_{ss}}{C'_{oz}} < 0
\]

\[
n\text{-channel: } V_{th} = \Phi_{MS} - 2\phi_p - \frac{Q'_B}{C'_{oz}} - \frac{Q'_{ss}}{C'_{oz}} ??
\]

In order to ensure that \(V_{thn} > 0\), need:

- \(|\phi_p|\) large \(\Rightarrow\) high doping \(\Rightarrow\) large \(C_s\), low junction \(BV\), low \(\mu_n'\)
- \(C'_{oz}\) small \(\Rightarrow\) thick oxide \(\Rightarrow\) low gain (\(g_m \propto C'_{oz}\))

Body Bias (Back Gate Bias)

In the analysis to this point we have assumed that \(V_B = 0\), i.e. \(V_{SB} = 0\). If this is not true and the substrate is reverse biased, i.e. \(V_{SB} > 0\) for an n-channel device, then \(|Q'_B|\) will increase.

\[
Q'_B = \mp \sqrt{2K_1\epsilon_0 q N_B |V_{SB} - 2\phi_F|}.
\]

(13)

The change in \(Q'_B\) due to \(V_{SB}\) is supported by a change in the gate voltage, i.e. a change in \(V_{th}\).

\[
V_{th} = \Phi_{MS} - \frac{Q'_{ss}}{C'_{oz}} - 2\phi_F \pm \sqrt{2\epsilon_0 N_B |V_{SB} - 2\phi_F| \frac{C'_{oz}}{C'_{oz}}}
\]

(14)

The equations we derived for I-V relations etc., are still valid provided we replace \(V_{th}\) by the above expression.

Therefore, the threshold voltage can be increased by reverse biasing the substrate. The disadvantage of this approach is the requirement for an additional voltage source.

This back gate bias is also used in modern MOS ICs to reduce junction capacitances.
Channel Implants

The most widespread approach used for control of threshold voltage in MOS transistors is the shallow-implantation of a surface layer of dopants in the channel.

If we assume that the implanted layer is very thin and close to the surface, then the region can be assumed to be depleted yet have no impact on the depletion region width (the electric field is changed only over a negligible distance).

Therefore the depletion charge is is changed from $Q'_B$ to $Q'_B + qN'$ where $N'$ is the implanted dose per unit area and

$$V_{th} = V_{FB} - 2\phi_F - \frac{Q'_B}{C_{ox}'} - \frac{qN'}{C_{oz}}$$ (15)

In order to increase the threshold voltage (as is required to make n-channel enhancement mode devices) acceptor impurities (−) must be implanted, while to reduce $V_{th}$ donor impurities are implanted.

If the implant profile is not actually of negligible depth, an additional voltage drop will occur across the implanted layer reducing the effectiveness of the implant on $V_{th}$. In that case, Poisson's equation would have to be used to determine the threshold shift (see text).

Implantation is an effective method of controlling the threshold voltage because the critical parameter, total dose, is very well controlled (\(\sim \pm 3\%\)).
The Universal Mobility Curve

\[ \mu(V_g, T_{ox}, N_{sub}, V_{sb}) = \mu(E_{eff}) \]

\[ \mu_{eff} = \frac{\mu_0}{1 + (E_{eff} / E_0)^n} \]

\[ E_{eff} = \frac{Q_{ap} + Q_{mv}}{\varepsilon_s} \]

\[ E_b = \frac{Q_{ap}}{\varepsilon_s} \]

\[ E_t = \frac{Q_{ap} + Q_{mv}}{\varepsilon_s} \]

\[ \frac{E_b + E_t}{2} = \frac{Q_{ap} + Q_{mv}}{2\varepsilon_s} = E_{eff} \]

Effective Field

\[ E_{eff} = \frac{Q_{ap} + Q_{mv}}{\varepsilon_s} \]

\[ Q_{mv} = C_{ox}(V_g - V_I) \]

\[ V_I = V_F + 2\varphi_a + \frac{Q_{ap}}{C_{ox}} \]

For \( n^+ \)/NMOS or \( p^+ \)/PMOS

\[ V_F + 2\varphi_a \approx 0, V_I \approx \frac{Q_{ap}}{C_{ox}} \]

therefore \( Q_{ap} = C_{ox} V_I \)

\( n^+ \)/PMOS

\[ E_{eff} = \frac{V_g + V_I - 2.3V}{6T_{ox}} \]

\[ E_{eff} = \frac{C_{ox} V_I + C_{ox} V_s - V_I}{2\varepsilon_s} \]

\[ E_{eff} = \frac{C_{ox} V_I + V_s - V_I}{2\varepsilon_s} \]

\[ E_{eff} = \frac{C_{ox} V_I + V_s - V_I}{2\varepsilon_s} \]

\[ V_F + V_I = \frac{V_s + V_I}{T_{ox} 2\varepsilon_s} \]

\[ n^+ \text{ poly} \]

\[ p^+ \text{ poly} \]
**Universal Surface Mobilities**

- Phonon Scattering
- Coulombic Scattering
- Surface Roughness Scattering

\[(V_g + V_t + 0.2V)/6T_{ox}\] can be shown to be the average vertical electric field in the inversion layer.

---

**Velocity Saturation**

Model A: \[ v = \frac{\mu_{eff} E}{1 + E/E_{sat}} \]

At \( E << E_{sat} \)  \[ v = \mu_{eff} E \]

At \( E >> E_{sat} \)  \[ v = \text{constant} = E_{sat} \mu_{eff} \]

Velocity overshoot
Channel Length Modulation

In the saturation region, as $V_D$ increases, the pinch-off point of the channel moves back towards the source (the drain depletion region expands). Therefore the current $I$ increases since it is $\propto 1/L_{\text{eff}}$. The depletion region expands as $\sqrt{V_D - V_{D_{\text{sat}}}}$ assuming constant doping. Provided the device has a channel length $L \gg \Delta x_d$ and provided $x_{oz} \ll \Delta x_d$, then the change in channel length is approximately

$$\Delta L \cong \sqrt{\frac{2\epsilon_s}{qN_B}} |V_D - V_{D_{\text{sat}}}|$$

(16)

The decrease in $L$ is responsible for an increase in $I$ in the saturation region, therefore a finite output impedance results. Channel width modulation is most pronounced in devices with lightly doped substrates and short channels.
Short Channel Effects

In devices with long channel lengths, the gate is completely responsible for depleting the semiconductor ($Q_B$). In very short channel devices, part of the depletion is accomplished by the drain and source depletion regions.

Therefore $V_{th} \downarrow$ as $L \downarrow$ since less gate voltage is required to support $Q_B$.

These effects are particularly pronounced in lightly doped substrates. Ion implanting a more heavily doped region under the gate (same type as substrate) can be used to minimize the threshold shifts.
The performance of the intrinsic MOS transistor may be expressed in terms of the following parameters.

- **Device gain** ($g_m$) \( \propto \frac{1}{x_{oz}} \)
- **Input capacitance** \( \propto \frac{WL}{x_{oz}} \)
- **Miller capacitance** \( \propto \frac{Wl_0}{x_{oz}} \)
- **Punch through voltage** \( \propto N_B L^2 \)
- **Threshold voltage** \( \propto \sqrt{N_B x_{oz}} \)

In order to maintain the same I-V characteristics as size is shrunk,

- $L, x_{oz}, x_j, l_0 \downarrow$ by $S$,
- $N_B \uparrow$ by $S$,

where $S$ = scaling factor.
<table>
<thead>
<tr>
<th>Device / Circuit Parameter</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensions : $x_{oz}, L, l_0, W, x_j$</td>
<td>$\frac{1}{S}$</td>
</tr>
<tr>
<td>Substrate Doping : $N_B$</td>
<td>$S$</td>
</tr>
<tr>
<td>Supply voltage : $V$</td>
<td>$\frac{1}{S}$</td>
</tr>
<tr>
<td>Supply current : $I$</td>
<td>$\frac{1}{S}$</td>
</tr>
<tr>
<td>Parasitic Capacitance : $\frac{W L e_{oz}}{x_{oz}}$</td>
<td>$\frac{1}{S}$</td>
</tr>
<tr>
<td>Gate delay : $\frac{V C}{I}$</td>
<td>$\frac{1}{S}$</td>
</tr>
<tr>
<td>Power dissipation : $J V$</td>
<td>$\frac{1}{S^2}$</td>
</tr>
<tr>
<td>Power delay product : $V^2 C$</td>
<td>$\frac{1}{S^3}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>1977</th>
<th>1980</th>
<th>1986</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layout density (gates/mm²)</td>
<td>200</td>
<td>450</td>
<td>1000</td>
</tr>
<tr>
<td>Power delay product (pJ)</td>
<td>1</td>
<td>0.2</td>
<td>0.01</td>
</tr>
<tr>
<td>Gate delay (ns)</td>
<td>1.0</td>
<td>0.2</td>
<td>0.01</td>
</tr>
<tr>
<td>Power supply voltage</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Channel length (µm)</td>
<td>3.5</td>
<td>2.0</td>
<td>0.7</td>
</tr>
<tr>
<td>$x_{oz}$ (Å)</td>
<td>700</td>
<td>400</td>
<td>100</td>
</tr>
</tbody>
</table>
Circuit Models

The MOS transistor may be modeled in the following manner, by inspection of its physical structure.

Of the elements in the model, only the gate to channel capacitance is essential; the rest are parasitic and degrade performance. Technology improvements are generally designed to reduce these parasitics.

In many cases, the equivalent circuit can be reduced to:

Note that many of the parameters in the model are voltage sensitive. Therefore accurate large signal modeling usually requires computer techniques.
Summary

1. MOS devices operate by controlling surface conductance.

2. Performance is greatly improved by shrinking dimensions (L, etc.). Therefore VLSI IC's will likely use this type of device because:
   
   (a) Small device size ⇒ high performance.
   
   (b) Small device size ⇒ lots of devices in a given chip area.
   
   (c) MOS devices are physically smaller for the same design rules than bipolar devices.

3. Small MOS devices (L < 1μm) require very tight process controls to achieve reproducible characteristics.