The MOS Transistor

CROSS-SECTION of NMOS Transistor
Cross-Section of CMOS Technology
MOS transistors
Types and Symbols

NMOS Enhancement

PMOS Enhancement

NMOS Depletion

NMOS with Bulk Contact

Channel
MOSFET Static Behavior

With drain and source grounded, and $V_{GS} = 0$, both back-to-back (sub-source, sub-drain) junctions have 0V bias and are OFF.
MOSFET Static Behavior

Positive voltage applied to the gate ($V_{GS} > 0$)

- The gate and substrate form the plates of a capacitor.
- Negative charges accumulate on the substrate side (repels mobile holes)
- A depletion region is formed under the gate (like $pn$ junction diode)

[Diagram of MOSFET with labels for gate (G), source (S), drain (D), and substrate (B).]
As the $V_{GS}$ increases, the surface under the gate undergoes *inversion* to *n*-type material. This is the beginning of a phenomenon called *strong inversion*.

Further increases in $V_{GS}$ do not change the width of the depletion layer, but result in more electrons in the thin inversion layer, producing a continuous *channel* from source to drain.
The value of $V_{GS}$ where strong inversion occurs is called the *Threshold Voltage*, $V_T$, and has several components:

- The flat-band voltage, $V_{FB}$, represents the built-in voltage offset across the MOS structure.
- $V_B$ represents the voltage drop across the depletion layer at inversion.
- $V_{ox}$ represents the potential drop across the gate oxide.

$$V_T = V_{FB} + V_B + V_{ox}$$
The Threshold Voltage

\[ V_T = \phi_{ms} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{SS}}{C_{ox}} - \frac{Q_I}{C_{ox}} \]

Where:

- \( \phi_F \) is the Fermi potential (\( \sim -0.3 \)V for p-type substrates)
- Cox is the gate oxide capacitance
- \( V_{SB} \) is the substrate bias voltage
- \( V_{T0} \) is \( V_T \) at \( V_{SB} = 0 \)

Note:

- \( V_T \) is positive for NMOS transistors and negative for PMOS

Workfunction Difference

Body Effect Coefficient

Surface Charge

Depletion Layer Charge

Implants
Linear (triode) Region

- When $V_{GS} - V_{DS} > V_T$, the channel can exist from drain to source
- Transistor behaves like voltage controlled resistor
Saturation region

- When $V_{GS} - V_{DS} \leq V_T$, there is not enough voltage near the drain to maintain the channel (channel is *pinched off*).
- Electrons arriving from source are injected into depletion region and accelerated towards drain by electric field.
- Transistor behaves like voltage-controlled current source.
Current-Voltage Relations

**Linear Region:** \( V_{DS} \leq V_{GS} - V_T \)

\[
I_D = k'_n \frac{W}{L} (V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2}
\]

with

\[
k'_n = \mu_n C_{ox} = \frac{\mu_n \varepsilon_{ox}}{t_{ox}}
\]

**Process Transconductance Parameter**

**Saturation Mode:** \( V_{DS} \geq V_{GS} - V_T \)

Channel Length Modulation

\[
I_D = \frac{k_n W}{2L} (V_{GS} - V_T)^2 (1 + \lambda V_{DS})
\]
I-V Relation

NMOS Enhancement Transistor: \( W = 100 \ \mu m, \ L = 20 \ \mu m \)
A model for manual analysis

\[ V_{DS} > V_{GS} - V_T \quad \text{Saturation} \]

\[ I_D = \frac{k_n' W}{2 L} (V_{GS} - V_T)^2 (1 + \lambda V_{DS}) \]

\[ V_{DS} < V_{GS} - V_T \quad \text{Linear} \]

\[ I_D = k_n' \frac{W}{L} (V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \]

with

\[ V_T = V_{T0} + \gamma \left( \sqrt{-2 \Phi_F + V_{SB}} - \sqrt{-2 \Phi_F} \right) \]
Dynamic Behavior of MOS Transistor

\[ C_{GS} \quad C_{GD} \quad C_{SB} \quad C_{GB} \quad C_{DB} \]

TODAY!
Dynamic Behavior of MOS Transistor

- MOSFET is a majority carrier device (unlike \textit{pn} junction diode)
- Dynamic response solely dependant on the time taken to (dis)charge the capacitances between device ports and interconnections
- Capacitances originate from three sources:
  - basic MOS \textit{structure} (layout of the device)
  - charge present in the \textit{channel}
  - \textit{depletion regions} of the reverse-biased \textit{pn}-junctions of drain and source
- Capacitances associated with the \textit{structure} are \textit{constant} for the device
- All other capacitances are \textit{non-linear} and vary with applied voltage
MOS Structure Capacitances

Gate Capacitance

• Gate isolated from channel by gate oxide

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

• \( t_{ox} \) small as possible

• Results in gate capacitance \( C_g \)

\[ C_g = C_{ox} \cdot WL \]
Gate Capacitance

CROSS-SECTION of NMOS Transistor

Gate Oxide

Source

Polysilicon

Gate

Drain

n+

n+

p-substrate

Field-Oxide (SiO₂)

p+ stopper

Bulk Contact
The Gate Capacitance

(a) Top view.

(b) Cross-section

\[ C_{gate} = \frac{\varepsilon_{ox}}{t_{ox}} WL \]
The Gate Capacitance

Gate Capacitance has two components
• due to channel charge (non-linear)
• due to topology

Capacitance due to topology
• Source and drain extend below the gate oxide by \( x_d \) (*lateral diffusion*)
• Effective length of the channel \( L_{\text{eff}} \) is shorter than the drawn length (designed length) of channel by factor of \( 2x_d \)
• Cause of parasitic capacitance (overlap capacitance, \( C_{gsO} \)) between gate and source (drain)
• \( C_{gsO} \) strictly linear and has fixed value
The Gate Capacitance

Overlapping Capacitance

Channel Capacitance

(a) Top view.

(b) Cross-section

\[ C_{\text{gate}} = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} WL \]
The Channel Capacitance

Channel Capacitance has three components
• capacitance between gate and source, $C_{gs}$
• capacitance between gate and drain, $C_{gd}$
• capacitance between gate and bulk region, $C_{gb}$

Channel Capacitance values
• non-linear, depends on operating region
• averaged to simplify analysis
The Channel Capacitance

Different distributions of gate capacitance for varying operating conditions

<table>
<thead>
<tr>
<th>Operation Region</th>
<th>$C_{gb}$</th>
<th>$C_{gs}$</th>
<th>$C_{gd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cutoff</td>
<td>$C_{oxWLeff}^*$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Triode</td>
<td>0</td>
<td>$C_{oxWLeff}/2$</td>
<td>$C_{oxWLeff}/2$</td>
</tr>
<tr>
<td>Saturation</td>
<td>0</td>
<td>$(2/3)C_{oxWLeff}$</td>
<td>0</td>
</tr>
</tbody>
</table>

Most important regions in digital design: saturation and cut-off
Diffusion (junction) Capacitance

\[ C_{\text{diff}} = C_{\text{bottom}} + C_{\text{sw}} = C_j \times \text{AREA} + C_{jsw} \times \text{PERIMETER} \]
\[ = C_j L_S W + C_{jsw} (2L_S + W) \]
Junction Capacitance

\[ C_j = \frac{C_{j0}}{(1 - V_D/\phi_0)^m} \]
Diffusion Capacitance

- Caused by the reverse-biased source-bulk and drain-bulk $pn$ junctions (same as junction capacitance $C_j$ for diode)
- Non-linear, depends on source (drain) to bulk voltage
- Has two components:

**Bottom plate** junction capacitance
- formed between source and substrate
- $C_{\text{bottom}} = C_j W L_s$, $C_j$ is junction capacitance per unit area

**Side-wall** junction capacitance
- formed between source and channel-stop implant
- larger than **bottom-plate** contribution
- $C_{\text{sw}} = C'_{\text{jsw}} x_j (W+2 * L_s)$
Diffusion Capacitance

\[
C_{\text{diff}} = C_{\text{bottom}} + C_{\text{sw}} = C_j \times \text{AREA} + C_{jsw} \times \text{PERIMETER}
\]

\[
= C_j L_s W + C_{jsw} (2L_s + W)
\]
Linearizing the Junction Capacitance

Replace non-linear capacitance by large-signal equivalent linear capacitance which displaces equal charge over voltage swing of interest

\[ C_{eq} = \frac{\Delta Q_j}{\Delta V_D} = \frac{Q_j(V_{high}) - Q_j(V_{low})}{V_{high} - V_{low}} = K_{eq} C_{j0} \]

\[ K_{eq} = \frac{-\phi_0^m}{(V_{high} - V_{low})(1 - m)} \left[ (\phi_0 - V_{high})^{1-m} - (\phi_0 - V_{low})^{1-m} \right] \]
Capacitive Device Model

\[
C_{GS} = C_{gs} + C_{gsO}
\]
\[
C_{GD} = C_{gd} + C_{gdO}
\]
\[
C_{GB} = C_{gb}
\]
\[
C_{SB} = C_{Sdiff}
\]
\[
C_{DB} = C_{Ddiff}
\]

Work example 2.9
The Sub-Micron MOS Transistor

• Actual transistor deviates substantially from model
• Deviates more for channel lengths below 1um
• Channel length becomes comparable to other device params. Ex: depth of drain and source junctions
• Referred to as a short-channel device
• Influenced heavily by secondary effects
The Sub-Micron MOS Transistor

Secondary Effects:

• Threshold Variations

• Parasitic Resistances

• Velocity Saturation and Mobility Degradation

• Sub-threshold Conduction

• Latchup
Threshold Variations

Part of the region below gate is depleted by source and drain fields, which reduces the threshold voltage for short channel. Similar effect is caused by increase in Vds.
Parasitic Resistances

\[ R_{S,D} = \frac{L_{S,D}}{W} R_{SQ} + R_C \]

Silicide the bulk region

Polysilicon gate

Drain contact

increase \( W \)
Variations in I-V Characteristics

- *Velocity saturation* and *mobility degradation* cause the short-channel device to deviate considerably from ideal expressions (eq 2.47 and 2.51).
- The velocity of the carriers is proportional to the electric field up to a point. When electric field reaches a critical value, $E_{\text{sat}}$, the velocity of the carriers tend to saturate.
- When the channel length decreases, only a small $V_{DS}$ is needed for saturation.
- Causes a *linear dependence* of the saturation current wrt the gate voltage (in contrast to *squared dependence* of long-channel device).
- Current drive cannot be increased by decreasing L.
- Reduced L decreases the mobility of the carriers due to the vertical component of the electric field (decreases $I_D$).
Variations in I-V Characteristics

\( \nu_{sat} = 10^7 \) 

\begin{align*}
\text{Constant velocity} & \\
E_{sat} = 1.5 & \quad E (\text{V/\mu m})
\end{align*}

(a) Velocity saturation

\begin{align*}
\mu_n & \\
\mu_{n0} & \quad E_t (\text{V/\mu m})
\end{align*}

(b) Mobility degradation
Velocity Saturation

(a) $I_D$ as a function of $V_{DS}$

(b) $I_D$ as a function of $V_{GS}$ (for $V_{DS} = 5 \text{ V}$).

Linear Dependence on $V_{GS}$
I-V Relation

(a) $I_D$ as a function of $V_{DS}$

(b) $\sqrt{I_D}$ as a function of $V_{GS}$
(for $V_{DS} = 5V$).

NMOS Enhancement Transistor: $W = 100 \mu m$, $L = 20 \mu m$
Sub-Threshold Conduction

\[ \ln(I_D) (\text{A}) \]

\[ V_G (\text{V}) \]

Subthreshold exponential region

Linear region

\[ V_T \]
Latchup

(a) Origin of latchup

(b) Equivalent circuit
SPICE MODELS

Level 1: Long Channel Equations - Very Simple

Level 2: Physical Model - Includes Velocity Saturation and Threshold Variations

Level 3: Semi-Emperical - Based on curve fitting to measured devices

Level 4 (BSIM): Emperical - Simple and Popular

Berkeley Short-Channel IGFET Model
# MAIN MOS SPICE PARAMETERS

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Symbol</th>
<th>SPICE Name</th>
<th>Units</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPICE Model Index</td>
<td></td>
<td>LEVEL</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Zero-Bias Threshold Voltage</td>
<td>VT0</td>
<td>VT0</td>
<td>V</td>
<td>0</td>
</tr>
<tr>
<td>Process Transconductance</td>
<td>k'</td>
<td>KP</td>
<td>A/V2</td>
<td>2E-5</td>
</tr>
<tr>
<td>Body-Bias Parameter</td>
<td>g</td>
<td>GAMMA</td>
<td>V0.5</td>
<td>0</td>
</tr>
<tr>
<td>Channel Modulation</td>
<td>1</td>
<td>LAMBDA</td>
<td>1/V</td>
<td>0</td>
</tr>
<tr>
<td>Oxide Thickness</td>
<td>tox</td>
<td>TOX</td>
<td>m</td>
<td>1E-7</td>
</tr>
<tr>
<td>Lateral Diffusion</td>
<td>xd</td>
<td>LD</td>
<td>m</td>
<td>0</td>
</tr>
<tr>
<td>Metallurgical Junction Depth</td>
<td>xj</td>
<td>XJ</td>
<td>m</td>
<td>0</td>
</tr>
<tr>
<td>Surface Inversion Potential</td>
<td>2</td>
<td>fF</td>
<td></td>
<td>PHI</td>
</tr>
<tr>
<td>Substrate Doping</td>
<td>NA,ND</td>
<td>NSUB</td>
<td>cm-3</td>
<td>0</td>
</tr>
<tr>
<td>Surface State Density</td>
<td>Qss/q</td>
<td>NSS</td>
<td>cm-3</td>
<td>0</td>
</tr>
<tr>
<td>Fast Surface State Density</td>
<td>NFS</td>
<td>cm-3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total Channel Charge Coefficient</td>
<td>NEFF</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Type of Gate Material</td>
<td>TPG</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Surface Mobility</td>
<td>m0</td>
<td>U0</td>
<td>cm2/V-sec</td>
<td>600</td>
</tr>
<tr>
<td>Maximum Drift Velocity</td>
<td>umax</td>
<td>VMAX</td>
<td>m/s</td>
<td>0</td>
</tr>
<tr>
<td>Mobility Critical Field</td>
<td>xcrit</td>
<td>UCRIT</td>
<td>V/cm</td>
<td>1.0E4</td>
</tr>
<tr>
<td>Critical Field Exponent in Mobility Degradation</td>
<td>UEXP</td>
<td>-</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Transverse Field Exponent (mobility)</td>
<td>UTRA</td>
<td>-</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
# SPICE Parameters for Parasitics

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Symbol</th>
<th>SPICE Name</th>
<th>Units</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source resistance</td>
<td>$R_S$</td>
<td>RS</td>
<td>Ω</td>
<td>0</td>
</tr>
<tr>
<td>Drain resistance</td>
<td>$R_D$</td>
<td>RD</td>
<td>Ω</td>
<td>0</td>
</tr>
<tr>
<td>Sheet resistance (Source/Drain)</td>
<td>$R_o$</td>
<td>RSH</td>
<td>Ω/$\omega$</td>
<td>0</td>
</tr>
<tr>
<td>Zero Bias Bulk Junction Cap</td>
<td>$C_{j0}$</td>
<td>CJ</td>
<td>F/m$^2$</td>
<td>0</td>
</tr>
<tr>
<td>Bulk Junction Grading Coeff.</td>
<td>$m$</td>
<td>MJ</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>Zero Bias Side Wall Junction Cap</td>
<td>$C_{jsw0}$</td>
<td>CJSW</td>
<td>F/m</td>
<td>0</td>
</tr>
<tr>
<td>Side Wall Grading Coeff.</td>
<td>$m_{sw}$</td>
<td>MJSW</td>
<td>-</td>
<td>0.3</td>
</tr>
<tr>
<td>Gate-Bulk Overlap Capacitance</td>
<td>$C_{gb0}$</td>
<td>CGBO</td>
<td>F/m</td>
<td>0</td>
</tr>
<tr>
<td>Gate-Source Overlap Capacitance</td>
<td>$C_{gs0}$</td>
<td>CGSO</td>
<td>F/m</td>
<td>0</td>
</tr>
<tr>
<td>Gate-Drain Overlap Capacitance</td>
<td>$C_{gdo}$</td>
<td>CGDO</td>
<td>F/m</td>
<td>0</td>
</tr>
<tr>
<td>Bulk Junction Leakage Current</td>
<td>$I_S$</td>
<td>IS</td>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>Bulk Junction Leakage Current Density</td>
<td>$J_S$</td>
<td>JS</td>
<td>A/m$^2$</td>
<td>1E-8</td>
</tr>
<tr>
<td>Bulk Junction Potential</td>
<td>$\phi_0$</td>
<td>PB</td>
<td>V</td>
<td>0.8</td>
</tr>
</tbody>
</table>
## SPICE Transistors Parameters

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Symbol</th>
<th>SPICE Name</th>
<th>Units</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drawn Length</td>
<td>L</td>
<td>L</td>
<td>m</td>
<td>-</td>
</tr>
<tr>
<td>Effective Width</td>
<td>W</td>
<td>W</td>
<td>m</td>
<td>-</td>
</tr>
<tr>
<td>Source Area</td>
<td>AREA</td>
<td>AS</td>
<td>m(^2)</td>
<td>0</td>
</tr>
<tr>
<td>Drain Area</td>
<td>AREA</td>
<td>AD</td>
<td>m(^2)</td>
<td>0</td>
</tr>
<tr>
<td>Source Perimeter</td>
<td>PERIM</td>
<td>PS</td>
<td>m</td>
<td>0</td>
</tr>
<tr>
<td>Drain Perimeter</td>
<td>PERIM</td>
<td>PD</td>
<td>m</td>
<td>0</td>
</tr>
<tr>
<td>Squares of Source Diffusion</td>
<td>NRS</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Squares of Drain Diffusion</td>
<td>NRD</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

See Ex 2.12 (HW problem)
Fitting level-1 model for manual analysis

Select $k'$ and $\lambda$ such that best matching is obtained @ $V_{gs} = V_{ds} = V_{DD}$
# Technology Evolution

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel length (μm)</td>
<td>0.4</td>
<td>0.3</td>
<td>0.25</td>
<td>0.18</td>
<td>0.13</td>
<td>0.1</td>
</tr>
<tr>
<td>Gate oxide (nm)</td>
<td>12</td>
<td>7</td>
<td>6</td>
<td>4.5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$V_{DD}$ (V)</td>
<td>3.3</td>
<td>2.2</td>
<td>2.2</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>$V_T$ (V)</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>NMOS $I_{Dsat}$ (mA/μm) (@ $V_{GS} = V_{DD}$)</td>
<td>0.35</td>
<td>0.27</td>
<td>0.31</td>
<td>0.21</td>
<td>0.29</td>
<td>0.33</td>
</tr>
<tr>
<td>PMOS $I_{Dsat}$ (mA/μm) (@ $V_{GS} = V_{DD}$)</td>
<td>0.16</td>
<td>0.11</td>
<td>0.14</td>
<td>0.09</td>
<td>0.13</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Read Section 2.6
Process Variations

Devices parameters vary between runs and even on the same die!

Variations in the process parameters, such as impurity concentration densities, oxide thicknesses, and diffusion depths. These are caused by non-uniform conditions during the deposition and/or the diffusion of the impurities. This introduces variations in the sheet resistances and transistor parameters such as the threshold voltage.

Variations in the dimensions of the devices, mainly resulting from the limited resolution of the photolithographic process. This causes ($W/L$) variations in MOS transistors and mismatches in the emitter areas of bipolar devices.

Read Section 2.5
Impact of Device Variations

Delay of Adder circuit as a function of variations in $L$ and $V_T$

Read Section 2.5