MOSFET: Introduction

- Metal oxide semiconductor field effect transistor (MOSFET) or MOS is widely used for implementing digital designs
  - Its major assets are:
    - Higher integration density, and
    - Relatively simple manufacturing process

- As a consequence, it is possible to realize $10^6-7$ transistors on an integrated circuit (IC) economically
For an n-channel MOS transistor (NMOS)

- Heavily doped n-type source and drain regions are implanted (diffused) into a lightly doped p-type substrate (body)
- A thin layer (approx. 50 Å) of silicon dioxide (SiO₂) is grown over the region between source and drain and is called thin or gate oxide

Gate oxide is covered by a conductive material, often polycrystalline silicon (polysilicon) and forms the gate of the transistor

MOS transistors are insulated from each other by thick oxide (SiO₂) and reverse biased p-n+ diode

Adding p+ field implant (channel stop implant) makes sure a parasitic MOS transistor is not formed

MOS Transistor as a switch

- \( V_{in} > V_T \): a conducting channel is formed between source and drain and current flows
- \( V_{in} < V_T \): the channel does not form and switch is said to be open
- \( V_{in} > V_T \) current is a function of gate voltage
NMOS, PMOS, and CMOS Technology

- In an NMOS transistor, current is carried by electrons (from source, through an n-type channel to the drain)
  - Different than diode where both holes and electrons contribute to the total current
  - Therefore, MOS transistor is also known as unipolar device

- Another MOS device can be formed by having p+ source and drain and n-substrate (PMOS)
  - Current is carried by holes through a p-type channel

- A technology that uses NMOS (PMOS) transistors only is called NMOS (PMOS) technology
  - In NMOS or PMOS technologies, substrate is common and is connected to +ve voltage, VDD (NMOS) or GND (PMOS)

In a complementary MOS (CMOS) technology, both PMOS and NMOS transistors are used

- NMOS and PMOS devices are fabricated in isolated region from each other (i.e., no common substrate for all devices)

- MOS transistor is a 4 terminal device, if 4th terminal is not shown it is assumed to be connected to appropriate voltage
Static Behavior

- Only the NMOS transistor is discussed, however, arguments are valid for PMOS transistor as well

**The threshold voltage**

- Consider the case where \( V_{gs} = 0 \) and drain, source and bulk are connected to ground

\[ V_{gs} = 0, \] \[ V_{sd} = 0, \] \[ V_{sb} = 0. \]

- Hence, high resistance between source and drain \( (10^7 \, \Omega) \)

**If now the gate voltage \( (V_{gs}) \) is increased, gate and substrate form plates of a capacitor with oxide as dielectric**

- +ve gate voltage causes +ve charge on gate and -ve charge on the substrate side
- In substrate it occurs in two steps (i) depletion of mobile holes, (ii) accumulation of -ve charge (inversion)
At certain $V_{GS}$, potential at the interface reaches a critical value, where surface inverts to n-type (start of strong inversion).

Further $V_{GS}$ increase does not increase the depletion width but increases electrons in the inversion layer.

Threshold Voltage

$$V_T = V_{TO} + \gamma \left[ \sqrt{-2\phi_F + V_{SB}} - \sqrt{-2\phi_F} \right]$$

Where

$$\gamma = \frac{\sqrt{2qEsiNA}}{C_{ox}}$$

$V_T$ is +ve for NMOS and -ve for PMOS devices.

Current-Voltage Relationship

When $V_{GS} > V_T$

Let at any point along the channel, the voltage is $V(x)$ and gate to channel voltage at that point is $V_{GS} - V(x)$.
If the $V_{GS} - V(x) > V_T$ for all $x$, the induced channel charge per unit area at $x$

$$Q_i(x) = -C_{ox}[V_{gs} - V(x) - V_T]$$

Current is given by

$$I_D = -\nu(x)Q_i(x)W$$

The electron velocity is given by

$$\nu_n = -\mu_n E(x) = \mu_n \frac{dV}{dx}$$

Therefore,

$$I_D dx = \mu_n C_{ox} W (V_{gs} - V - V_T) dV$$

Integrating the equation over the length $L$ yields

$$I_D = K'_n \frac{W}{L} \left[ (V_{gs} - V_T) V_{ds} \frac{V^2_{ds}}{2} \right]$$

or

$$I_D = K'_n \left[ (V_{gs} - V_T) V_{ds} \frac{V^2_{ds}}{2} \right]$$

$k'_n$ is known as the process trans-conductance parameter and equals

$$K'_n = \mu_n C_{ox} = \mu_n \frac{\varepsilon_{ox}}{t_{ox}}$$

If the $V_{GS}$ is further increased, then at some $x$, $V_{GS} - V(x) < V_T$ and that point the channel disappears and transistor is said to be pinched-off.

Close to drain no channel exists, the pinched-off condition in the vicinity of drain is $V_{GS} - V_{DS} <= V_T$.

Under these conditions, transistor is in the saturation region.

If a complete channel exists between source and drain, then transistors is said to be in triode or linear region.

Replacing $V_{DS}$ by $V_{GS} - V_T$ in the current equation we get, MOS current-voltage relationship in saturation region

$$I_D = \frac{K_n W}{2 L} (V_{gs} - V_T)^2$$
This equation is not entirely correct, the position of pinch-off point and hence the effective channel length is a function of \( V_{ds} \), a more accurate equation is given as

\[
I_D = \frac{K_n W}{2 L} (V_{gs} - V_T)^2 \left( 1 + \lambda V_{ds} \right)
\]

where \( \lambda \) is an empirical constant parameter called channel length modulation factor.

Dynamic Behavior

MOS transistor is a unipolar (majority carrier) device, therefore, its dynamic response is determined by time to (dis)charge various capacitances.

MOS capacitances

- **Gate oxide capacitance**: \( C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \) per unit area,
- For a transistor of width, \( W \) and length, \( L \), the \( C_g = WL \frac{\varepsilon_{ox}}{t_{ox}} \)
- From current equation it is apparent that \( C_{ox} \) should be high or gate oxide thickness should be small.
- Gate capacitance consists of several components.
- Source and drain diffusions extend below the thin oxide (lateral diffusion) giving rise to overlap capacitance.
MOSFET Overlap Capacitance

- Source and drain diffusions extend below the thin oxide (lateral diffusion) giving rise to overlap capacitance.
- $x_d$ is constant for a technology and this capacitance is linear and has a fixed value $C_{gsO} = C_{gdO} = C_{ox} x_d W = C_0 W$.

MOSFET Channel Capacitance

- Gate to channel capacitance consists of $C_{gs}$, $C_{gd}$ and $C_{gb}$ components.
- All these components are non-linear and their value depends on operation region of the device.
- Average/estimated values are used to simplify the analysis.

<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>$C_{ox} W L_{eff}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Triode</td>
<td>0</td>
<td>$C_{ox} W L_{eff}/2$</td>
<td>$C_{ox} W L_{eff}/2$</td>
</tr>
<tr>
<td>Saturation</td>
<td>0</td>
<td>$(2/3)C_{ox} W L_{eff}$</td>
<td>0</td>
</tr>
</tbody>
</table>
MOSFET: Junction Capacitances

- This component is contributed by the reverse biased source-bulk and drain-bulk pn-junctions
  - Depletion region (also known as diffusion) capacitance is non-linear and decreases as reverse bias is increased
  - **Bottom plate junction capacitance**: is formed by source \((N_D)\) and bulk regions \((N_A)\), \(C_{\text{bottom}} = C_jW_L_s\)
  - **Side wall junction capacitance**: is formed by source \((N_D)\) and \(p^+\) channel stop implant with doping \(N_A^+\)
  - Doping concentration is higher for channel stop implant hence the capacitance per unit area is also higher, \(C_{\text{sw}} = C_j\text{sw} x_j (W + 2L_s)\), since \(x_j\) is fixed for a technology \(C_{\text{sw}} = C_j\text{sw}(W + 2L_s)\)
  - Total diffusion capacitance \(C_{\text{diff}} = C_{\text{bottom}} + C_{\text{sw}} = C_j\cdot\text{area} + C_j\text{sw}\cdot\text{perimeter} = C_jW_L_s + C_j\text{sw}(W + 2L_s)\)

MOS: Capacitive Device Model

- \(C_{\text{GS}} = C_{gs} + C_{gsO}\)
- \(C_{\text{GD}} = C_{gd} + C_{gdO}\)
- \(C_{\text{GB}} = C_{gb}\)
- \(C_{\text{SB}} = D_{\text{diff}}\)
- \(C_{\text{DB}} = C_{\text{diff}}\)
Actual MOS Transistor: Short Channel Effects

- Realistic MOS transistor behaves differently from an ideal one owing to several factors
  - Owing to scaling, transistor channel length becomes comparable to other device parameters (e.g., junction depth, depletion width)
  - Assumptions such as, current flows only on surface, electric field is only in the direction of current flow, etc., are no longer true
  - Such a short channel device can not be adequately described by simple one dimensional model
  - Hence, a two dimensional model is widely used

Short Channel Effects: $V_T$ Variations

$$V_T = V_{TO} + \gamma [\sqrt{-2\phi_F + V_{SB}} - \sqrt{-2\phi_F}]$$

- Equation suggests $V_T$ is a function of technology and applied VSB
  - $V_T$ should be constant for all NMOS and all PMOS transistors
  - As dimensions are reduced, threshold potential becomes a function of $W$, $L$ and $V_{DS}$
  - Influence of Source and Drain over channel helps in depleting the charge from channel
  - As a consequence, a lower $V_T$ is required to cause strong inversion
  - **Drain induced barrier lowering**: as $V_{DS}$ increases, the depletion region width also becomes wider resulting in lower $V_T$
Hence $V_T$ is a function of operating voltage

- **Hot carrier effect**
  - As transistor dimensions are scaled, electric field strength is increased significantly
  - Higher electric field enables electrons (holes) to acquire high energy so that they can tunnel into thin oxide and modify the $V_T$
  - For NMOS $V_T$ is increased and for PMOS $V_T$ is reduced

- **Hot carrier damage remains a long term reliability threat**

---

**Source-Drain Resistance**

- With transistor scaling, junctions are made shallower & contacts windows are made smaller while their depth is increased

$$R_{S,D} = \frac{L_{S,D}}{W} R_0 + R_C$$

- Technology and design objective is to reduce source-drain resistance
- Often source drain regions are covered by titanium or tungsten (silicidation) to reduce the resistance
Variation in I-V Characteristics

- While developing the I-V equation we assumed that carrier velocity is proportional to $E$
  - However, as $E = E_{sat}$ (approx. $10^4$/micron), the carrier velocity saturates, as a consequence
    $$I_{DSAT} = \upsilon_{sat}C_{ox}W(V_{GS} - V_{DSAT} - V_T)$$
- In long channel MOSFET we also assumed that there is no vertical electric field
  - However, as transistor scales, $E_{vertical}$ can not be ignored
  - Carrier mobility is decreased as vertical electric field is increased

Sub-threshold Conduction

- MOS transistor partially conducts for $V_{gs} < V_T$
  - Known as sub-threshold conduction or weak-inversion conduction
  - Very small for long-channel ($10^{-12}$A/micron)
  - The inverse rate of decrease in current with respect to $V_{gs}$ is given by
    $$S = \left(\frac{d}{dV_{gs}}\ln(I_D)\right)^{-1} = \frac{kT}{q}\ln 10(1 + \alpha)$$
  - $kT\ln 10 = 60$ mV/decade and $\alpha$ is 0 for an ideal transistor
  - However, $\alpha$ is greater than 1 for real transistor making $S = 80$ mV/decade
Narrow Channel Effects

- Owing to small width, transistor exhibits non-ideal behavior

- LOCOS isolation
  - Depletion region is not limited to the area just under the thin oxide,
  - **If W is large**: part of the depletion region on the sides is small fraction and may be neglected
  - **If W is small**: gate also depletes the sides, hence larger $V_T$

![Diagram of LOCOS isolation](image)

- Shallow trench isolation
  - Field lines beyond gate region helps in depleting the channel & causing the inversion at lower gate voltage
  - Hence, lower $V_T$

![Diagram of Shallow trench isolation](image)
Spice Model for the MOS Transistor

- Several MOS models have been developed
  - Model complexity is a trade-off between accuracy and run time in simulator
  - In SPICE, model complexity is set by LEVEL parameter
  - Level 1: spice model is based on long channel MOS I-V equation; no longer used
  - Level 2: geometry, physics based; uses several short channel effects; complex and inaccurate; no longer used
  - Level 3: semi-empirical model
  - Level 4: empirical model based on extracted values from experimental data; widely used

- Several other models are available; virtually every semiconductor fab has some model development group

Technology Scaling & CMOS

- Ever since ICs were invented, dimensions are scaled to
  - Integrated more transistors in the same area
  - Allow higher operational speed

- Scaling has profound impact on many aspects of ICs

- Constant Voltage Scaling
  - All device dimensions are scaled by a factor S
  - Voltage (i.e., $V_{DD}$) after the scaling is same as before
  - This method of scaling is followed till 0.8 micron
  - However for lower geometries, higher electric field resulted in poor device reliability
Therefore, for advanced technologies today Constant Field Scaling is followed

- All dimensions including power supply is scaled by a factor $S$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Relation</th>
<th>CVS</th>
<th>CFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W, L, t_{ox}$</td>
<td>$1/S$</td>
<td>$1/S$</td>
<td></td>
</tr>
<tr>
<td>$V_{DD}, V_T$</td>
<td>$1$</td>
<td>$1/S$</td>
<td></td>
</tr>
<tr>
<td>Area</td>
<td>$WL$</td>
<td>$1/S^2$</td>
<td>$1/S^2$</td>
</tr>
<tr>
<td>$C_{ox}$</td>
<td>$t_{ox}$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>$C_L$</td>
<td>$C_{ox}W/L$</td>
<td>$1/S$</td>
<td>$1/S$</td>
</tr>
<tr>
<td>$k_n, k_p$</td>
<td>$C_{ox}W/L$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>$I_{av}$</td>
<td>$k_{n,p}V^2$</td>
<td>$S$</td>
<td>$1/S$</td>
</tr>
<tr>
<td>$J_{av}$</td>
<td>$I_{av}/Area$</td>
<td>$S^3$</td>
<td>$S$</td>
</tr>
<tr>
<td>$t_p$ (intrinsic)</td>
<td>$C_LV^2/I_{av}$</td>
<td>$1/S^2$</td>
<td>$1/S$</td>
</tr>
<tr>
<td>$P_{av}$</td>
<td>$C_LV^2/t_p$</td>
<td>$S$</td>
<td>$1/S^2$</td>
</tr>
<tr>
<td>PDP</td>
<td>$C_LV^2$</td>
<td>$1/S$</td>
<td>$1/S^3$</td>
</tr>
</tbody>
</table>

Concluding Remarks

- MOS transistor is the backbone of contemporary VLSIs
- Constant motivation for scaling
  - Scaling improves, Power, switching delay and PDP
- Experts predict slow down in scaling below 0.10 micron
  - Transistor characteristics are influenced by several short and narrow channel effects