MOSFETS: Gain & non-linearity

MOSFETs (metal-oxide-semiconductor field-effect transistors) are four-terminal voltage-controlled switches. Current flows between the diffusion terminals if the voltage on the gate terminal is large enough to create a conducting “channel”, otherwise the mosfet is off and the diffusion terminals are not connected.

Why are MOS devices King?
FETs as switches

The four terminals of a Field Effect Transistor (gate, source, drain and bulk) connect to conducting surfaces that generate a complicated set of electric fields in the channel region which depend on the relative voltages of each terminal.

CONDUCTION:
If a channel exists, a horizontal field will cause a drift current from the drain to the source.

INVERSION:
A sufficiently strong vertical field will attract enough electrons to the surface to create a conducting n-type channel between the source and drain.
"Linear" operating region

Larger $V_{GS}$ creates deeper channel which increases $I_{DS}$

Larger $V_{DS}$ increases drift current but also reduces vertical field component which in turn makes channel less deep. At some point, electrons are traveling as fast as possible through the channel ("velocity saturation") and the current stops growing linearly.

$V_{GS} > V_{TH}$

$0 < V_{DS} < V_{Dsat}$

$I_{DS}$ proportional to $\mu_0(W/L)$
Saturated operating region

When $V_{DS} = V_{GS} - V_{TH}$ the vertical field component is reduced and the channel is pinched-off. Electrons just keep traveling across depletion region…

This looks just like a FET with a channel length of $L' < L$. Shorter $L'$ implies greater $I_{DS}$. As $V_{DS}$ increases, $\delta L$ gets larger.

When $V_{GS} > V_{TH}$, the vertical field component is reduced and the channel is pinched-off. Electrons just keep traveling across depletion region…

$V_{DS} = V_{GS} - V_{TH}$

Increasing $V_{GS}$
**NFET Summary**

**Operating regions:**

**cut-off:**
\[ V_{GS} < V_{TH} \]
\[ V_{DS} < V_{Dsat} \]

**linear:**
\[ V_{GS} \geq V_{TH} \]
\[ V_{DS} < V_{Dsat} \]

**saturation:**
\[ V_{GS} \geq V_{TH} \]
\[ V_{DS} \geq V_{Dsat} \]

---

**Diagram:**

- **NFET symbol:**
  - G (Gate)
  - D (Drain)
  - S (Source)

- **Operating regions diagram:**
  - **Cut-off:**
    - \( V_{GS} < V_{TH} \)
  - **Linear:**
    - \( V_{GS} \geq V_{TH} \)
    - \( V_{DS} < V_{Dsat} \)
  - **Saturation:**
    - \( V_{GS} \geq V_{TH} \)
    - \( V_{DS} \geq V_{Dsat} \)

---

**Graph:**

- **Current (I_{DS})** vs. **Voltage (V_{GS})**
- **Current (I_{DS})** vs. **Voltage (V_{DS})**
FETs come in two flavors

By embedding p-type source and drain in a n-type substrate, we can fabricate a complement to the N-FET:

![Diagram of NMOS and PMOS transistors]

The use of both NFETs and PFETs – complementary transistor types – is a key to CMOS (complementary MOS) logic families.
PFET Summary

Operating regions:

**cut-off:**
\[ V_{GS} > V_{TH} \]

**linear:**
\[ V_{GS} \leq V_{TH} \]
\[ V_{DS} > V_{Dsat} \]

**saturation:**
\[ V_{GS} \leq V_{TH} \]
\[ V_{DS} \leq V_{Dsat} \]
**CMOS Inverter**

![CMOS Inverter Diagram]

- **Vin** and **Vout**
- **S = power supply**
- **I_{PU}** and **I_{PD}**

**Graphs:**

- **I_{PU} vs VOUT for PULLUP**
  - \(V_{IN} = 0v\)
  - \(V_{IN} = 1v\)
  - \(V_{IN} = 2v\)
  - \(V_{IN} = 3v\)
  - \(V_{IN} = 4v\)

- **I_{PD} vs VOUT for PULLDOWN**
  - \(V_{IN} = 5v\)
  - \(V_{IN} = 4v\)
  - \(V_{IN} = 3v\)
  - \(V_{IN} = 2v\)
  - \(V_{IN} = 1v\)
CMOS Inverter VTC

When both fets are saturated, small changes in $V_{in}$ produce large changes in $V_{out}$.

Steady state reached when $V_{out}$ reaches value where $I_{pu} = I_{pd}$.
Think Switches

**pullup:** make this connection when $V_{IN}$ near $0$ so that $V_{OUT} = V_{DD}$

**pulldown:** make this connection when $V_{IN}$ near $V_{DD}$ so that $V_{OUT} = 0$
Physical design of a CMOS gate is represented by a mask layout showing where material on each layer (ndiff, pdiff, poly, m1, m2, …) should be placed on the silicon wafer. Each manufacturing process has a set of design rules that determine minimum widths, spacings, overlaps, etc.

Use two narrow mosfets in parallel instead of one wide mosfet

SW: “scaled width” used in Process-independent design
Beyond Inverters: Complementary pullups and pulldowns

Now you know what the “C” in CMOS stands for!

We want complementary pullup and pulldown logic, i.e., the pulldown should be “on” when the pullup is “off” and vice versa.

<table>
<thead>
<tr>
<th>pullup</th>
<th>pulldown</th>
<th>F(A₁,...,Aₙ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>on</td>
<td>off</td>
<td>driven “1”</td>
</tr>
<tr>
<td>off</td>
<td>on</td>
<td>driven “0”</td>
</tr>
<tr>
<td>on</td>
<td>on</td>
<td>driven “X”</td>
</tr>
<tr>
<td>off</td>
<td>off</td>
<td>no connection</td>
</tr>
</tbody>
</table>

Since there’s plenty of capacitance on the output node, when the output becomes disconnected it “remembers” its previous voltage -- at least for a while. The “memory” is the load capacitor’s charge. Leakage currents will cause eventual decay of the charge (that’s why DRAMs need to be refreshed!).
What a nice $V_{oh}$ you have...

Thanks. It runs in the family...

---

CMOS complements

- Conducts when $V_{GS}$ is high
- Conducts when $V_{GS}$ is low

A
B

- Conducts when $A$ is high and $B$ is high: $A \cdot B$
- Conducts when $A$ is low or $B$ is low: $\overline{A} + \overline{B} = A \cdot B$

A
B

- Conducts when $A$ is high or $B$ is high: $A + B$
- Conducts when $A$ is low and $B$ is low: $\overline{A} \cdot \overline{B} = A + B$
A pop quiz!

What function does this gate compute?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

NAND
Here's another...

What function does this gate compute?

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

NOR
General CMOS gate recipe

Step 1. Figure out pulldown network that does what you want, *e.g.*, $F = A*(B+C)$ (What combination of inputs generates a low output)

Step 2. Walk the hierarchy replacing nfets with pfets, series subnets with parallel subnets, and parallel subnets with series subnets

Step 3. Combine pfet pullup network from Step 2 with nfet pulldown network from Step 1 to form fully-complementary CMOS gate.

But isn’t it hard to wire it all up?
Emerging Big Issue: Power

Energy dissipated = \( C \cdot V_{DD}^2 \) per gate

Power consumed = \( f \cdot n \cdot C \cdot V_{DD}^2 \) per chip

where

\( f \) = frequency of charge/discharge

\( n \) = number of gates /chip
Unfortunately...

Modern chip (UltraSparc III, Power4, Itanium 2) dissipates from 80W to 150W with a Vdd \( \approx 1.2V \) (Power supply current chip is \( \approx 100 \) Amps)

Ampacity is similar to a big double oven!

Cooling challenge is like making the filament of a 100W incandescent lamp cool to the touch!

Worse yet...

Little room left to reduce Vdd
nC and f continue to grow
Emerging Big Issue: Wires

Today (i.e., 100nm):

\[ \tau_{RC} \approx 50 \text{ps/mm} \]

Implies 2ns to traverse a 20mm x 20mm chip

This is a long time in a 2GHz processor
Summary

• MOSFET features
  – PN junctions provide electrical isolation
  – Switch-like behavior controlled by $V_{GS}$
  – Shrinking geometries improves performance

• CMOS features
  – CMOS logic is “naturally” inverting: “1” inputs lead to “0” outputs
  – Good noise margins because
    • $V_{OL} = 0, V_{OH} = V_{DD}$
    • Complementary logic has high gain
  – No static power dissipation

• Next time: timing, converting functionality to logic