Lecture 2
CMOS Devices

Xuan ‘Silvia’ Zhang
Washington University in St. Louis

http://classes.engineering.wustl.edu/ese566/
Outline

Transistor Model

Wire Model

Delay Model

CMOS Fabrication
Fundamental Building Block: MOSFET
Accumulation:
Voltage source puts negative charge on gate, attracts positively-charged majority carriers in p-type silicon body.
Metal-Oxide-Semiconductor Structure

Depletion:
Voltage source puts positive charge on gate, pushes positively-charged carriers away from surface, uncovers some negatively-charged dopant atoms in substrate.
Inversion:
Voltage source puts more positive charge on gate, instead of pushing holes even further away, draws free electrons to surface.
NMOS Transistor

**Cutoff:** $V_{gs} = 0V$, $V_{ds}$ can be $0V$ or $V_{dd}$
No Channel, $I_{ds} = 0$

**Linear:** $V_{gs} = V_{dd}$, $V_{ds} = 0V$
Channel Formed, $I_{ds}$ increases with $V_{ds}$

**Linear:** $V_{gs} = V_{dd}$, $V_{ds} = V_{dd}$
Channel Formed, $I_{ds}$ increases with $V_{ds}$

**Saturation:** Channel Pinched Off,
$I_{ds}$ independent of $V_{ds}$
Simple NMOS Circuit

$V_{gs} > V_t$

$V_{gd} = V_{gs}$

$V_{ds} = 0$

$V_{gs} - V_t$

**Linear:** Channel Formed

$I_{ds}$ increases with $V_{ds}$

$V_{gs} = V_{dd}$

$V_t$

$V_{ds} = V_{dd}$

$V_{gs}$

$V_{ds}$

$V$ vs. time
Key Qualitative Characteristics of MOSFET

- $V_t$ sets when transistor turns on, impacts leakage current.
- $I_d \propto \mu \times \left( \frac{W}{L} \right)$
- $\mu_n > \mu_p \implies R_{N,\text{eff}} < R_{P,\text{eff}}$
- $C_g \propto (W \times L)$
- $C_d \propto W$
- $\uparrow W = \downarrow R_{\text{eff}} = \uparrow I_d = \uparrow C_d, C_g$
- $\uparrow L = \uparrow R_{\text{eff}} = \downarrow I_d = \uparrow C_g$
Quantitative CMOS Model

- **Threshold Voltage**
  \[
  V_T = V_{T0} + \gamma \left( \sqrt{|-2\phi_F|} + V_{SB} - \sqrt{|-2\phi_F|} \right)
  \]
  \[
  \phi_0 = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2}
  \]

- **I-V Curve**
  - linear/triode
    \[
    I_D = \frac{\mu_n C_{ox}}{2} \frac{W}{L} \left( V_{gs} - V_{th} \right) V_{ds} - \frac{V_{ds}^2}{2}
    \]
  - saturation
    \[
    I_D = \frac{\mu_n C_{ox}}{2} \frac{W}{L} \left( V_{gs} - V_{th} \right)^2 \left( 1 + \lambda V_{ds} \right)
    \]

- **Parasitic Capacitance**
  - gate cap
    \[
    C_{j0} = \sqrt{\frac{\varepsilon_{si} q}{2} \left( \frac{N_A N_D}{N_A + N_D} \right)} \frac{1}{\sqrt{\phi_0}}
    \]
  - junction cap
Outline

Transistor Model

Wire Model

Delay Model

CMOS Fabrication
Wire Resistance

- Thickness fixed in given manufacturing process
- Resistance quoted as ohm/square
- TSMC 180nm 6 Aluminum metal layers
  - M1-M5: 0.08 ohm/square (0.5um x 1mm wire = 160 ohm)
  - M6: 0.03 ohm/square (0.5um x 1mm wire = 60 ohm)

\[
\text{resistance} = \frac{(\text{length} \times \text{resistivity})}{(\text{height} \times \text{width})}
\]

- **bulk aluminum**: \(2.8 \times 10^{-8} \ \Omega\cdot\text{m}\)
- **bulk copper**: \(1.7 \times 10^{-8} \ \Omega\cdot\text{m}\)
- **bulk silver**: \(1.6 \times 10^{-8} \ \Omega\cdot\text{m}\)
Wire Capacitance

- Capacitance depends on geometry of surrounding wires and relative permittivity, $\varepsilon_r$, of dielectric
  - Silicon dioxide, $\text{SiO}_2$, $\varepsilon_r = 3.9$
  - Silicon flouride, $\text{SiOF}$, $\varepsilon_r = 3.1$
  - SiLK polymer, $\varepsilon_r = 2.6$
Outline

Transistor Model

Wire Model

Delay Model

CMOS Fabrication
Qualitative Characteristics of Wire Delay

Wire resistance $\propto$ length
Wire capacitance $\propto$ length
Wire delay $\propto$ length$^2$
Quantitative Delay Model

- RC Ladder
  - Elmore delay model

\[ t_{pd} = \sum_{\text{nodes } i} R_{i \text{-to-source}} C_i \]

\[ = R_1 C_1 + (R_1 + R_2)C_2 + \cdots + (R_1 + R_2 + \cdots + R_N)C_N \]
Outline

Transistor Model

Wire Model

Delay Model

CMOS Fabrication
Mask Set for NMOS (circa 1986)

Vd = 1V  Vg = 0V  Vs = 0V

I = nA

dielectric

p-

Masks
#1: n+ diffusion
#2: poly (gate)
#3: diff contact
#4: metal

Layers to do p-Fet not shown. Modern processes have 6 to 10 metal layers (or more) (in 1986: 2).

Top-down view:
Design Rules for Masks (circa 1986)

Poly overhang. So that if masks are misaligned, channel doesn’t short out.

Minimum gate length. So that the source and drain depletion regions do not meet!

Metal rules: Contact separation from channel, one fixed contact size, overlap rules with metal, etc...

1. n+ diffusion
2. poly (gate)
3. diff contact
4. metal
Start with an Un-Doped Wafer

UV hardens exposed resist. A wafer wash leaves only hard resist.

Steps

#1: dope wafer p-
#2: grow gate oxide
#3: deposit undoped polysilicon
#4: spin on photoresist
#5: place positive poly mask and expose with UV.
Wet Etch to Remove Unmasked Regions

HF acid etches through poly and oxide, but not hardened resist.

After etch and resist removal
Use Diffusion Mask to Implant N-Type

accelerated donor atoms

Notice how donor atoms are blocked by gate and do not enter channel.

Thus, the channel is "self-aligned", precise mask alignment is not needed!
Metallization Completes Device

Grow a thick oxide on top of the wafer.

Mask and etch to make contact holes.

Put a layer of metal on chip. Be sure to fill in the holes!
Final NMOS Transistor

"The planar process"
Jean Hoerni, Fairchild Semiconductor 1958

Top-down view:
PMOS—Dual of NMOS

\[ V_{\text{well}} = V_s = 1V \quad V_g = 0V \quad V_d = 0V \]

New “n-well” mask

“Mobility” of holes is slower than electrons.

p-Fets drive less current than n-Fets, all else being equal
Single- and Triple-Well Process
Local Interconnect

IBM 6-Transistor SRAM Cell
Intel Metal Stacks: 90nm and 45nm
Intel Metal Stacks: 45nm with M9 and I/O Bump
## Intel Metal Layer Dimensions in 45nm

<table>
<thead>
<tr>
<th>Layer</th>
<th>t (nm)</th>
<th>w (nm)</th>
<th>s (nm)</th>
<th>pitch (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M9</td>
<td>7μm</td>
<td>17.5μm</td>
<td>13μm</td>
<td>30.5μm</td>
</tr>
<tr>
<td>M8</td>
<td>720</td>
<td>400</td>
<td>410</td>
<td>810</td>
</tr>
<tr>
<td>M7</td>
<td>504</td>
<td>280</td>
<td>280</td>
<td>560</td>
</tr>
<tr>
<td>M6</td>
<td>324</td>
<td>180</td>
<td>180</td>
<td>360</td>
</tr>
<tr>
<td>M5</td>
<td>252</td>
<td>140</td>
<td>140</td>
<td>280</td>
</tr>
<tr>
<td>M4</td>
<td>216</td>
<td>120</td>
<td>120</td>
<td>240</td>
</tr>
<tr>
<td>M3</td>
<td>144</td>
<td>80</td>
<td>80</td>
<td>160</td>
</tr>
<tr>
<td>M2</td>
<td>144</td>
<td>80</td>
<td>80</td>
<td>160</td>
</tr>
<tr>
<td>M1</td>
<td>144</td>
<td>80</td>
<td>80</td>
<td>160</td>
</tr>
</tbody>
</table>
IBM Metal Stack

IBM 11-layer Copper Metal Stack

IBM 6-layer Copper Metal Stack
Bulk vs Silicon-on-Insulator (SOI) Processing

- Eliminate parasitic cap between S/D and body
  - lower energy, higher performance
- Lower sub-Vth leakage, but Vth varies over time
- 10-15% increase in total manufacturing cost
  - due to substrate cost
Lithography

- Pattern resolution exceeds wavelength of light
  - 193nm from argon fluoride laser
- Sophisticated patterning tricks
  - immersion lithography
  - optical proximity correction (OPC)
  - double patterning
Processing Enhancements

- **High-K Dielectrics and Metal Gates** – Replacing silicon dioxide gate dielectric with a high-K material allows increased vertical electric field without increasing gate leakage.

- **Strained Silicon** – Layer of silicon in which silicon atoms are stretched beyond their normal interatomic distance leading to better mobility.

- **Gate Engineering** – Multiple transistor designs with different threshold voltages to allow optimization of delay or power.
FinFET Transistors

- Small footprint, good gate control
  - use the vertical dimension
  - Intel starts using FinFET in 20nm
Questions?

Comments?

Discussion?
Acknowledgement

N. Weste and D. Harris, “CMOS VLSI Design”, 2011
Cornell University, ECE 5745
UC Berkeley, CS 230
MIT, 6.371