Lecture 9: CMOS
From the Bottom Up

- IC processing
- MOS transistors
- CMOS Circuits
Overview of Physical Implementations

The stuff out of which we make systems.

- Integrated Circuits (ICs)
  - Combinational logic circuits, memory elements, analog interfaces.

- Printed Circuits (PC) boards
  - Substrate for ICs and interconnection, distribution of CLK, Vdd, and GND signals, heat dissipation.

- Power Supplies
  - Converts line AC voltage or battery DC voltage to regulated DC low voltage levels.

- Chassis (rack, card case, ...)
  - Holds boards, power supply, fans, provides physical interface to user or other systems.

- Connectors and Cables.
- Peripheral and I/O components.
Printed Circuit Boards

- fiberglass or ceramic
- 1-25 conductive layers
- ~1-20in on a side
- IC packages are soldered down.

Multichip Modules (MCMs)

- Multiple chips directly connected to a substrate. (silicon, ceramic, plastic, fiberglass) without chip packages.
Integrated Circuits

- Primarily Crystalline Silicon
- 1mm - 25mm on a side
- 100 - 20B transistors
- (25 - 250M “logic gates”)
- 3 - 10 conductive layers
- 2019 state-of-the-art feature size
  7nm = 0.007 x 10^-6 m
- “CMOS” most common - complementary metal oxide semiconductor

Chip in Package

- Package provides:
  - spreading of chip-level signal paths to board-level
  - heat dissipation.
- Ceramic or plastic with gold wires.
Chip Fabrication
Silicon “ingots” are grown from a “perfect” crystal seed in a melt, and then purified to “nine nines”.

[Diagram showing the process of growing silicon ingots]

[Image of a clean room with workers in protective suits]

[Image of a crucible with molten silicon]
Ingots sliced into 450μm thick wafers, using a diamond saw.
CMOS Transistors

- MOSFET (Metal Oxide Semiconductor Field Effect Transistor).

The gate acts like a capacitor. A high voltage on the gate attracts charge into the channel. If a voltage exists between the source and drain a current will flow. In its simplest approximation, the device acts like a switch.
An n-channel MOS transistor (planar)

$V_d = 1\, \text{V}$

$V_g = 0\, \text{V}$

$V_s = 0\, \text{V}$

Polysilicon gate, dielectric, and substrate form a capacitor.

$n$Fet is off (I is "leakage")

$I \approx nA$

$V_d = 1\, \text{V}$

$V_g = 1\, \text{V}$

$V_s = 0\, \text{V}$

$V_g = 1\, \text{V}$, small region near the surface turns from p-type to n-type.

$n$Fet is on.

$I \approx \mu A$
Mask set for an n-Fet (circa 1986)

- **$V_d = 1V$**
- **$V_g = 0V$**
- **$V_s = 0V$**

- $I \approx nA$

**Dielectric**

**Top-down view:**

- **#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).
“Design rules” for masks, 1986 ...

Poly overhang. So that if masks are misaligned, we still get channel.

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

length
How a fab uses a mask set to make an IC

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

Top-down view:

Masks

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

Vg

Vd

Vs

I = μA

Ids

dielectric

p−

H+
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 polysilicon
#4: spin on photoresist
#5: place positive poly mask and expose with UV.
Wet etch to remove unmasked...

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 product...

Top-down view:

"The planar process"

Jean Hoerni, Fairchild Semiconductor 1958
Lithography

Current state-of-the-art photolithography tools use **deep ultraviolet (DUV)** light with wavelengths of 248 and 193 nm, which allow minimum feature sizes below 50 nm. Moving to **extreme ultraviolet (EUV)** with wavelength of 13.5nm.

Optical proximity correction (OPC) is an enhancement technique commonly used to compensate for image errors due to **diffraction** or process effects.
Transistor channel is a raised fin.
Gate controls channel from sides and top.
CMOS Transistors – State-of-the-Art
State of the art

7nm
As of September 2018, mass production of 7 nm devices has begun. The first mainstream 7 nm mobile processor intended for mass market use, the Apple A12 Bionic, was released at their September 2018 event. Although Huawei announced its own 7 nm processor before the Apple A12 Bionic, the Kirin 980 on August 31, 2018, the Apple A12 Bionic was released for public, mass market use to consumers before the Kirin 980. Both chips are manufactured by TSMC. AMD is currently working on their "Rome" workstation processors, which are based on the 7 nanometer node and feature up to 64 cores.

5nm
The 5 nm node was once assumed by some experts to be the end of Moore's law. Transistors smaller than 7 nm will experience quantum tunnelling through the gate oxide layer. Due to the costs involved in development, 5 nm is predicted to take longer to reach market than the two years estimated by Moore's law. Beyond 7 nm, it was initially claimed that major technological advances would have to be made to produce chips at this small scale. In particular, it is believed that 5 nm may usher in the successor to the FinFET, such as a gate-all-around architecture.


3.5nm
3.5 nm is a name for the first node beyond 5 nm. In 2018, IMEC and Cadence had taped out 3 nm test chips. Also, Samsung announced that they plan to use Gate-All-Around technology to produce 3 nm FETs in 2021.

* From Wikipedia
CMOS Transistors
1. Transistor “strength” proportional to $W/L$. In digital circuits, $L$ is almost always minimal allowed by process.

2. MOS transistors are symmetrical devices (Source and drain are interchangeable). But usually designed to be used in one direction.

Source is the node w/ the lowest voltage for (N-FET), in general the source node is connected to power rail.
Circuit Layout Examples

- **2-input NAND**

NAND gate layout from Lecture 3: CMOS Technology and Logic Gates. (Image by Professors Arvind and Asanovic.)

Finfet layout

NAND gate layout from Ji Li.
MOS Transistor as a Resistive Switch

Let’s look beneath the abstraction:
origins of $V_T$ and $R_{on}$
MOSFET Threshold Voltage

Transistor "turns on" when $V_{gs}$ is $>$ $V_t$.

$V_{ds} = V_{dd}$

$I_{ds} = 1.2$ mA $= I_{on}$

$I_{off} = 0$ ???

$0.25 = V_t$

$0.7 = V_{dd}$
Transistor “resistance”

- Actually, nonlinear I/V characteristic:

- But, linearizing makes all delay and power calculations simple (usually just 1st order ODEs):
ON/OFF Switch Model of MOS Transistor

$|V_{GS}| < |V_T|$  

$|V_{GS}| \geq |V_T|$
Plot on a “Log” Scale to See “Off” Current

Process engineers can:
- increase $I_{on}$ by lowering $V_t$ - but that raises $I_{off}$
- decrease $I_{off}$ by raising $V_t$ - but that lowers $I_{on}$.

$0.25 \approx V_t$

$1.2 \text{ mA} = I_{on}$

$I_{off} \approx 10 \text{ nA}$
Latest Modern Process

Transistor channel is a raised fin. Gate controls channel from sides and top.
A More Realistic Switch

Transistors in the sub 100 nm age

\[ |V_{GS}| < |V_T| \]

\[ R_{on} \]

\[ |V_{GS}| > |V_T| \]

\[ R_{off} \]
A Logic Perspective

NMOS Transistor

\[ V_{GS} > 0 \]

\[ Y = Z \text{ if } X = 1 \]
A Complementary Switch

Y = Z if X = 0

PMOS Transistor

\[ V_{GS} < 0 \]

Source is the node with the highest voltage!
The CMOS Inverter: A First Glance

Represents the sum of all the capacitance at the output of the inverter and everything to which it connects: (drains, interconnections gate capacitance of next gate(s))
The Switch Inverter
First-Order DC Analysis*

\[ V_{OL} = 0 \]
\[ V_{OH} = V_{DD} \]

*First-order means we will ignore Capacitance.
Switch logic
Static Logic Gate

- At every point in time (except during the switching transients) each gate output is connected to either $V_{DD}$ or $V_{GND}$ via a low resistive path.

- The output of the gate assumes at all times the value of the Boolean function implemented by the circuit (ignoring, once again, the transient effects during switching periods).

Example: CMOS Inverter
Building logic from switches (NMOS)

**Series**

\[
Y = X \text{ if } A \text{ OR } B
\]

**Parallel**

\[
Y = X \text{ if } A \text{ AND } B
\]

(output undefined if condition not true)
Logic using inverting switches (PMOS)

Series

\[ Y = X \text{ if } \overline{A} \text{ AND } \overline{B} \]
\[ = \overline{A + B} \]

Parallel

\[ Y = X \text{ if } \overline{A} \text{ OR } \overline{B} \]
\[ = \overline{AB} \]

(output undefined if condition not true)
Example Gate: NAND

- PDN: $G = AB \Rightarrow$ Conduction to GND
- PUN: $F = \overline{A} + \overline{B} = \overline{AB} \Rightarrow$ Conduction to $V_{DD}$

Truth Table of a 2 input NAND gate

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Out</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>
PUN and PDN are **dual logic networks**:
- series connections in the PUN are parallel connections in the PDN
- parallel connections in the PUN are series connections in the PDN

PUN and PDN functions are **complements**:
- guarantees they are mutually exclusive, under all input values, one or the other is conductive, but never both!
Example Gate: NOR

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Out</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>

Truth Table of a 2 input NOR gate

\[ \text{OUT} = \overline{A + B} \]
Complex CMOS Gate

\[ \text{OUT} = D + A \cdot (B + C) \]

\[ \text{OUT} = D \cdot A + B \cdot C \]
Graph Models for Duals

Pull-down circuit

I_3

I_1

d

c

I_2

b

a

Pull-up circuit

V_{DD}

a

b

d

c

Gnd

Pull-down

Pull-up

V_{DD}

Gnd

Non-inverting logic

PUN and PDN are dual logic networks
PUN and PDN functions are complementary

Why is this a bad idea?

Non-Inverting switches

\[ F(\text{In}_1, \text{In}_2, \ldots \text{In}_N) \]

Inverting switches
Switch Limitations

Good 1
\[ V_{DD} \rightarrow V_{DD} \]

Good 0
\[ V_{DD} \rightarrow 0 \]

Bad 1
\[ 0 \rightarrow V_{DD} - V_{Tn} \]

Bad 0
\[ V_{DD} \rightarrow |V_{Tp}| \]

Tough luck …
“Static” CMOS gates

Inverting switches (PMOS transistors)

Non-Inverting switches (NMOS transistors)

Static CMOS gates are always inverting

\[
\text{AND} = \text{NAND} + \text{INV}
\]
Complimentary CMOS Properties

- Full rail-to-rail swing
- Besides leakage (due to $I_{off}$), no static power dissipation
- Direct path current during switching
Transmission Gate

- Transmission gates are the way to build ideal “switches” in CMOS.
- In general, both transistor types are needed:
  - nFET to pass zeros.
  - pFET to pass ones.
- The transmission gate is bi-directional (unlike logic gates).

if \( \text{en} == 1 \) then \( A \) connects to \( B \)

- Does not directly connect to Vdd and GND, but can be combined with logic gates or buffers to simplify many logic structures.
Transmission-gate Multiplexor

2-to-1 multiplexor:
\[ c = sa + s'b \]

Switches simplify the implementation:

Compare the cost to logic gate implementation.

Care must be taken to not string together many pass-transistor stages. Occasionally, need to “rebuffer” with static gate.
4-to-1 Transmission-gate Mux

- The series connection of pass-transistors in each branch effectively forms the AND of s1 and s0 (or their complement).

- Compare cost to logic gate implementation

Any alternate solutions?
Alternative 4-to-1 Multiplexor

- This version has less delay from in to out.

- In both versions, care must be taken to avoid turning on multiple paths simultaneously (shorting together the inputs).
**Tri-state Buffers**

**Tri-state Buffer:**

```
<table>
<thead>
<tr>
<th>OE</th>
<th>IN</th>
<th>OUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Z</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Z</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
```

“high impedance” (output disconnected)

**Variations:**

- **Inverting buffer**

- **Inverted enable**

**CMOS Implementation**

Transmission gates provide the isolation: usually designed this way.
Tri-state buffers are used when multiple circuits all connect to a common node or wire. Only one circuit at a time is allowed to drive the bus. All others “disconnect” their outputs, but can “listen”.

Tri-state buffers enable “bidirectional” connections.
Tri-state Based Multiplexor

Multiplexor:

If \( s=1 \) then \( c=a \) else \( c=b \)

Transistor Circuit for inverting-multiplexor:
Latches and Flip-flops

Positive Level-sensitive latch:

Latch Implementation:
Positive edge-triggered flip-flop

A flip-flop “samples” right before the edge, and then “holds” value.

Sampling circuit

Holds value

Delay in Flip-flops

• Setup time results from delay through first latch.

• Clock to Q delay results from delay through second latch.

D Q

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'

clk

clk'
Sensing: When clock is low

A flip-flop “samples” right before the edge, and then “holds” value.

**Sampling circuit**

**Holds value**

\[ \text{clk} = 0 \]
\[ \text{clk'} = 1 \]

Will capture new value on posedge.

Outputs last value captured.
Capture: When clock goes high

A flip-flop “samples” right before the edge, and then “holds” value.

Sampling circuit

Holds value

\[ \text{clk} = 1 \]
\[ \text{clk}' = 0 \]

Remembers value just captured.

Outputs value just captured.
**Tri-state-Inverter Latch**

- Commonly used in standard cell flip-flops.
- More transistors than pass-transistor version, but more robust.
- Lays out well with modern layout rules.

**Positive Level-sensitive latch:**