

CMOS Transistor Theory

Dr. Tassadaq Hussain

Computer Architect

Co-Founder PakAsic.com



Outline

Introduction

MOS Capacitor

nMOS I-V Characteristics

pMOS I-V Characteristics

Gate and Diffusion Capacitance

Pass Transistors

RC Delay Models

Introduction

So far, we have treated transistors as ideal switches

An ON transistor passes a finite amount of current

Depends on terminal voltages

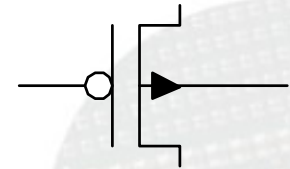
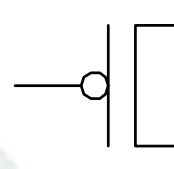
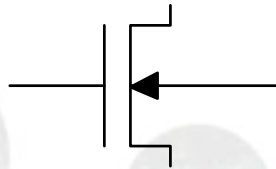
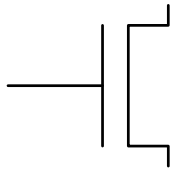
Derive current-voltage (I-V) relationships

Transistor gate, source, drain all have capacitance

$$I = C (\Delta V / \Delta t) \rightarrow \Delta t = (C / I) \Delta V$$

Capacitance and current determine speed

Also explore what a “degraded level” really means



MOS Capacitor

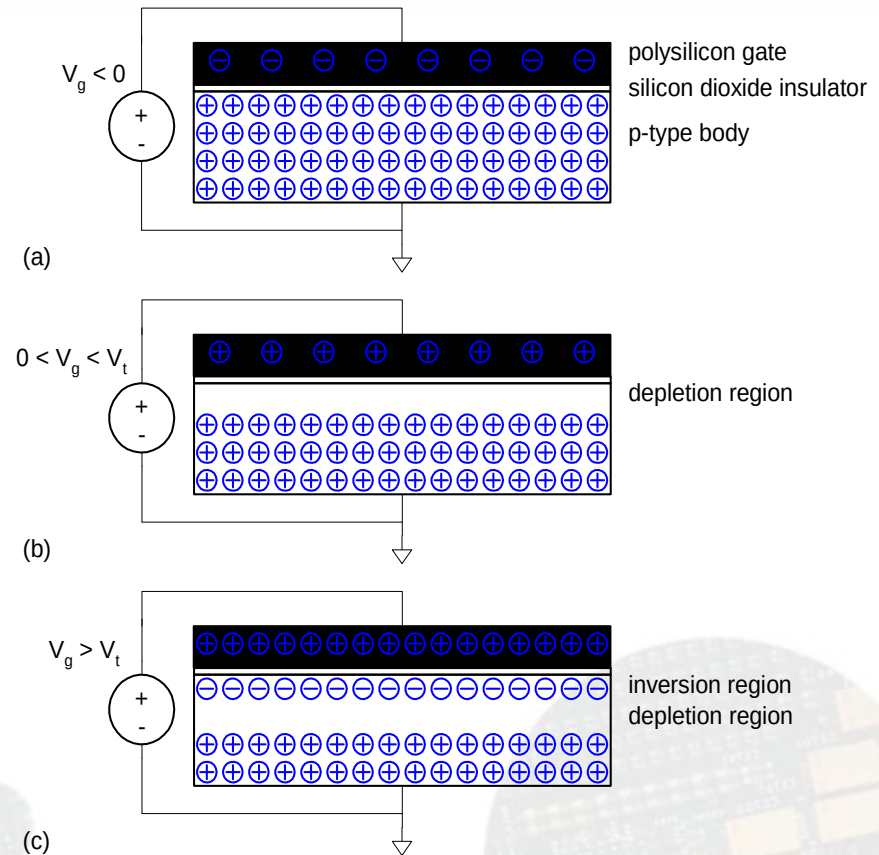
Gate and body form MOS capacitor

Operating modes

Accumulation

Depletion

Inversion



Example with an NMOS capacitor

Accumulation occurs typically for negative voltages where the negative charge on the gate attracts holes from the substrate to the oxide-semiconductor interface.

Depletion occurs for positive voltages; The positive charge on the gate pushes the mobile holes into the substrate, thereby depleting the semiconductor of the mobile carriers and leaving a negative charge in the space charge region which is due to the ionized acceptor ions.

The voltage separating the accumulation and depletion regime is referred to as the flatband voltage.

Inversion occurs at more positive voltages which are larger than the threshold voltage. In addition to the depletion layer charge a negatively charged inversion layer forms at the oxide-semiconductor interface.

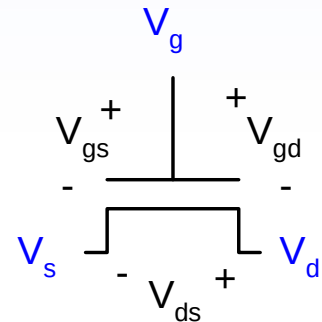
Terminal Voltages

Mode of operation depends on V_g , V_d , V_s

$$V_{gs} = V_g - V_s$$

$$V_{gd} = V_g - V_d$$

$$V_{ds} = V_d - V_s = V_{gs} - V_{gd}$$



Source and drain are symmetric diffusion terminals

However, $V_{ds} \geq 0$

NMOS body is grounded. First assume source may be grounded or may be at a voltage above ground.

Three regions of operation

Cutoff

Linear

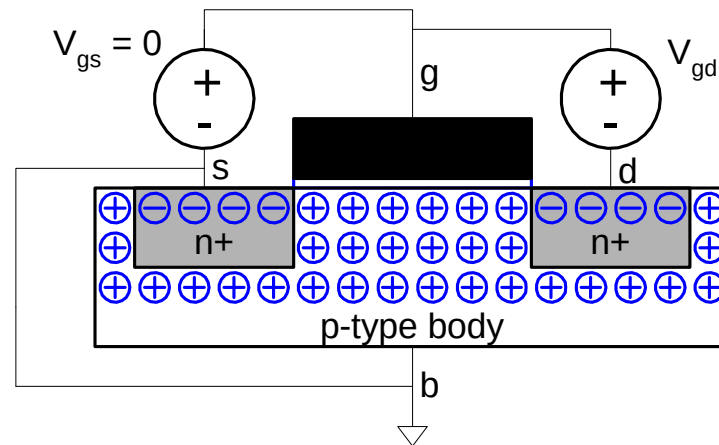
Saturation

nMOS Cutoff

Let us assume $v_s = v_d$

No channel, if $v_{gs} = 0$

$$I_{ds} = 0$$



NMOS Linear

Channel forms if $V_{gs} > V_t$

No Current if $V_{ds} = 0$

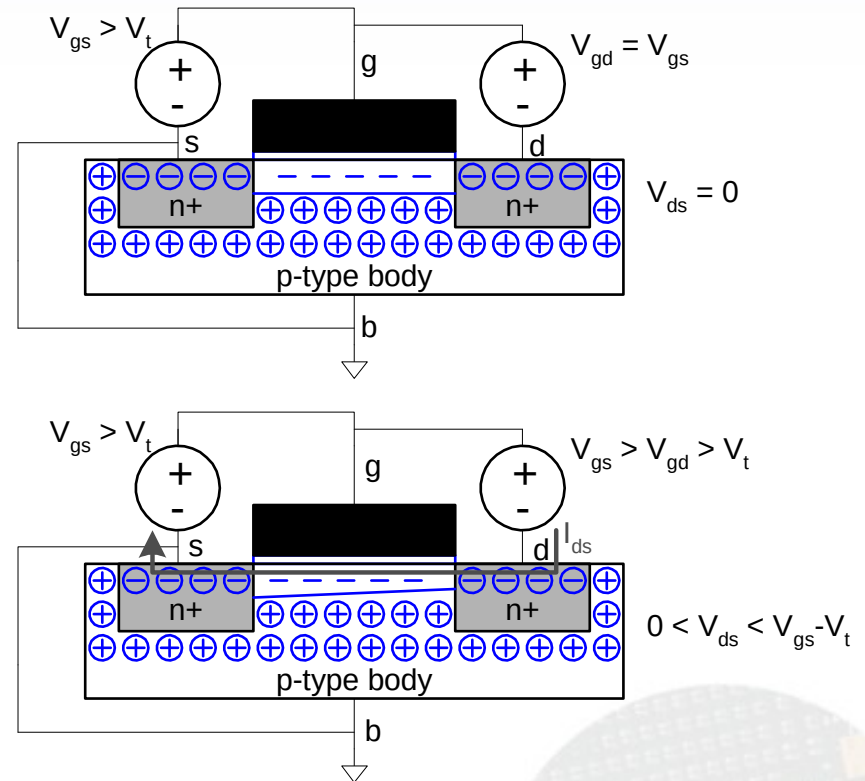
Linear Region:

If $V_{ds} > 0$, Current flows from d to s (e^- from s to d)

I_{ds} increases linearly

with V_{ds} if $V_{ds} > V_{gs} - V_t$

Similar to linear resistor

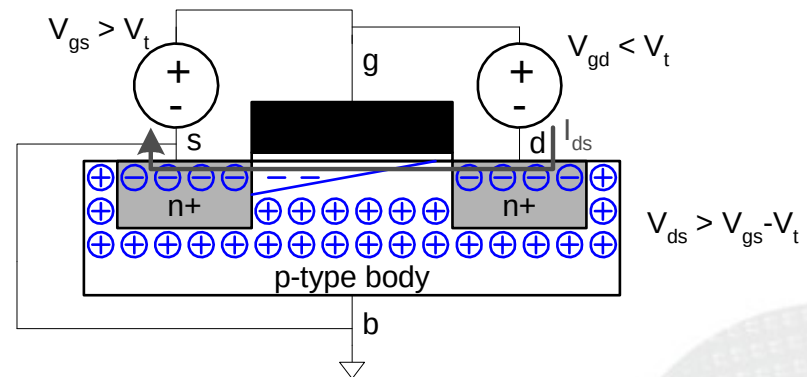


NMOS Saturation

Channel pinches off if $V_{ds} > V_{gs} - V_t$.

I_{ds} “independent” of V_{ds} , i.e., current saturates

Similar to current source



I-V Characteristics

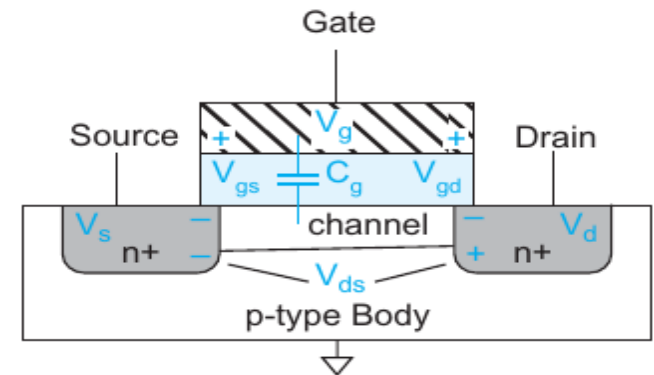
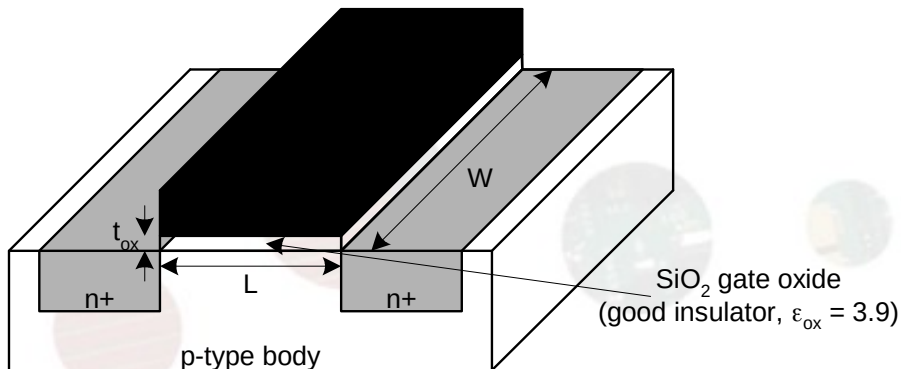
In Linear region, I_{ds} depends on
How much charge is in the channel
How fast is the charge moving

Channel Charge

MOS structure looks like parallel plate capacitor while operating in inversion

Gate – oxide – channel

$$Q = CV$$



Average gate to channel potential:

$$V_{gc} = (V_{gs} + V_{gd})/2 = V_{gs} - V_{ds}/2$$

Channel Charge

MOS structure looks like parallel plate capacitor while operating in inversion

Gate – oxide – channel

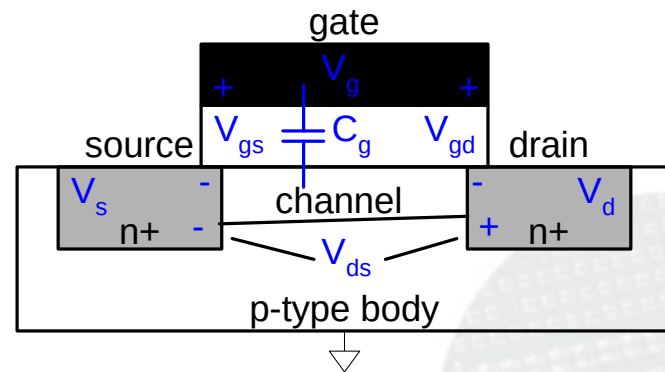
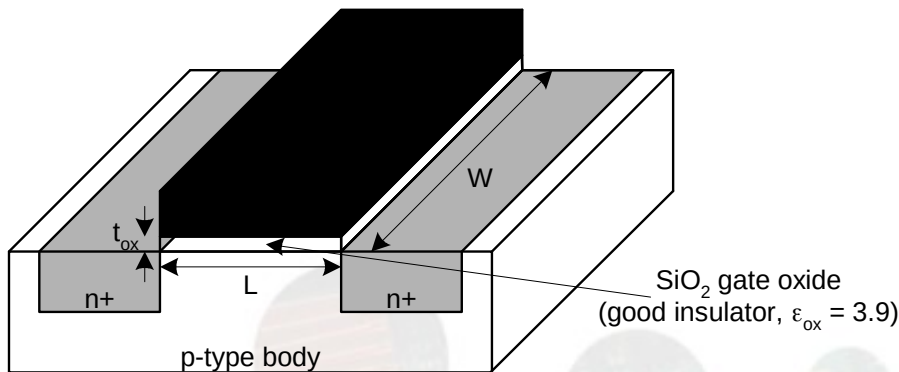
$$Q_{\text{channel}} = CV$$

$C = C_g = \epsilon_{\text{ox}} WL/t_{\text{ox}} = C_{\text{ox}} WL$ (where ϵ_{ox} is the permittivity of free space, 8.85×10^{-14} F/cm, and the permittivity of SiO₂ is $k_{\text{ox}} = 3.9$ times as great. Often, the $\epsilon_{\text{ox}}/t_{\text{ox}}$ term is called C_{ox} , the capacitance per unit area of the gate oxide)

$$B = \mu C_{\text{ox}} W/L ; V_{\text{GT}} = V_{\text{gs}} - V_{\text{T}}$$

$$V_{\text{gc}} = (V_{\text{gs}} + V_{\text{gd}})/2 = V_{\text{gs}} - V_{\text{ds}}/2$$

$$V = V_{\text{gc}} - V_{\text{t}} = (V_{\text{gs}} - V_{\text{ds}}/2) - V_{\text{t}}$$



Carrier velocity

Charge is carried by e-

Carrier velocity v proportional to lateral E-field between source and drain

$$v = \mu E \quad \mu \text{ called mobility}$$

$$E = V_{ds}/L$$

Time for carrier to cross channel:

$$t = L / v$$

$$Q_{\text{channel}} = C_g (V_{gs} - V_t) \quad (2.1)$$

$$C_g = k_{\text{ox}} \epsilon_0 \frac{WL}{t_{\text{ox}}} = \epsilon_{\text{ox}} \frac{WL}{t_{\text{ox}}} = C_{\text{ox}} WL \quad (2.2)$$

$$v = \mu E \quad (2.3)$$

$$E = \frac{V_{ds}}{L} \quad (2.4)$$

$$\begin{aligned} I_{ds} &= \frac{Q_{\text{channel}}}{L/v} \\ &= \mu C_{\text{ox}} \frac{W}{L} (V_{gs} - V_t - V_{ds}/2) V_{ds} \\ &= \beta (V_{GT} - V_{ds}/2) V_{ds} \end{aligned} \quad (2.5)$$

$$\beta = \mu C_{\text{ox}} \frac{W}{L}; \quad V_{GT} = V_{gs} - V_t \quad (2.6)$$

NMOS Linear I-V

Now we know

How much charge Q_{channel} is in the channel

How much time t each carrier takes to cross

$$I_{ds} = \frac{Q_{\text{channel}}}{t}$$

$$= \mu C_{\text{ox}} \frac{W}{L} \left(V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds}$$

$$= \beta \left(V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds}$$

$$\beta = \mu C_{\text{ox}} \frac{W}{L}$$

NMOS Saturation I-V

If $V_{gd} < V_t$, channel pinches off near drain

When $V_{ds} > V_{dsat} = V_{gs} - V_t$

Now drain voltage no longer increases current

$$I_{ds} = \beta \left(V_{gs} - V_t - \frac{V_{dsat}}{2} \right) V_{dsat}$$
$$= \frac{\beta}{2} (V_{gs} - V_t)^2$$

NMOS I-V Summary

Shockley 1st order transistor models (valid for Large channel devices only)

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t & \text{cutoff} \\ \beta \left(V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds} & V_{ds} < V_{dsat} & \text{linear} \\ \frac{\beta}{2} (V_{gs} - V_t)^2 & V_{ds} > V_{dsat} & \text{saturation} \end{cases}$$

Example

For a 0.6 μm process ([MOSIS site](#))

From AMI Semiconductor

$$t_{\text{ox}} = 100 \text{ \AA}$$

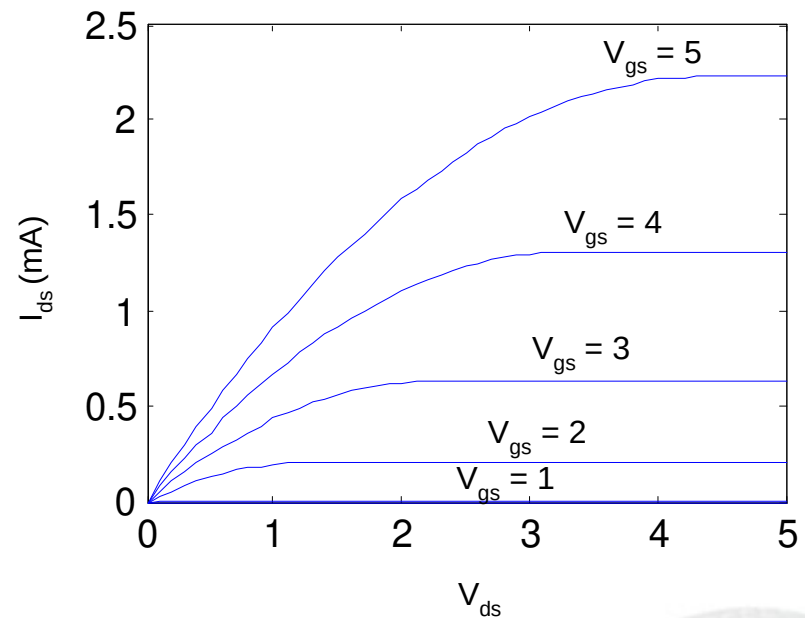
$$\mu = 350 \text{ cm}^2/\text{V}^*\text{s}$$

$$V_t = 0.7 \text{ V}$$

Plot I_{ds} vs. V_{ds}

$$V_{\text{gs}} = 0, 1, 2, 3, 4, 5$$

Use $W/L = 4/2 \lambda$



$$\beta = \mu C_{\text{ox}} \frac{W}{L} = (350) \left(\frac{3.9 \cdot 8.85 \cdot 10^{-14}}{100 \cdot 10^{-8}} \right) \left(\frac{W}{L} \right) = 120 \frac{W}{L} \mu\text{A}/\text{V}^2$$

PMOS I-V

All dopings and voltages are inverted for PMOS

Mobility μ_p is determined by holes

Typically 2-3x lower than that of electrons μ_n

120 cm²/V*s in AMI 0.6 μm process

Thus PMOS must be wider to provide same current

In this class, assume $\mu_n / \mu_p = 2$

Factors: Power, Speed and Area

Gate Capacitance (C_{gate}):

Gate capacitance refers to the capacitance between the gate terminal and the substrate of a MOSFET. It affects the charging and discharging time of the gate, influencing the switching speed of the transistor.

Higher gate capacitance results in longer charging and discharging times, slowing down the transistor's switching speed.

Gate Length (L_{gate}):

Gate length is the physical length of the gate electrode in a MOSFET. Shorter gate lengths typically result in faster switching speeds due to reduced channel resistance and gate capacitance.

Transistor Width (W):

Transistor width (W) refers to the width of the transistor channel along the direction of current flow. Increasing transistor width can reduce resistance and improve drive strength, leading to faster switching speeds. However, wider transistors consume more area and may increase parasitic capacitance, impacting power consumption and speed.

Threshold Voltage (V_{th}):

Threshold voltage is the minimum voltage required to turn on a MOSFET. Lower threshold voltages facilitate faster switching speeds by reducing the delay in turning the transistor on and off. Lower threshold voltages may also increase leakage currents, especially in subthreshold regions, impacting power consumption.

Gate Oxide Thickness:

Gate oxide thickness refers to the thickness of the insulating layer (oxide) between the gate and the channel. Thinner gate oxide reduces gate capacitance and improves transistor speed by allowing stronger electric fields. Thinner gate oxides are more susceptible to gate oxide leakage, which can increase static power consumption.

Channel Doping Profile:

The doping profile of the semiconductor channel affects the threshold voltage and conductivity of the transistor. Optimizing the doping profile can improve drive strength and reduce resistance, enhancing circuit speed and efficiency.

Interconnect Length and Metal Layers:

The length and routing of interconnects between transistors impact signal propagation delay and RC time constants. Longer interconnects introduce additional resistance and capacitance, slowing down signal propagation and increasing power consumption. Increasing the number of metal layers allows for better routing density and reduces the area footprint of interconnects.

Capacitance

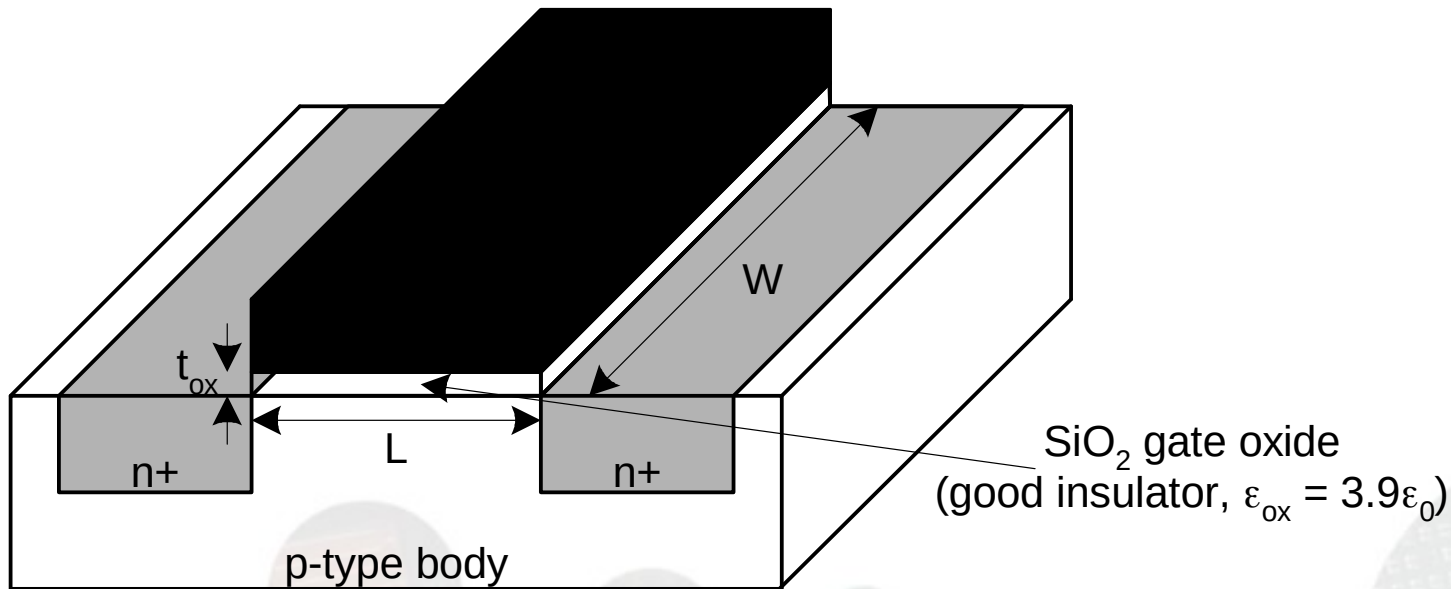
- Any two conductors separated by an insulator have capacitance
- Gate to channel capacitor is very important
- Creates channel charge necessary for operation
- Source and drain have capacitance to body
- Across reverse-biased diodes
- Called diffusion capacitance because it is associated with source/drain diffusion

Gate Capacitance

Approximate channel as connected to source

$$C_{gs} = \epsilon_{ox} WL/t_{ox} = C_{ox} WL = C_{\text{permicron}} W$$

$C_{\text{permicron}}$ is typically about 2 fF/ μm



Diffusion Capacitance

$$C_{sb}, C_{db}$$

Undesirable, called *parasitic* capacitance

Capacitance depends on area and perimeter

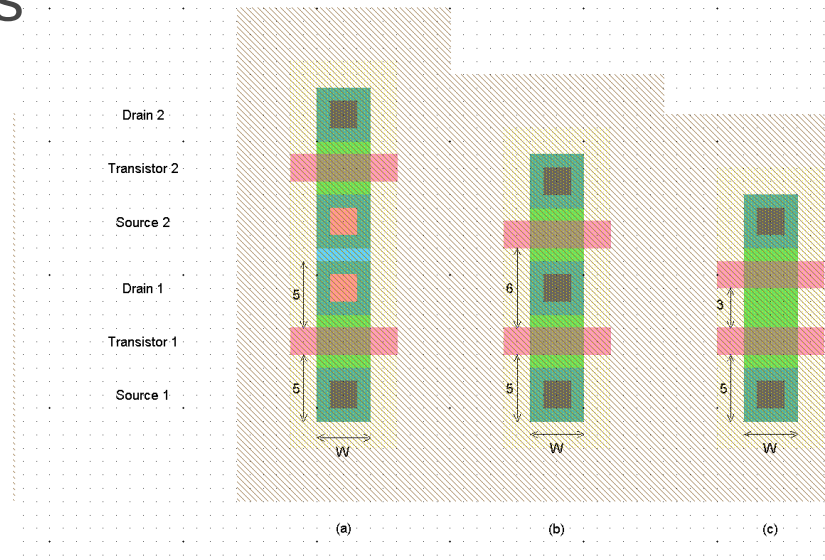
Use small diffusion nodes

Comparable to C_g

for contacted diff

$\frac{1}{2} C_g$ for uncontacted

Varies with process



Digital Circuit Design and Optimization

Following techniques play important role in digital circuit design and optimization, contributing to the overall correctness, reliability, and performance of the final design.

Impact of Interconnects, Transistor Switches, and Clock Distribution

Interconnects (wires) connect different components such as transistors, gates, and other circuit elements. The resistance of these interconnects affects signal propagation delays and power dissipation. As the feature sizes of VLSI technologies shrink, the resistance of interconnects becomes more significant due to increased wire lengths and reduced wire widths. Designers need to consider the effective resistance of interconnects to ensure signal integrity and minimize delays.

Transistor Switch: The effective resistance of a transistor affects its switching behavior and power consumption. For example, in NMOS transistors, the on-resistance (or channel resistance) determines how effectively the transistor can pull down a node to logic '0'. In PMOS (P-channel Metal-Oxide-Semiconductor) transistors, the effective resistance similarly affects pull-up behavior. Minimizing the effective resistance of transistor switches is essential for optimizing circuit performance and reducing power consumption.

Clock Distribution: Clock networks synchronize the clock signals used to operate different components. The effective resistance of clock distribution networks impacts clock skew, which refers to the variation in arrival times of clock signals at different parts of the circuit. High resistance in clock distribution networks can lead to increased skew, affecting circuit timing and performance.

Static Timing Analysis (STA): STA is closely related to formal verification as it aims to ensure that the timing constraints of a digital design are met. While STA is not strictly formal verification in itself, it employs mathematical methods to analyze the timing behavior of a design and detect violations of timing constraints.

Gate-level Power Estimation: Power estimation techniques help in analyzing the power consumption of a design, which is crucial for ensuring reliability and optimizing power usage. While not directly part of formal verification, power estimation contributes to overall design correctness and reliability.

RC Delay Analysis: RC delay analysis considers the effects of resistance and capacitance in the interconnects of digital circuits. While not a formal verification technique, it helps estimate signal propagation delay due to interconnect parasitics, which can impact the correctness of the design.

Path Delay Analysis: Path delay analysis identifies critical timing paths in a digital circuit. While not formal verification in itself, it helps in identifying potential timing violations that could affect the correctness of the design.

Logical Effort Analysis: Logical effort analysis aids in selecting optimal gate sizes and transistor configurations to minimize delay and power consumption. While not directly part of formal verification, it contributes to the optimization of design performance and correctness.

Monte Carlo Analysis: Monte Carlo analysis helps in evaluating the impact of manufacturing variations on circuit performance and reliability. While not formal verification, it contributes to ensuring the robustness and reliability of the design.

High-Level Synthesis (HLS): HLS translates high-level descriptions of algorithms into RTL designs. While not strictly formal verification, HLS tools optimize circuit performance and resource utilization while considering timing constraints and architectural trade-offs, which indirectly contribute to design correctness.

Effective Resistance

Effective resistance is used to determine the performance and behavior of VLSI circuit components, particularly interconnects and transistor switches.

Effective Resistance

Shockley models have limited value

Not accurate enough for modern transistors

Too complicated for much hand analysis

Simplification: treat transistor as resistor

Replace $I_{ds}(V_{ds}, V_{gs})$ with effective resistance R

- $I_{ds} = V_{ds}/R$

R averaged across switching of digital gate

Too inaccurate to predict current at any given time

But good enough to predict RC delay

RC Delay Analysis:

RC delay analysis focuses on estimating the delay in signal propagation caused by resistance (R) and capacitance (C) in interconnects and wiring within a digital circuit.

It helps designers understand and optimize the timing behavior of the interconnects, which can be significant contributors to overall signal propagation delay in complex digital circuits.

The purpose of RC delay analysis is to identify and minimize signal propagation delay due to interconnect parasitics, thus improving the overall timing performance of the circuit.

RC Network

RC (resistor-capacitor) network involves understanding CMOS VLSI behavior regarding charging, discharging, time constants, frequency response, and transient response

RC Delay Model

Use equivalent circuits for MOS transistors

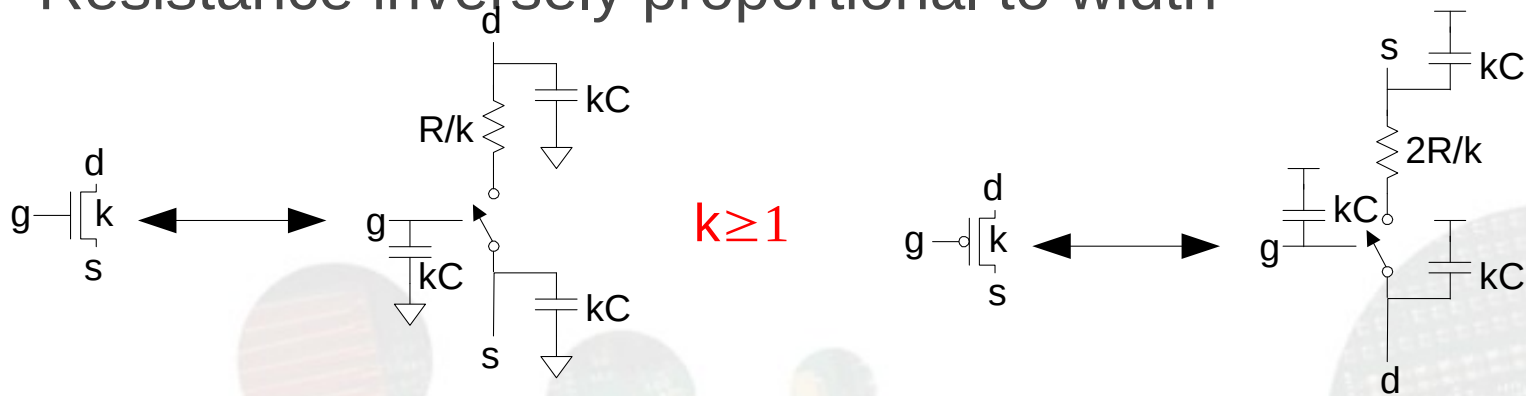
Ideal switch + capacitance and ON resistance

Unit nMOS has resistance R , capacitance C

Unit pMOS has resistance $2R$, capacitance C

Capacitance proportional to width

Resistance inversely proportional to width



RC Values

Capacitance

$C = C_g = C_s = C_d = 2 \text{ fF}/\mu\text{m}$ of gate width

Values similar across many processes

Resistance

$R \approx 6 \text{ K}\Omega \cdot \mu\text{m}$ in 0.6 μm process

Improves with shorter channel lengths

Unit transistors

May refer to minimum contacted device ($4/2 \lambda$)

Or maybe 1 μm wide device

Doesn't matter as long as you are consistent

Path Delay Analysis

Path delay analysis involves identifying the critical timing paths within a digital circuit where the signal propagation delay is the longest.

It helps designers prioritize optimization efforts and focus on critical paths to ensure that the overall circuit meets timing requirements.

Path delay analysis considers the combination of logic gates and interconnects along specific signal paths, typically from input to output or between specific registers.

The purpose of path delay analysis is to identify and optimize the timing of critical paths to ensure that the circuit meets its timing constraints and performance requirements.

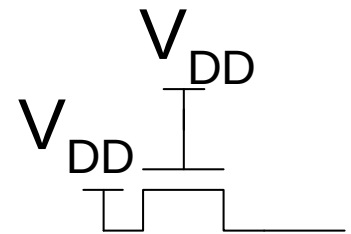
Pass Transistors

The pass transistor network is typically created during the physical design stage of VLSI (Very Large Scale Integration) analysis. This stage involves translating the logical representation of the circuit (which is typically described in a hardware description language like Verilog or VHDL) into the physical layout of transistors and interconnections on the silicon die.

We have assumed source is grounded

What if source > 0 ?

e.g. pass transistor passing V_{DD}



A pull-up switch is used to connect a node to a high voltage (logic '1') when activated, while a pull-down switch connects a node to ground (logic '0') when activated.

NMOS Pass Transistors

We have assumed source is grounded

What if source > 0 ?

e.g. pass transistor passing V_{DD}

Let $V_g = V_{DD}$

Now if $V_s > V_{DD} - V_t$, $V_{gs} < V_t$

Hence transistor would turn itself off

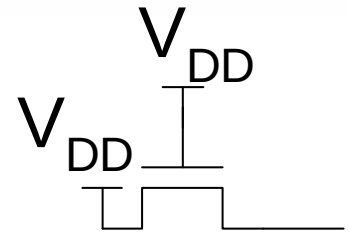
NMOS pass transistors **pull-up** no higher than $V_{DD} - V_{tn}$

Called a degraded “1”

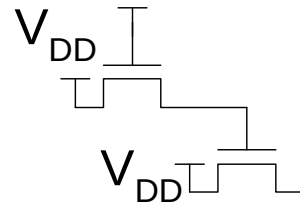
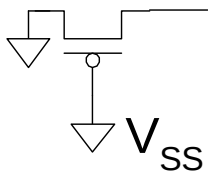
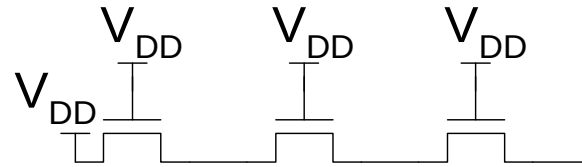
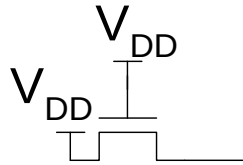
Approach degraded value slowly (low I_{ds})

PMOS pass transistors **pull-down** no lower than V_{tp}

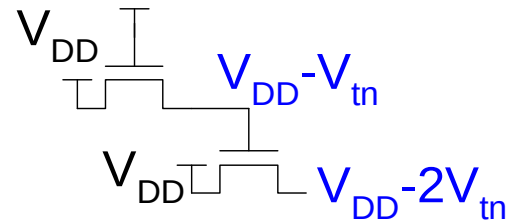
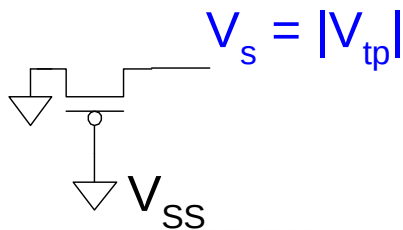
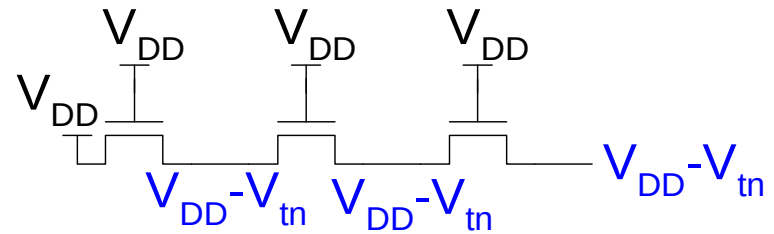
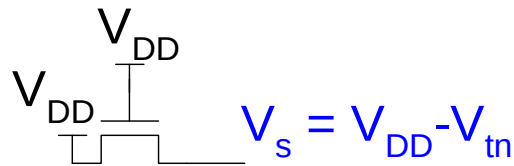
Called a degraded “0”



Pass Transistor Ckts



Pass Transistor Ckts



Inverter Delay Estimation

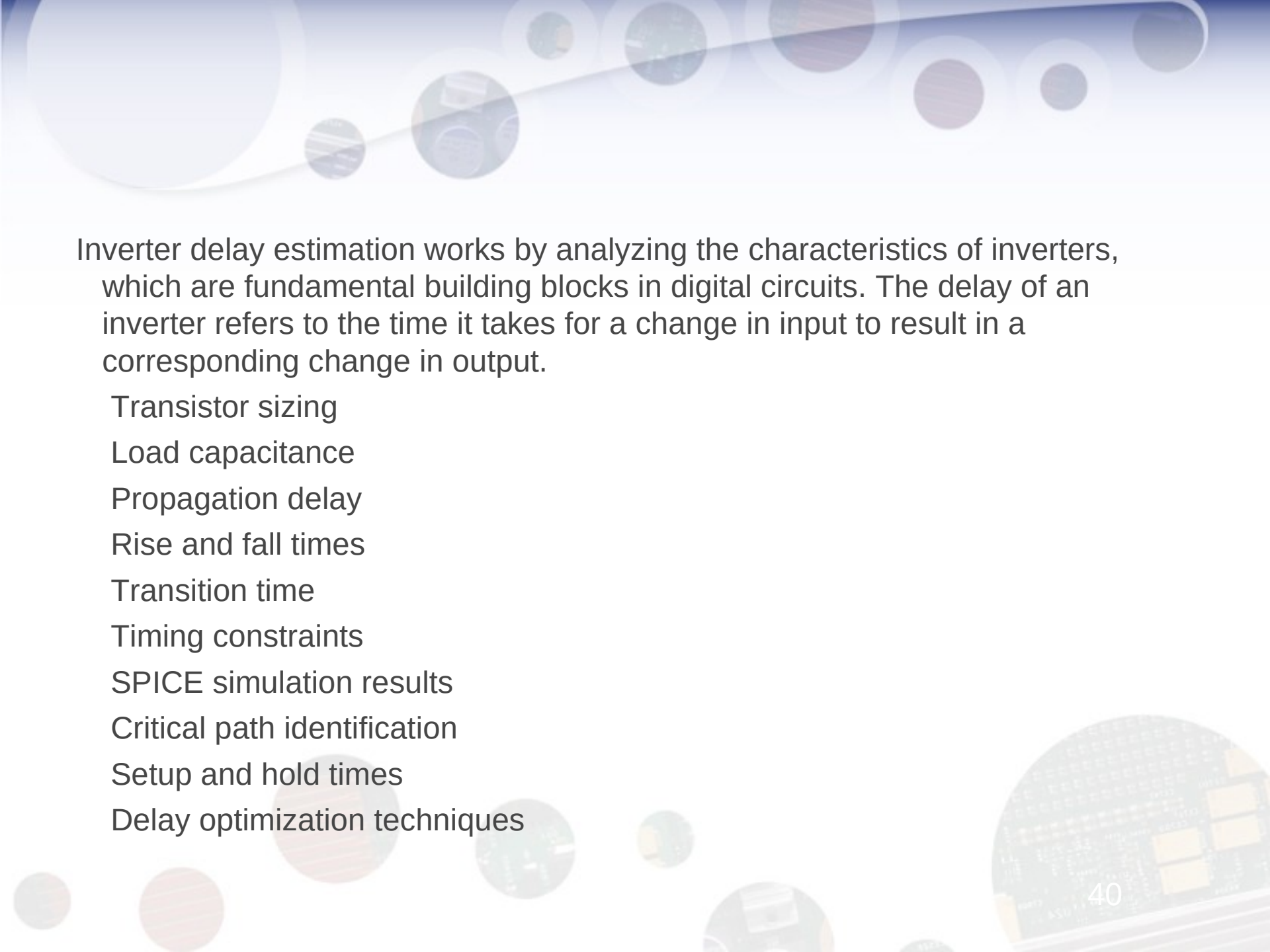
Inverter delay estimation focuses specifically on estimating the delay characteristics of inverters, which are fundamental building blocks in digital circuits.

It helps designers understand the timing behavior of individual inverters and optimize their sizing and configuration to meet timing constraints.

The purpose of inverter delay estimation is to accurately predict the delay of inverters within the circuit, which is essential for overall timing analysis and optimization.

Inverter delay estimation

Inverter delay estimation is closely related to ASIC design, particularly in the context of layout and physical design verification processes such as LVS (Layout vs. Schematic) and DRC (Design Rule Checking).



Inverter delay estimation works by analyzing the characteristics of inverters, which are fundamental building blocks in digital circuits. The delay of an inverter refers to the time it takes for a change in input to result in a corresponding change in output.

Transistor sizing

Load capacitance

Propagation delay

Rise and fall times

Transition time

Timing constraints

SPICE simulation results

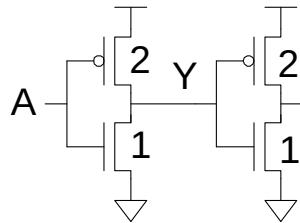
Critical path identification

Setup and hold times

Delay optimization techniques

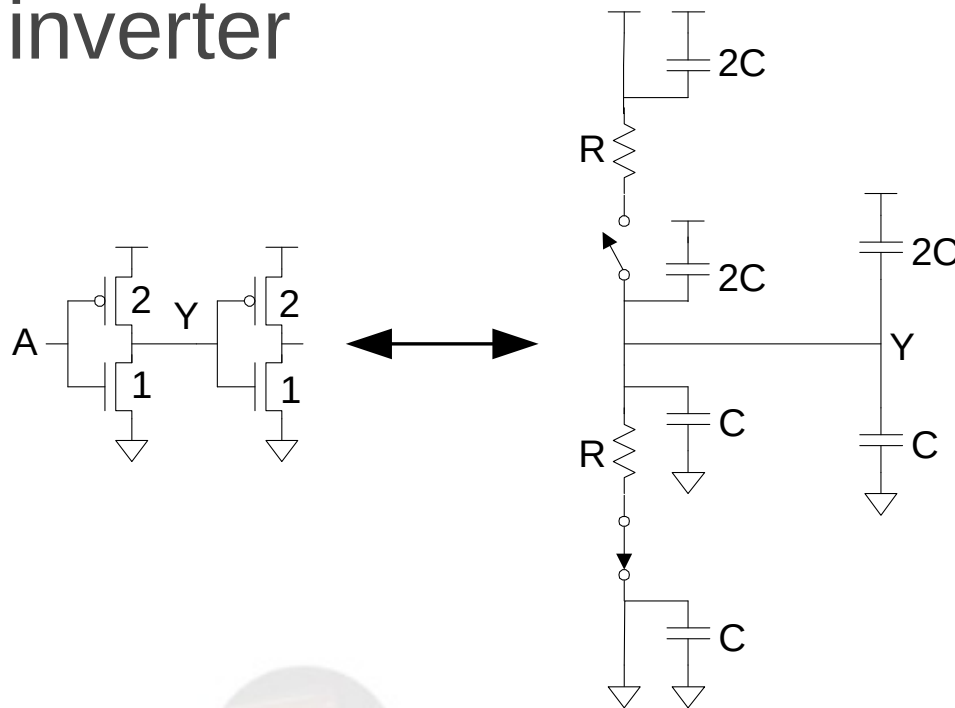
Modeling: Inverter Delay Estimate

Model and estimate the delay of a fanout-of-1 inverter



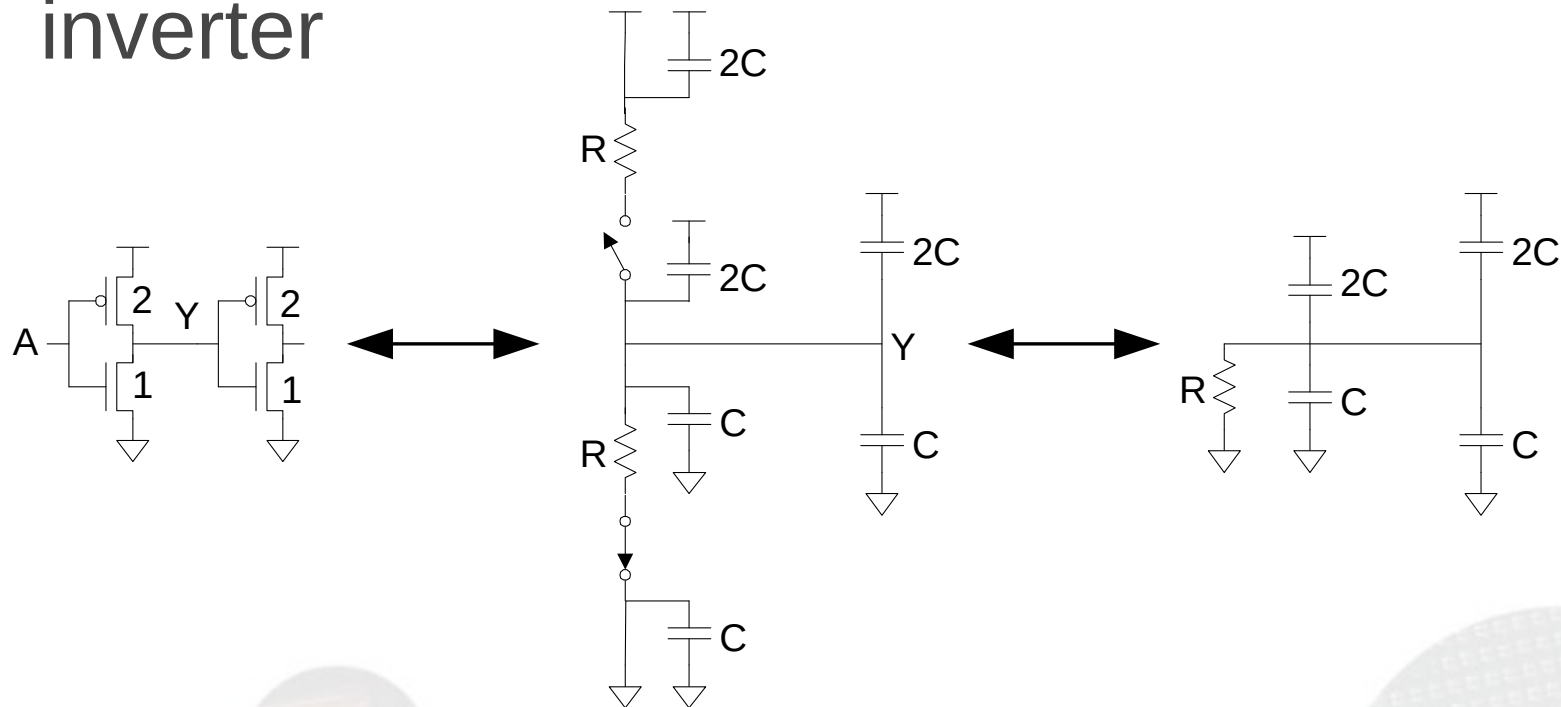
Inverter Delay Estimate

Estimate the delay of a fanout-of-1 inverter

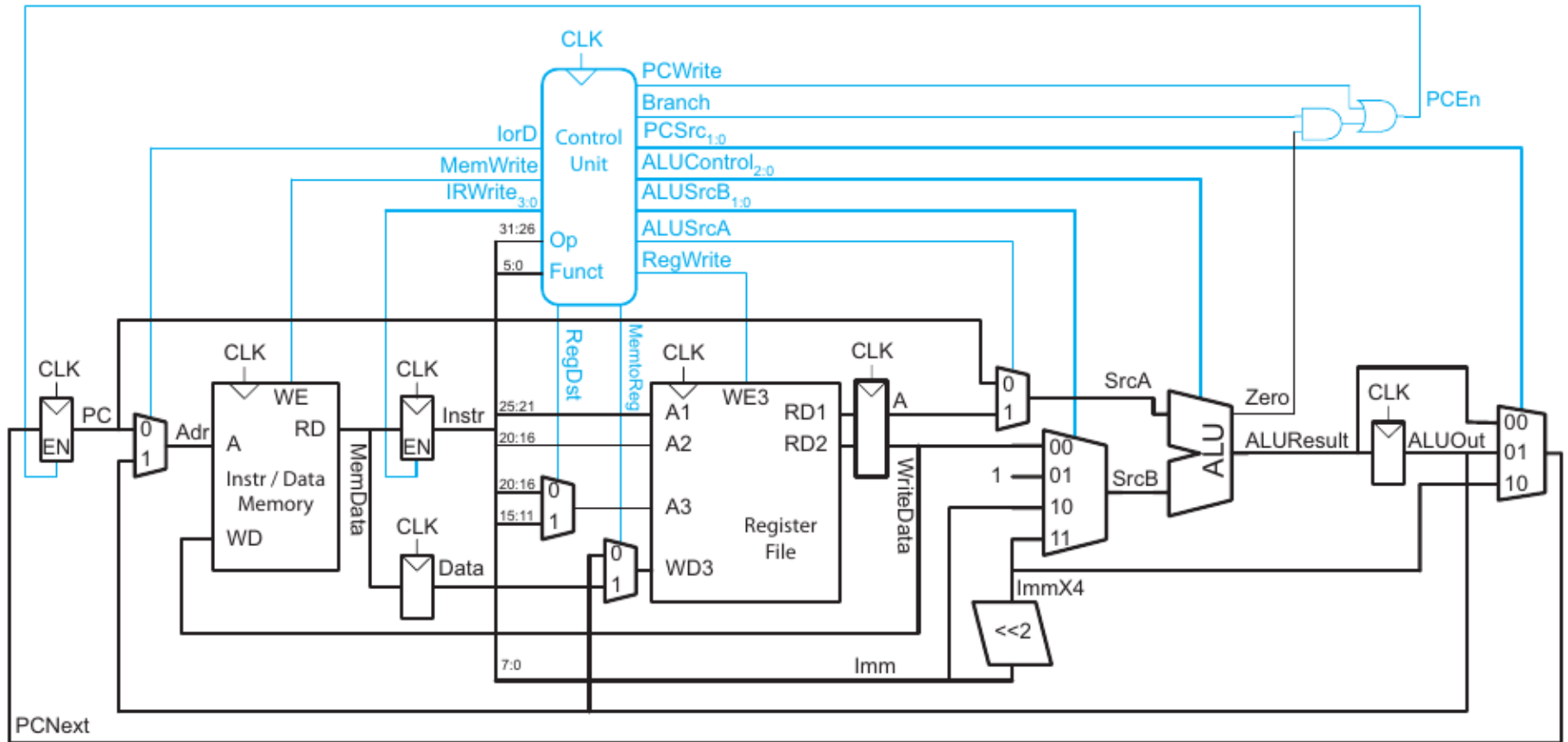


Inverter Delay Estimate

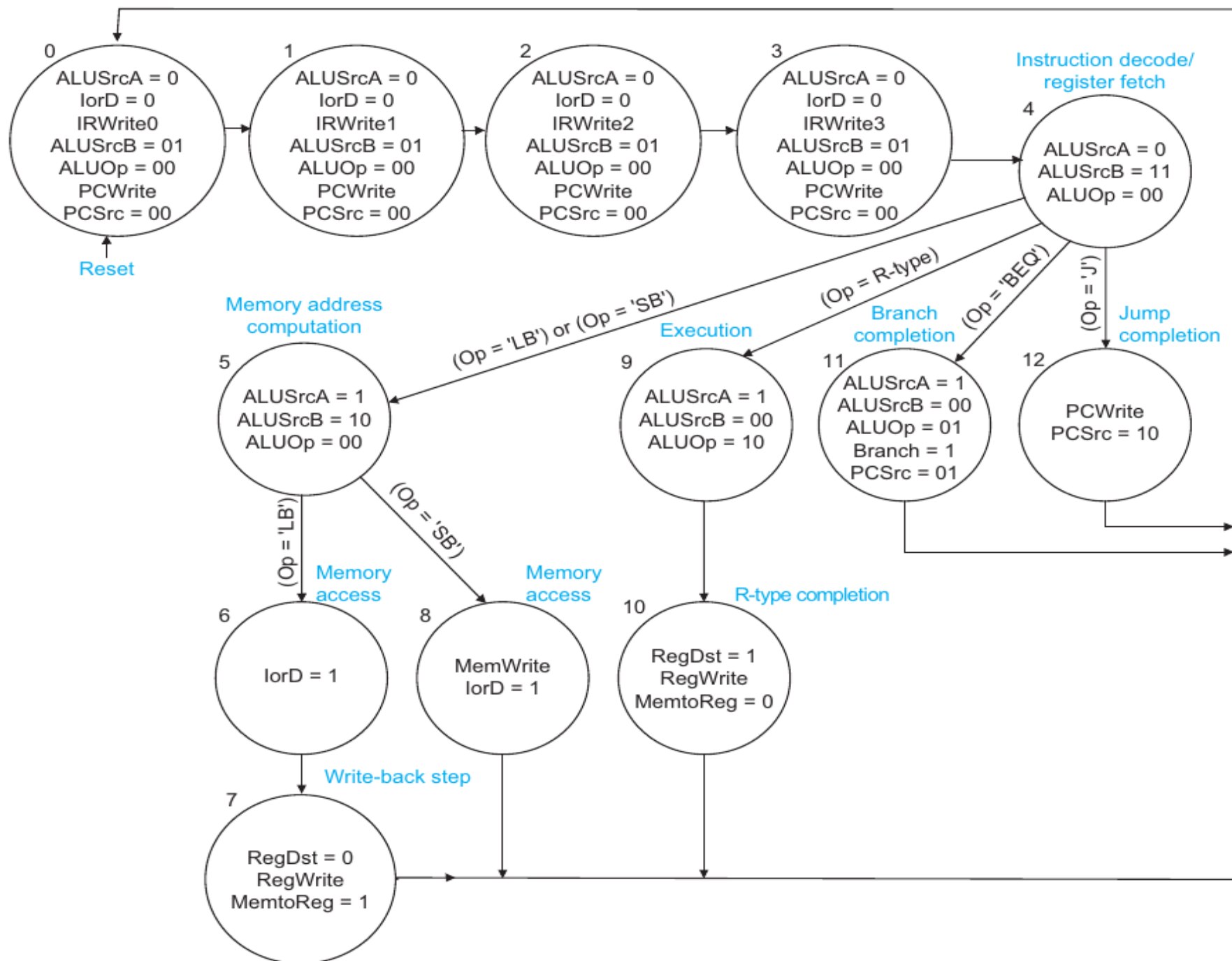
Estimate the delay of a fanout-of-1 inverter

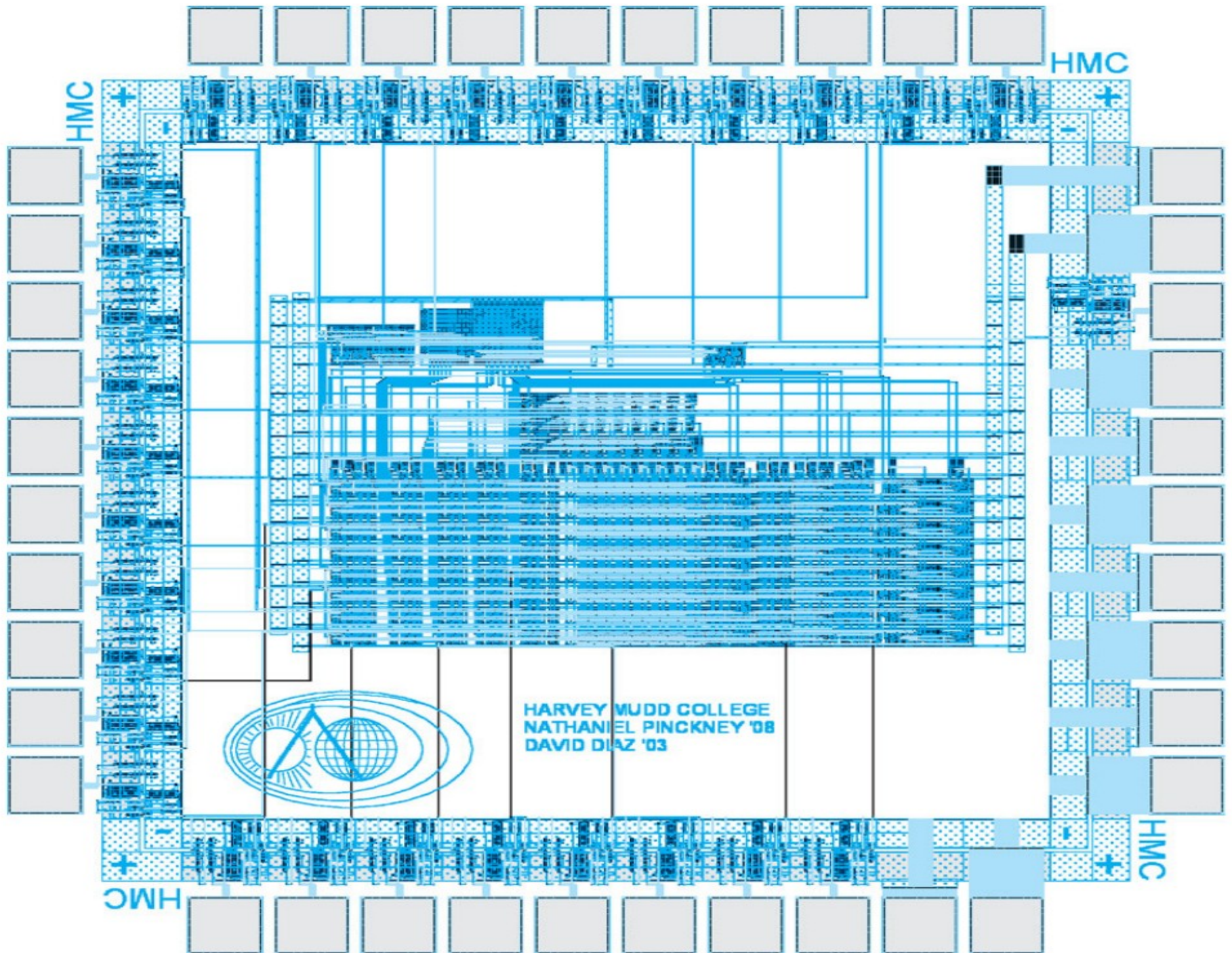


Task Discussion



Instruction fetch





HMC

HMC

HMC

HMC



HARVEY MUDD COLLEGE
NATHANIEL PINCKNEY '08
DAVID DIAZ '03

Synthesis Layout

