Lecture 2: Computer Elements

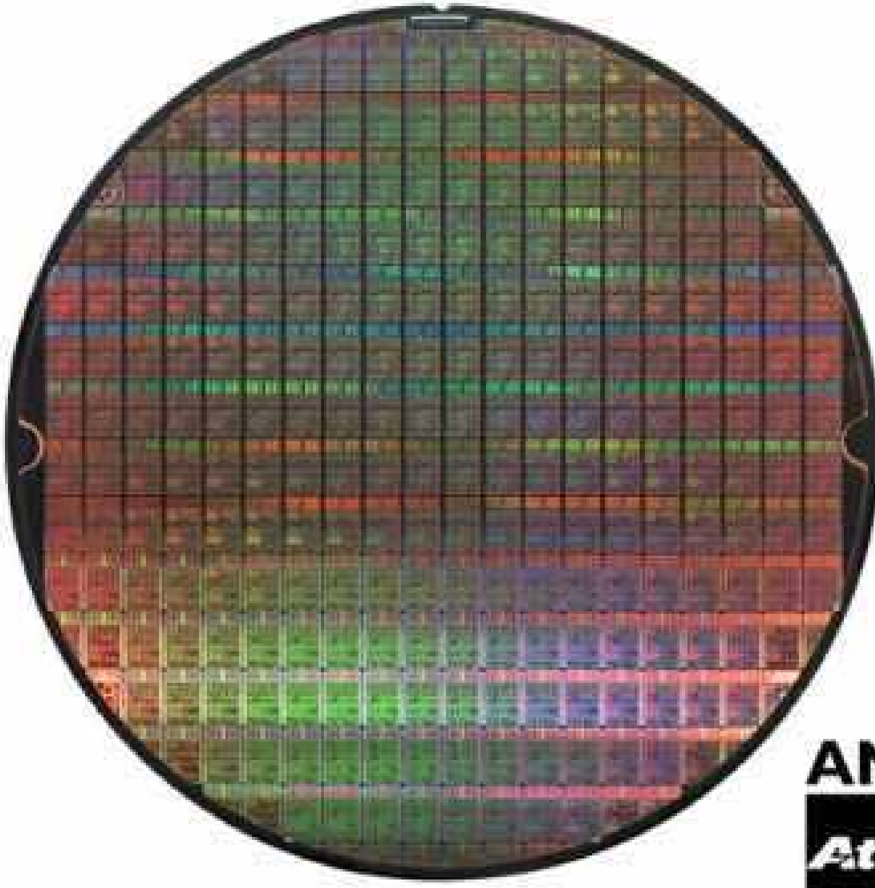
- Last Time
  - Course Overview
  - Introduction to Computer Architecture

- Today
  - Announcements
  - Computer elements
    - Transistors, wires, pins
  - Introduction to performance

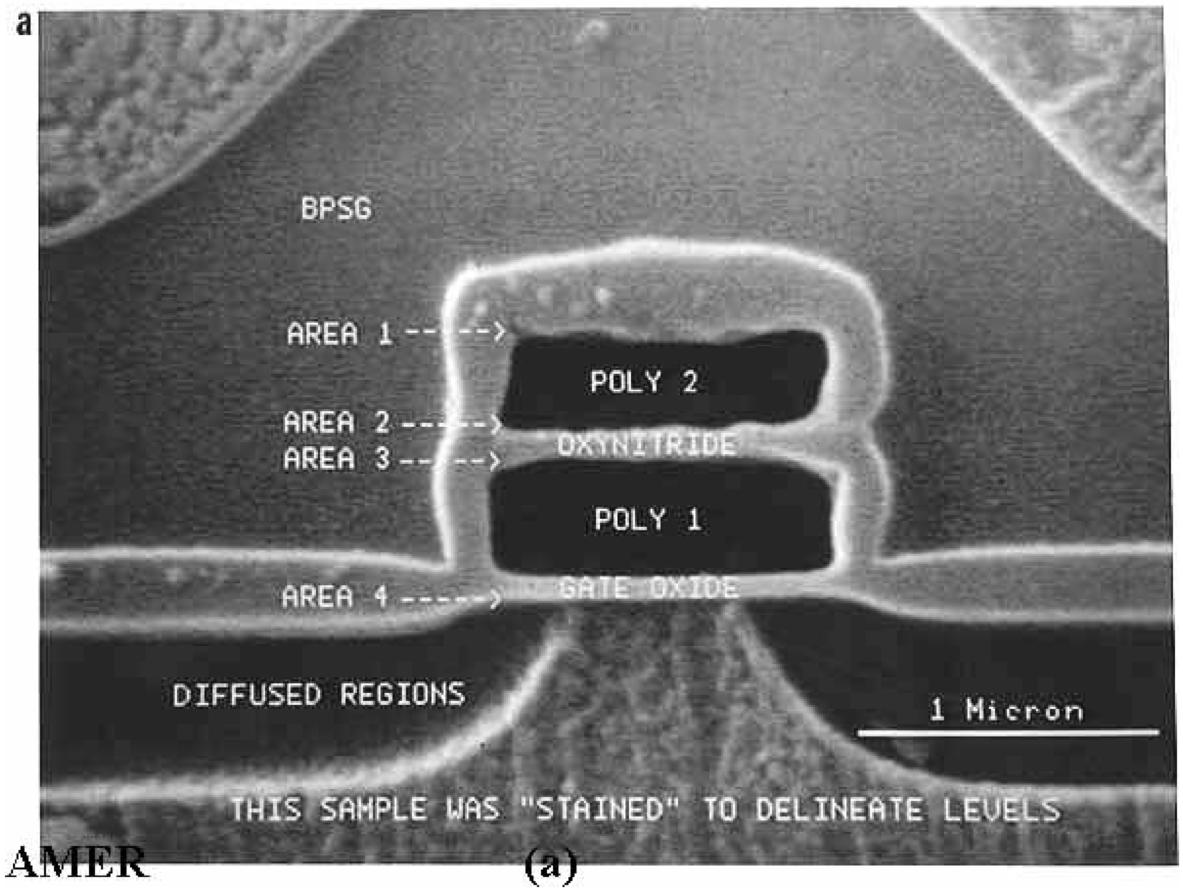
Computer Elements

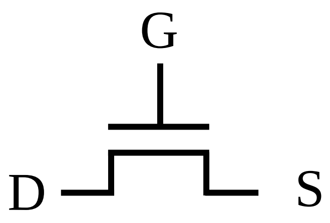
- Transistors (computing)
  - How can they be connected to do something useful?
  - How do we evaluate how fast a logic block is?
- Wires (transporting)
  - What and where are they?
  - How can they be modeled?
- Memories (storing)
  - SRAM vs. DRAM

What Comes out of the Fab?



The Mighty Transistor!

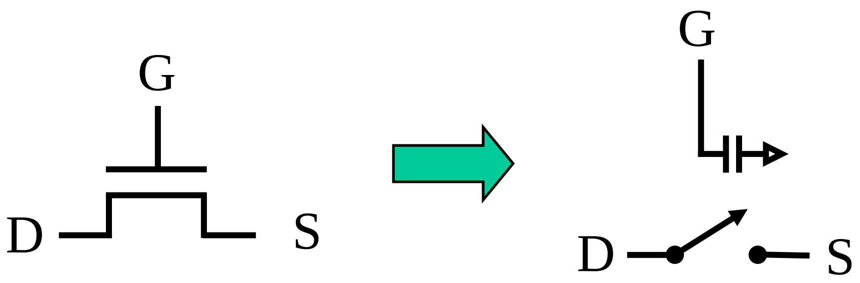


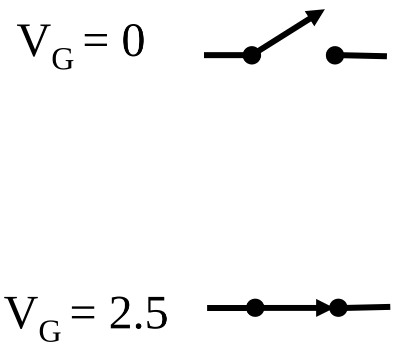


Transistor As a Switch

- Ideal Voltage Controlled Switch

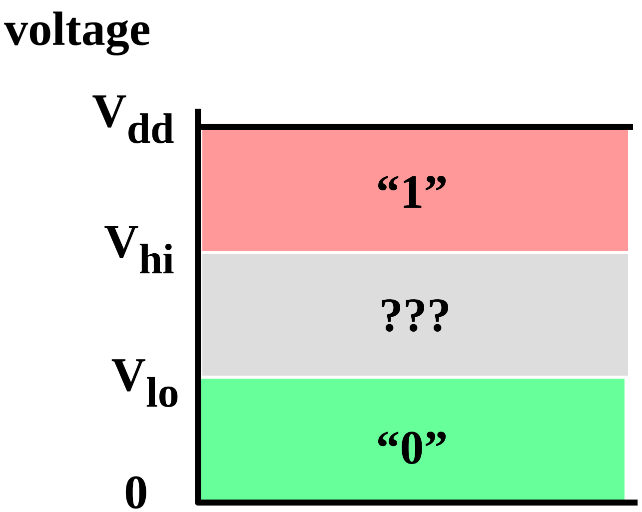
- Three terminals
  - Gate
  - Drain
  - Source





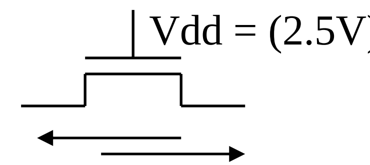
Abstractions in Logic Design

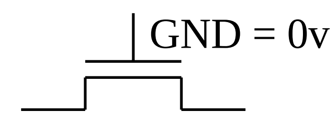
- In physical world
  - Voltages, Currents
  - Electron flow
- In logical world - abstraction
  - $V < V_{lo} \Rightarrow$ "0" = FALSE
  - $V > V_{hi} \Rightarrow$ "1" = TRUE
  - In between - forbidden
- Simplify design problem



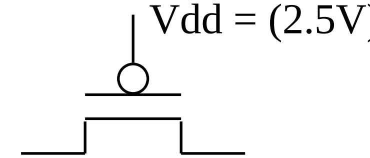
Basic Technology: CMOS

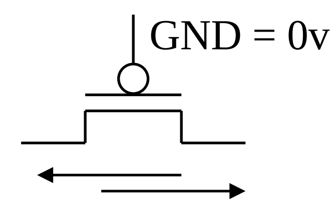
- CMOS: Complementary Metal Oxide Semiconductor
  - NMOS (N-Type Metal Oxide Semiconductor) transistors
  - PMOS (P-Type Metal Oxide Semiconductor) transistors
- NMOS Transistor
  - Apply a HIGH (Vdd) to its gate
    turns the transistor into a “conductor”
  - Apply a LOW (GND) to its gate
    shuts off the conduction path





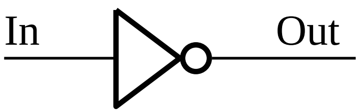
- PMOS Transistor
  - Apply a HIGH (Vdd) to its gate
    shuts off the conduction path
  - Apply a LOW (GND) to its gate
    turns the transistor into a “conductor”



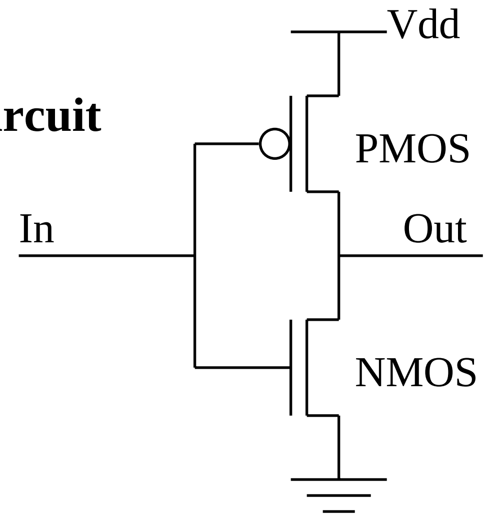


Basic Components: CMOS Inverter

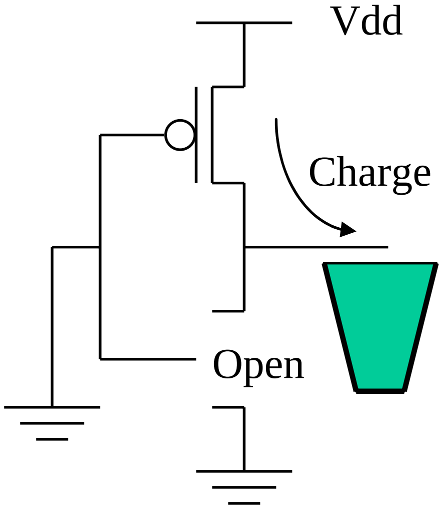
Symbol

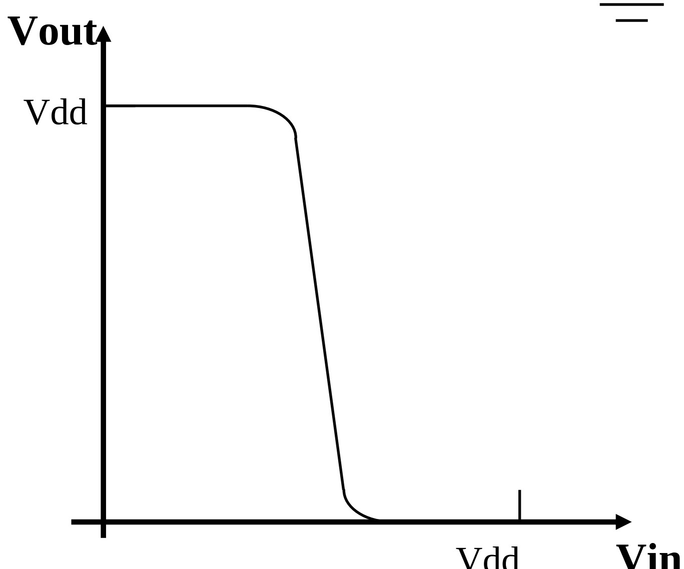


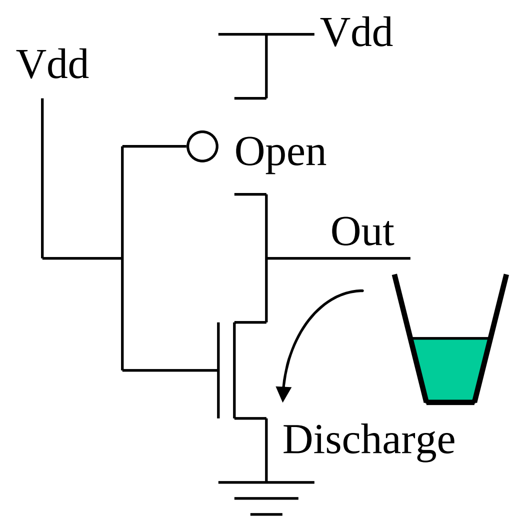
Circuit



- Inverter Operation

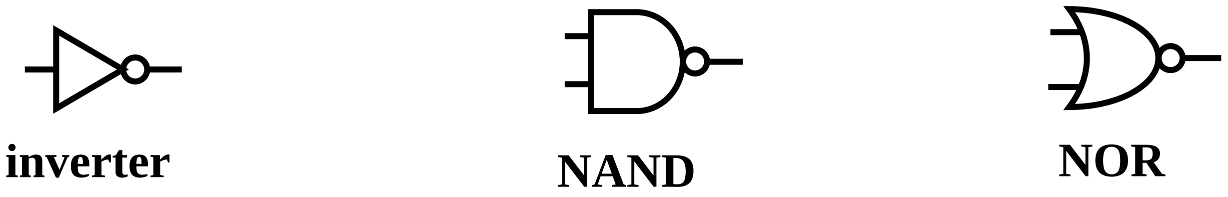






Composition of Transistors

- Logic Gates
  - Inverters, AND, OR, arbitrary

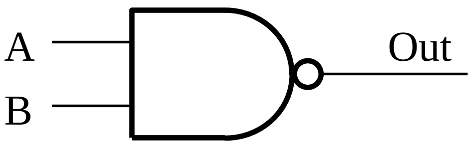


- Buffers (drive large capacitances, long wires, etc.)

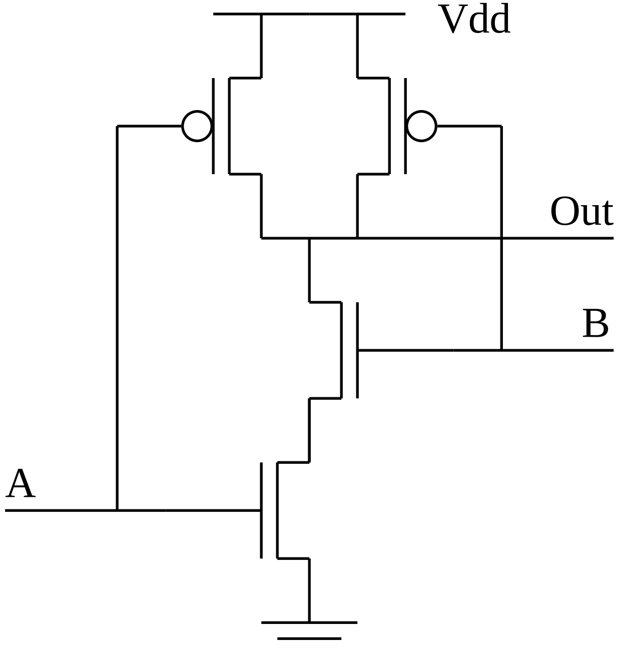
- Memory elements
  - Latches, registers, SRAM, DRAM

Basic Components: CMOS Logic Gates

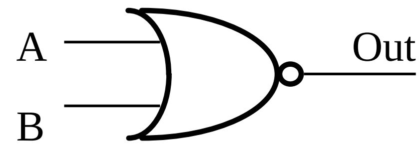
NAND Gate



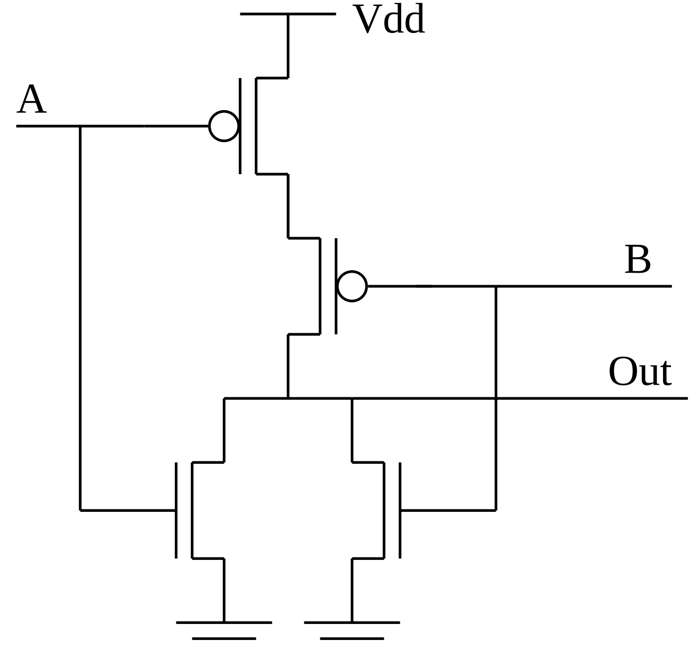
A	B	Out
0	0	1
0	1	1
1	0	1
1	1	0



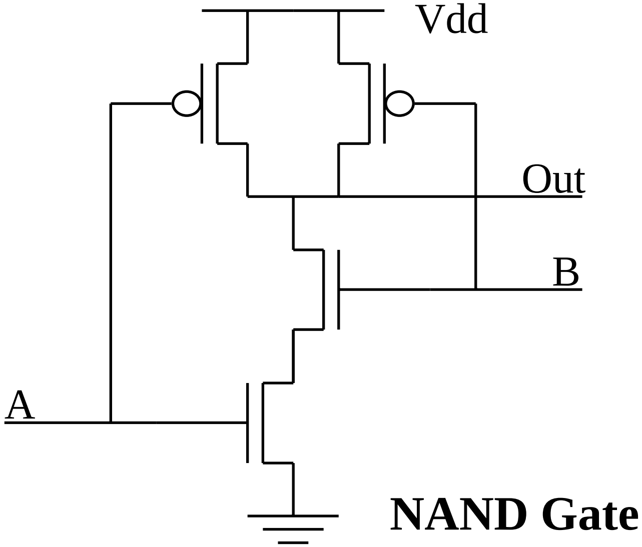
NOR Gate

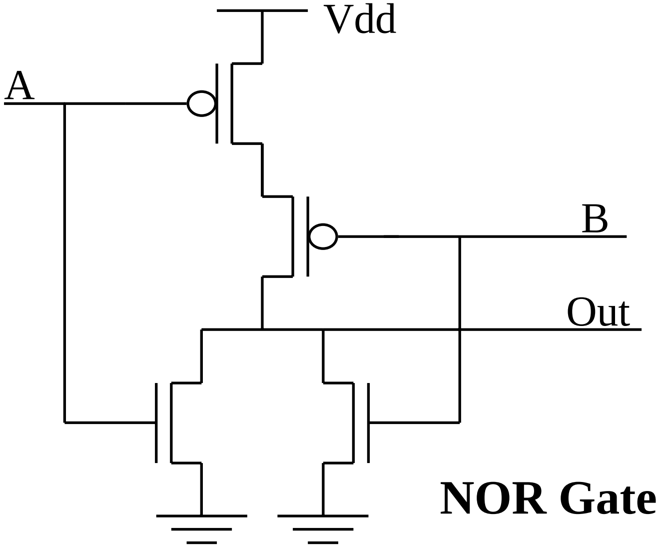


A	B	Out
0	0	1
0	1	0
1	0	0
1	1	0



Gate Comparison





- If PMOS transistors is faster:
  - It is OK to have PMOS transistors in series
  - NOR gate is preferred
  - NOR gate is preferred also if H -> L is more critical than L -> H
- If NMOS transistors is faster:
  - It is OK to have NMOS transistors in series
  - NAND gate is preferred
  - NAND gate is preferred also if L -> H is more critical than H -> L

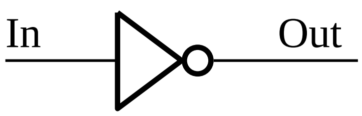
The Ugly Truth

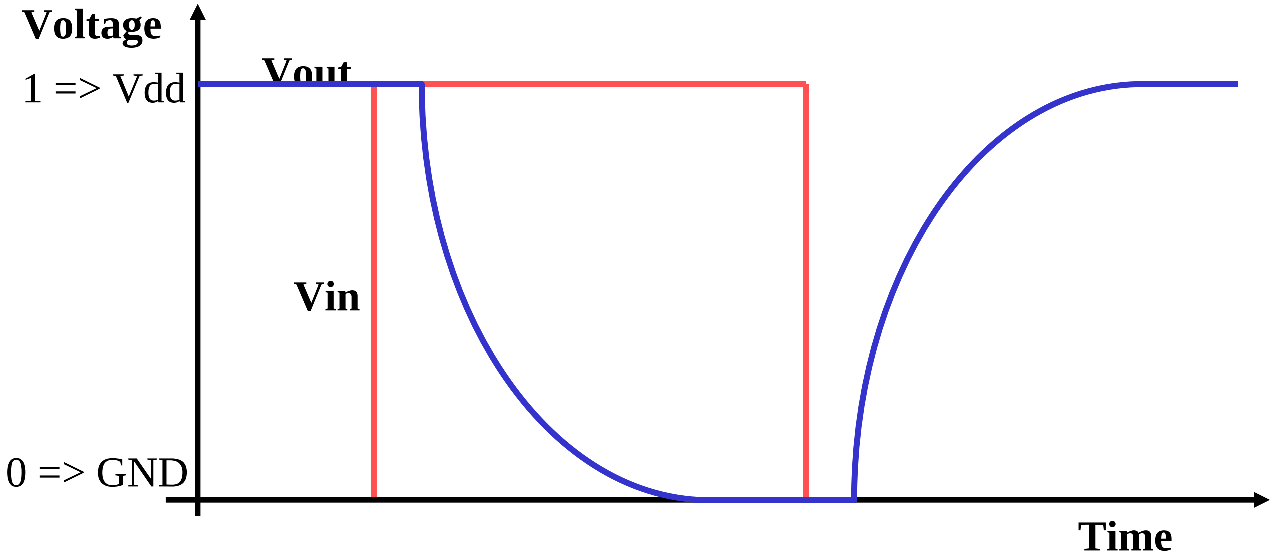
- Transistors are not ideal switches!
  - Gate Capacitance (C_g)
  - Source-to-Drain resistance (R)
  - Drain capacitance
- Issues
  - Delay - actually takes real time to turn transistors on and off
  - Power/Energy
  - Noise (from transistors, power rails)

- But - we can change transistor size
  - Increase C_g, but decrease R

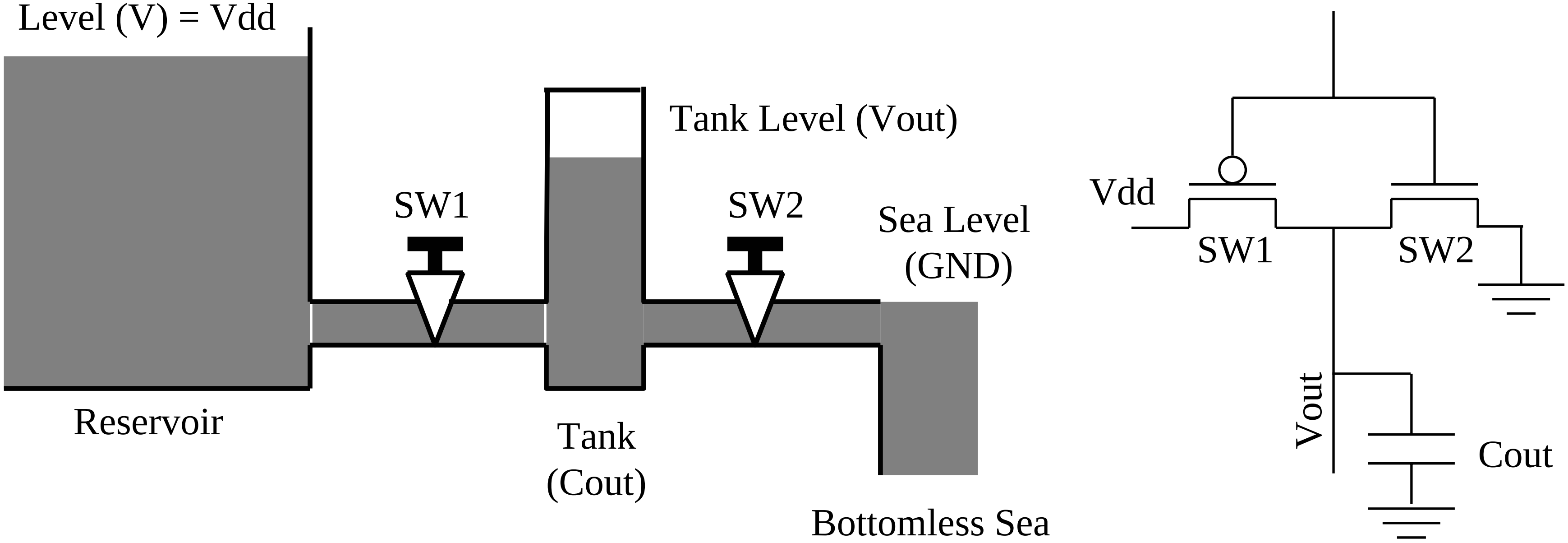
Ideal (CS) versus Reality (EE)

- When input 0 -> 1, output 1 -> 0 but NOT instantly
  - Output goes 1 -> 0: output voltage goes from Vdd (2.5v) to 0v
- When input 1 -> 0, output 0 -> 1 but NOT instantly
  - Output goes 0 -> 1: output voltage goes from 0v to Vdd (2.5v)
- Voltage does not like to change instantaneously



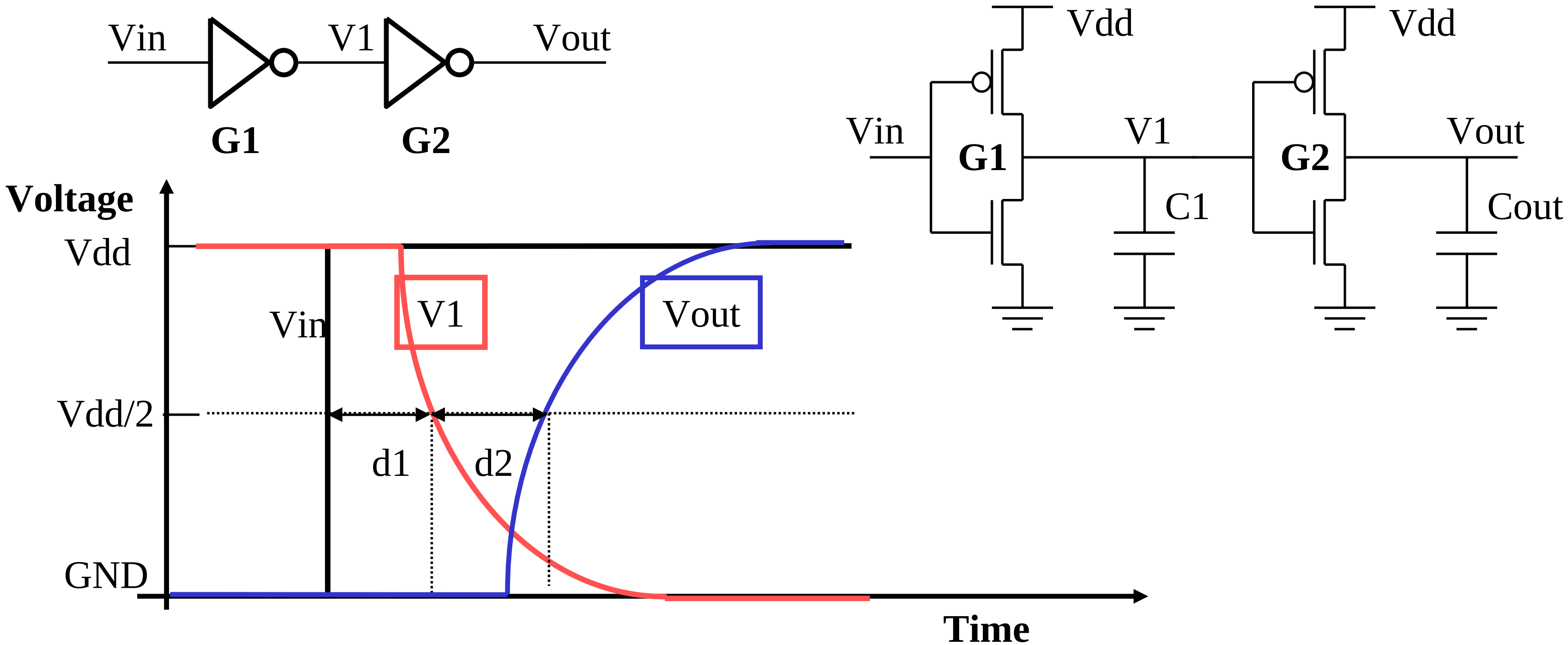


Fluid Timing Model



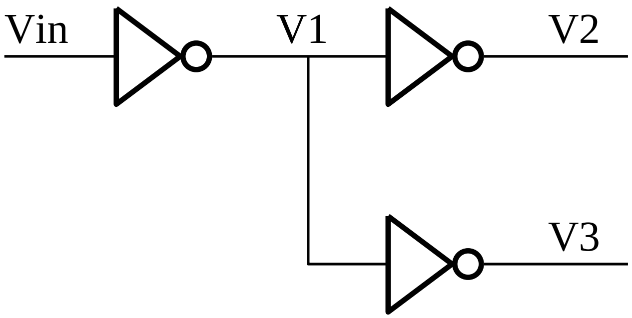
- Water <-> Electrical Charge Tank Capacity <-> Capacitance (C)
- Water Level <-> Voltage Water Flow <-> Charge Flowing (Current)
- Size of Pipes <-> Strength of Transistors (G)
- Time to fill up the tank ~ C / G **Resistance R = 1/G**

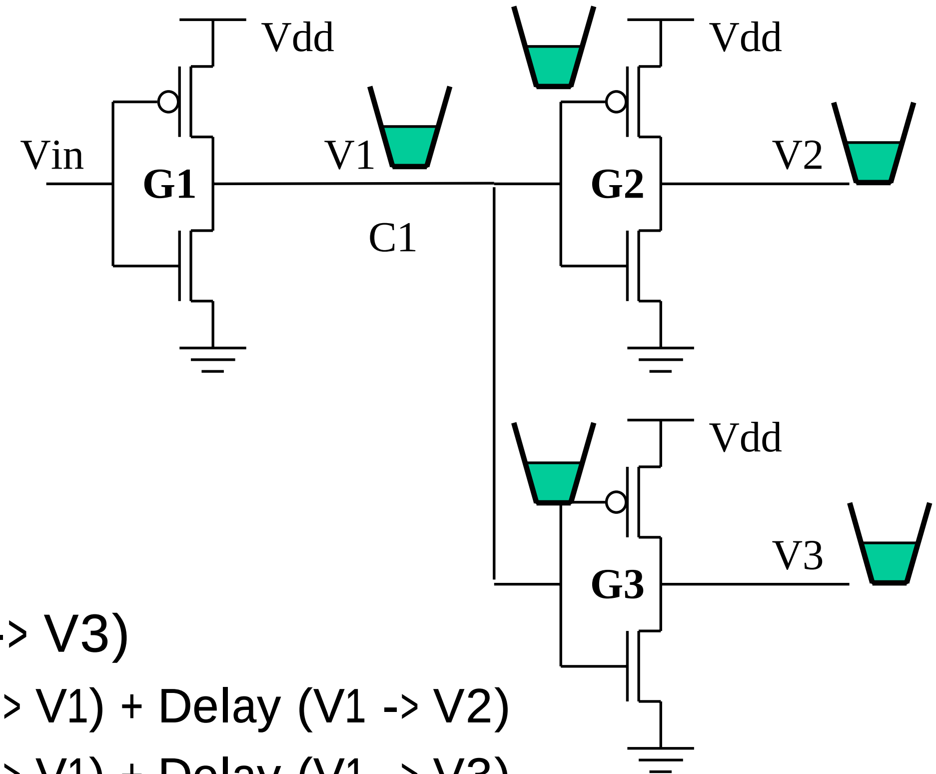
Series Connection



- Total Propagation Delay = Sum of individual delays = d1 + d2
- Capacitance C1 has two components:
  - Capacitance of the wire connecting the two gates
  - Input capacitance of the second inverter

Review: Calculating Delays

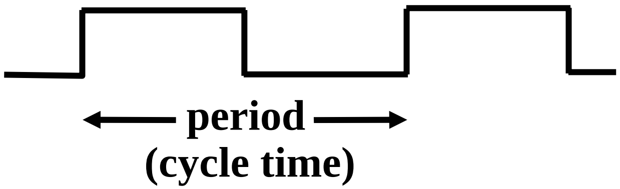




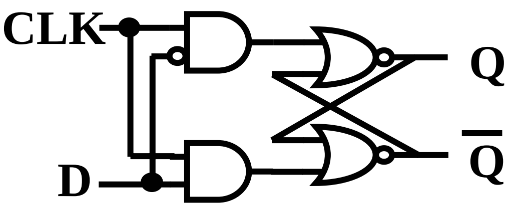
- Sum delays along serial paths
- Delay (Vin -> V2) ! = Delay (Vin -> V3)
  - Delay (Vin -> V2) = Delay (Vin -> V1) + Delay (V1 -> V2)
  - Delay (Vin -> V3) = Delay (Vin -> V1) + Delay (V1 -> V3)
- Critical Path = The longest among the N parallel paths
- C1 = Wire C + Cin of Gate 2 + Cin of Gate 3

Clocking and Clocked Elements

- Typical Clock
  - 1Hz = 1 cycle per second

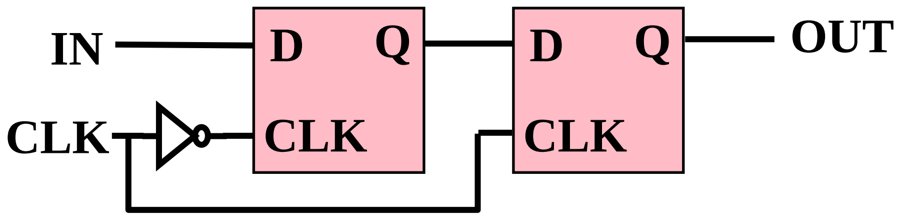


- Transparent Latch

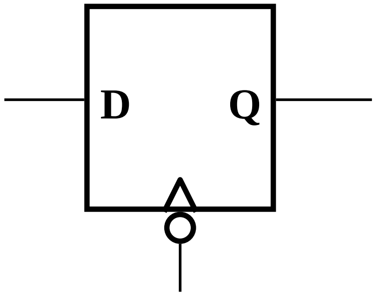


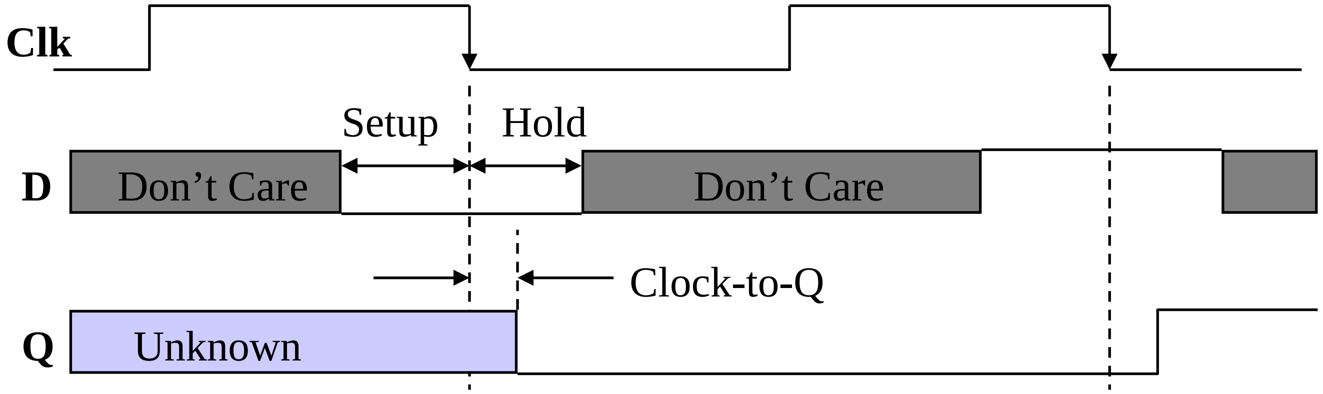
CLK=0, Q=oldQ
CLK=1, Q=D

- Edge Triggered Flip-Flop



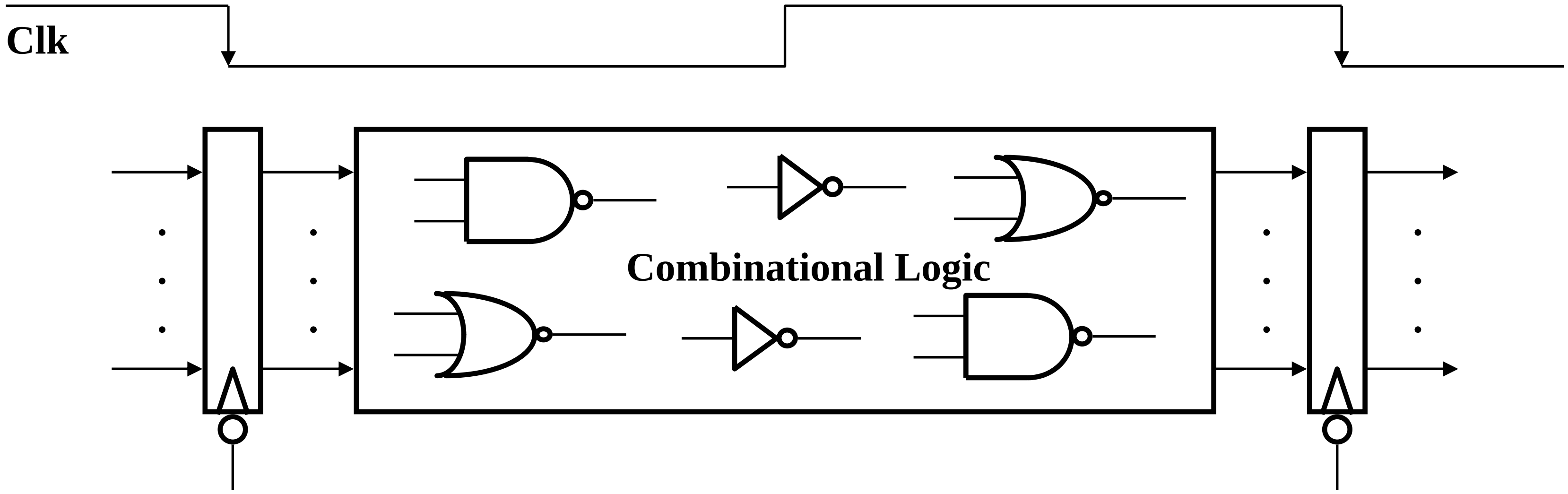
Storage Element's Timing Model





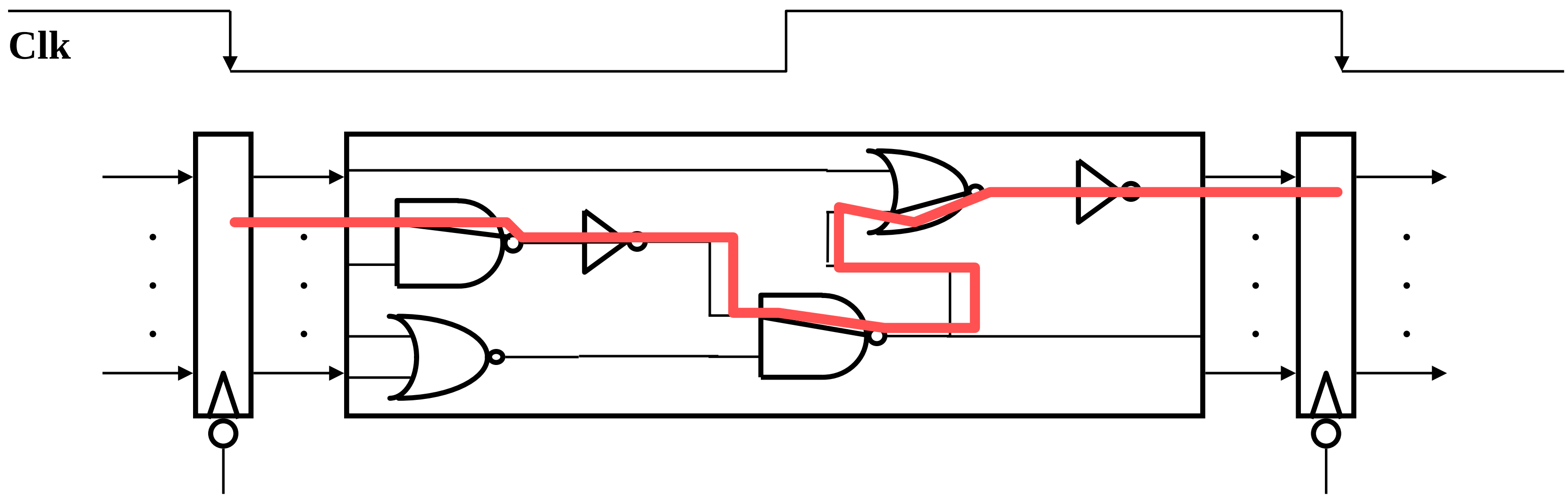
- Setup Time: Input must be stable BEFORE the trigger clock edge
- Hold Time: Input must REMAIN stable after the trigger clock edge
- Clock-to-Q time:
  - Output cannot change instantaneously at the trigger clock edge
  - Similar to delay in logic gates, two components:
    - Internal Clock-to-Q
    - Load dependent Clock-to-Q

Clocking Methodology



- All storage elements are clocked by the same clock edge
- The combination logic block's:
  - Inputs are updated at each clock tick
  - All outputs MUST be stable before the next clock tick

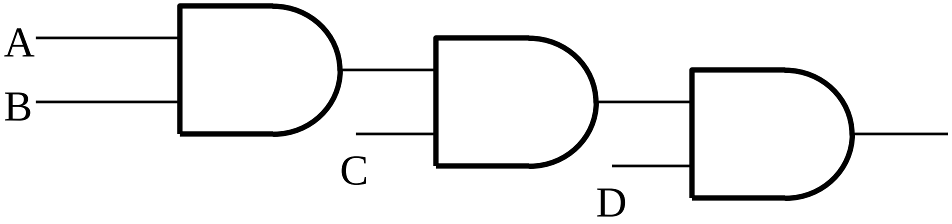
Critical Path & Cycle Time

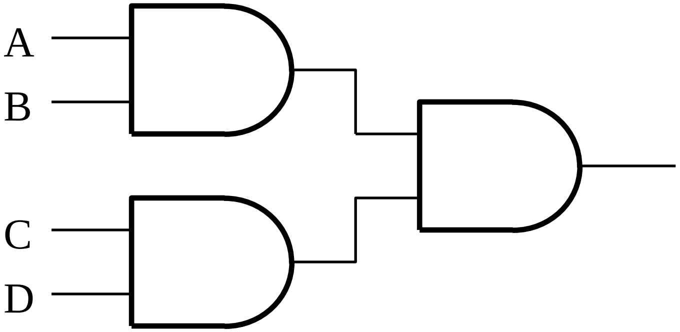


- Critical path: the slowest path between any two storage devices
- Cycle time is a function of the critical path
- must be greater than:
  - Clock-to-Q + Longest Path through the Combination Logic + Setup

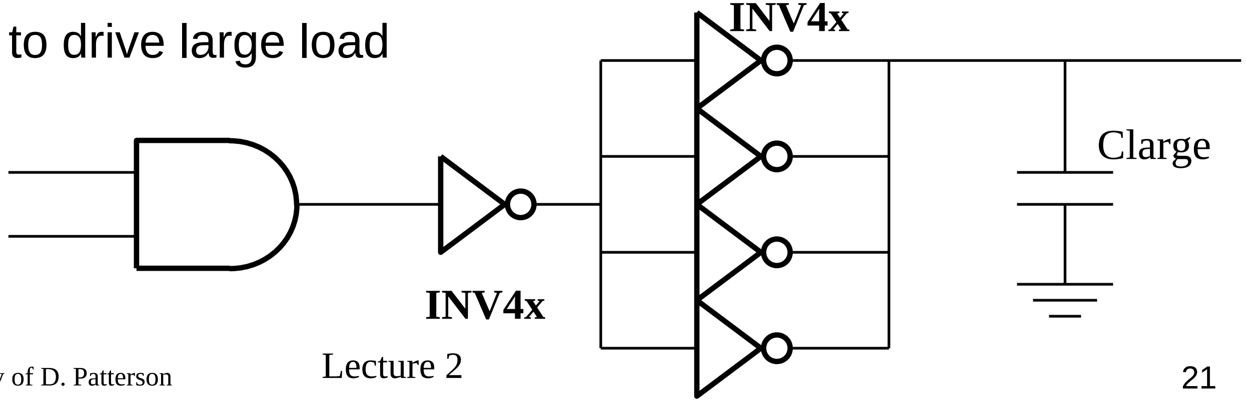
Tricks to Reduce Cycle Time

- Reduce the number of gate levels



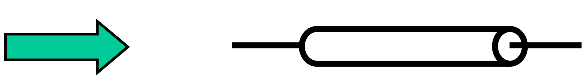


- Pay attention to loading
  - One gate driving many gates is a bad idea
  - Avoid using a small gate to drive a long wire
- Use multiple stages to drive large load



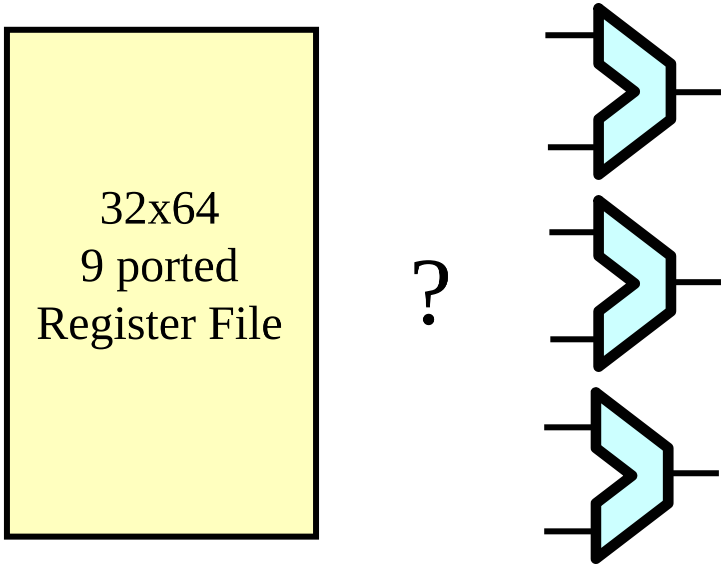
Wires

- Limiting Factor
  - Density
  - Speed
  - Power

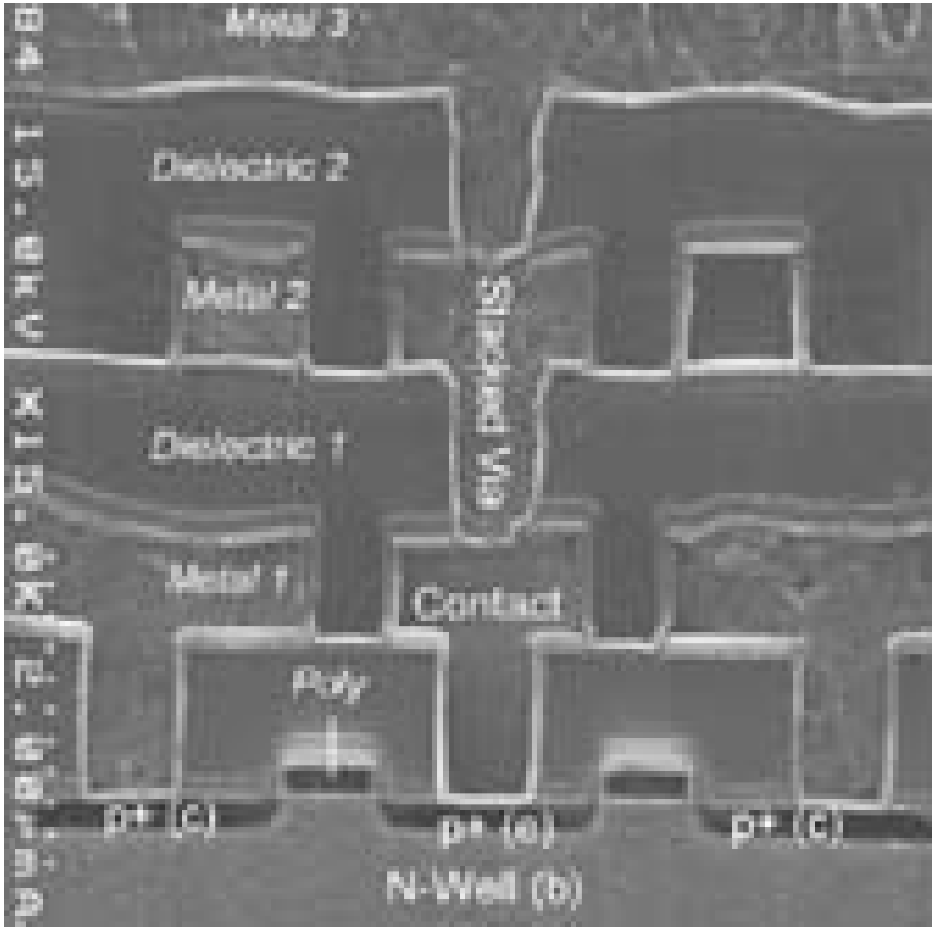
- 3 models for wires (model to use depends on switching frequency)
  - Short 
  - Lossless 
  - Lossy 

Wire Density

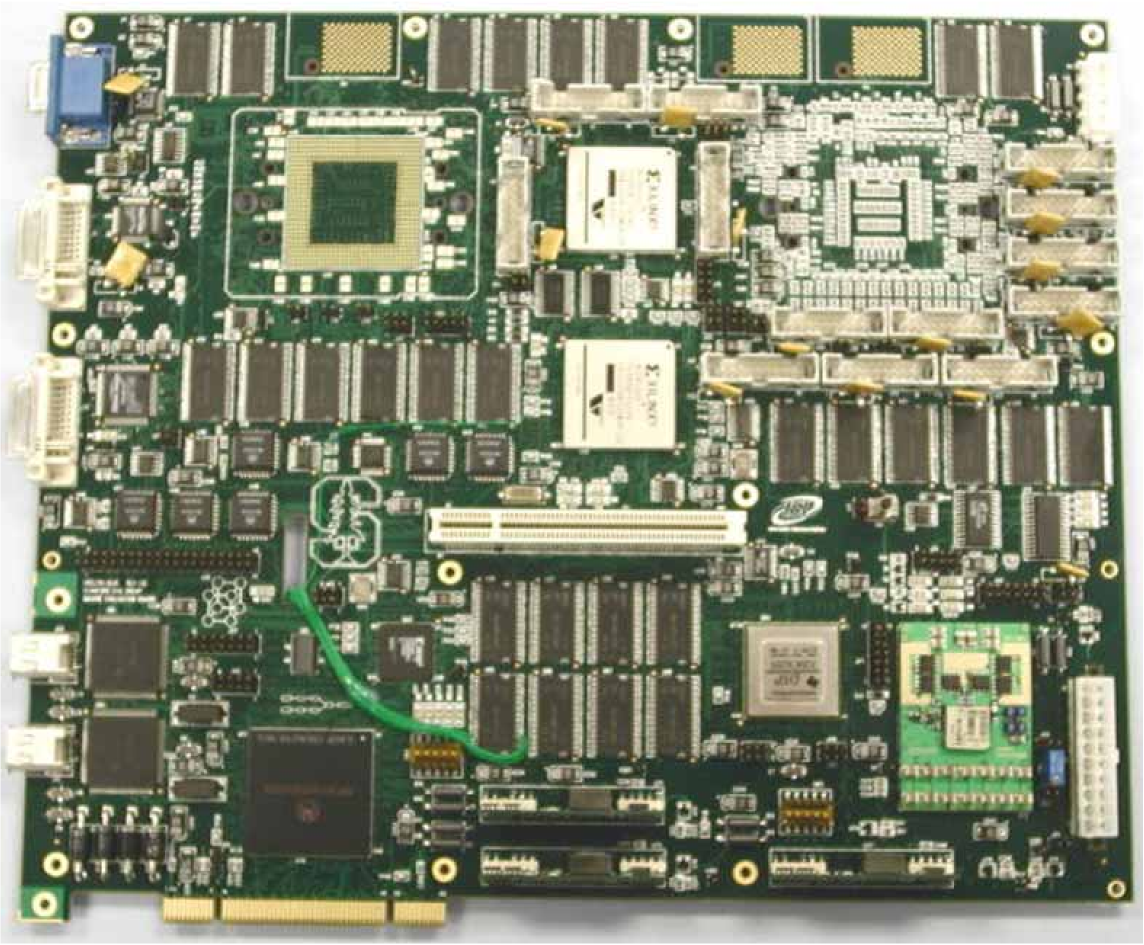
- Communication constraints
  - Must be able to move bits to/from storage and computation elements
- Example: 9 ported register file



Chip Level

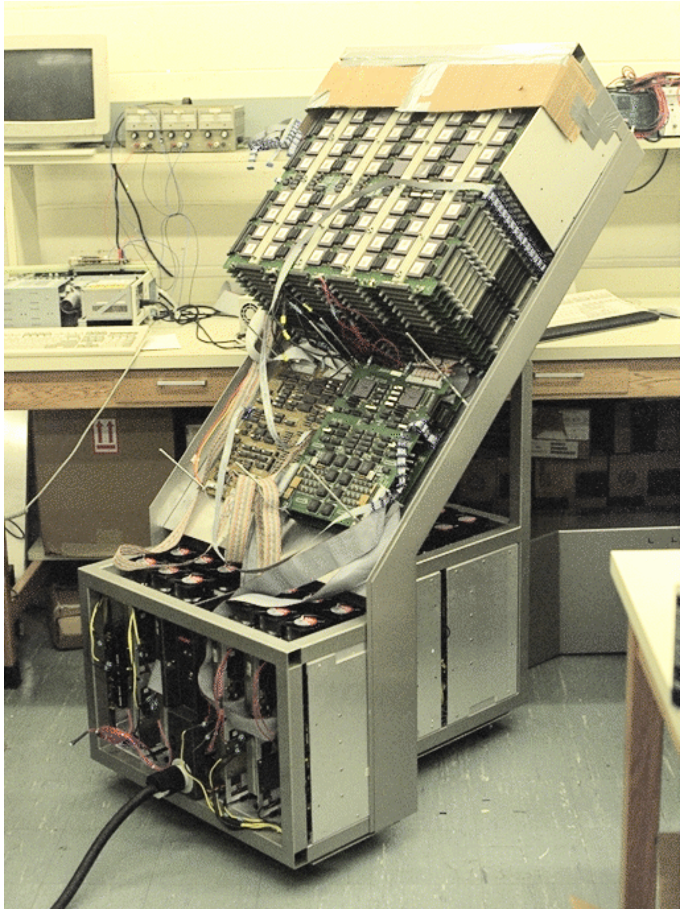

Metal 3
Dielectric 2
Metal 2
Stacked Via
Dielectric 1
Metal 1
Contact
Poly
p+ (c)
p+ (a)
p+ (c)
N-Well (b)

Board Level



Stanford Imagine Board

Rack Level

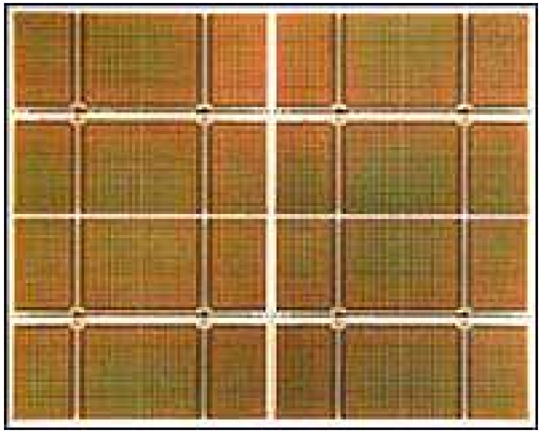


MIT J-Machine

DOE ASCI White

Memory

- Moves information in time (wires move it in space)
- Provides state
- Requires energy to change state
  - Feedback circuit - SRAM
  - Capacitors – DRAM
  - Magnetic media - disk
- Required for memories
  - Storage medium
  - Write mechanism
  - Read mechanism



4Gb DRAM Die

Technology Scaling Trends

- CPU Transistor density – 60% per year
- CPU Transistor speed – 15% per year
- DRAM density – 60% per year
- DRAM speed – 3% per year

- On-chip wire speed – decreasing relative to transistors (witness the Pentium 4 pipeline)
- Off-chip pin bandwidth – increasing, but slowly
- Power – approaching costs limits
  - $P = CV^2 f + I_{leak} V$

- All of these factors affect the end system architecture

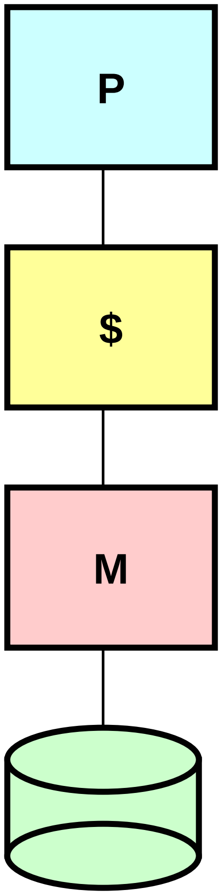
Summary

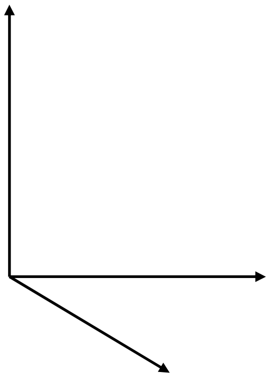
- Logic Transistors + Wires + Storage = Computer!
- Transistors
  - Composable switches
  - Electrical considerations
    - Delay from parasitic capacitors and resistors
    - Power ($P = CV^2 f$)
- Wires
  - Becoming more important from delay and BW perspective
- Memories
  - Density, Access time, Persistence, BW

Performance Measurement and Evaluation

Many Dimensions to Performance

- CPU execution time
  - by instruction or sequence
    - floating point
    - integer
    - branch performance
- Cache bandwidth
- Main memory bandwidth
- I/O performance
  - bandwidth
  - seeks
  - pixels or polygons per second
- Relative importance depends on applications



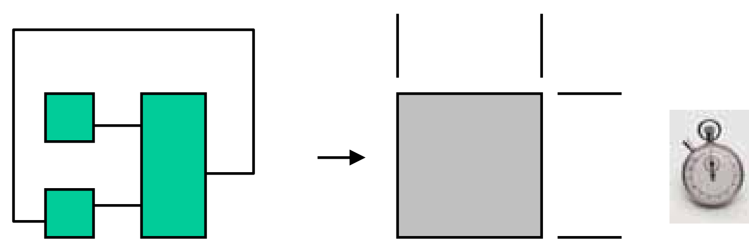


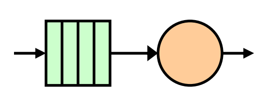
Evaluation Tools

- Benchmarks, traces, & mixes
  - macrobenchmarks & suites
    - application execution time
  - microbenchmarks
    - measure one aspect of performance
  - traces
    - replay recorded accesses
    - cache, branch, register
- Simulation at many levels
  - ISA, cycle accurate, RTL, gate, circuit
    - trade fidelity for simulation rate
- Area and delay estimation
- Analysis
  - e.g., queuing theory

MOVE	39%
BR	20%
LOAD	20%
STORE	10%
ALU	11%

LD 5EA3
ST 31FF
....
LD 1EA2
....





Next Time

- Evaluation of Systems
  - Performance
    - Amdahl's Law, CPI
  - Cost
  - Benchmark Examples

- Reading assignment
  - P&H Chapter 2

- HW #1