VLSI-1 Course Review

Mark McDermott
Electrical and Computer Engineering
The University of Texas at Austin

Final Exam

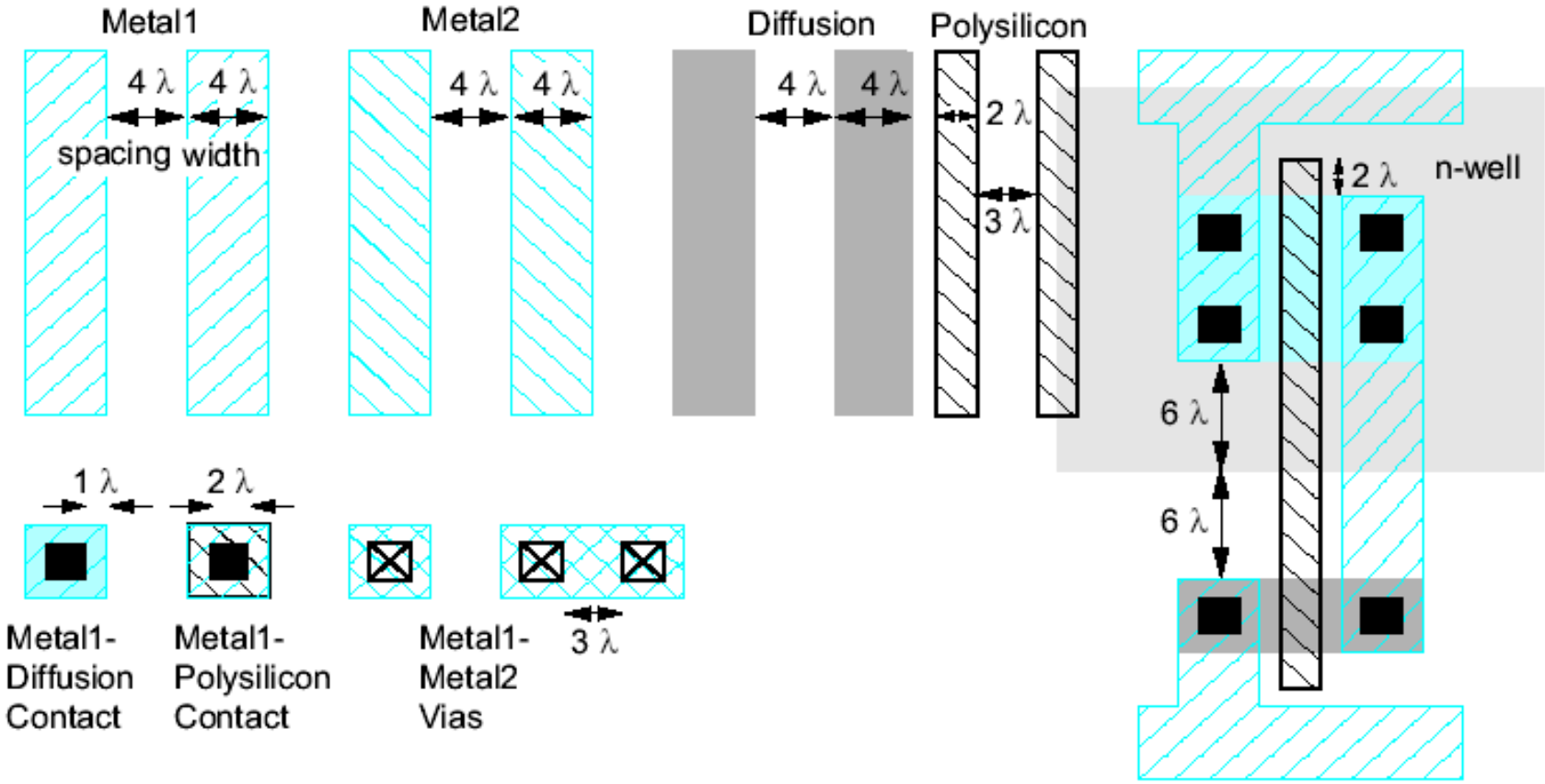
- **December 15th, RLP 0.126, 9:00AM - 11:59AM**
- **Topics may include (but are not limited to):**
  - D Flip-Flop timing analysis
  - Combinational logic timing analysis
  - Combination logic transistor sizing
  - Circuit optimization and analysis
  - Fault testing
  - State machines, state diagrams, state tables, PLAs
  - Memory design
  - General knowledge questions about transistors, wires, capacitors, power, energy, etc.

Overview

- Combinational logic
- Sequential logic
- Datapath
- Memories
- Scaling

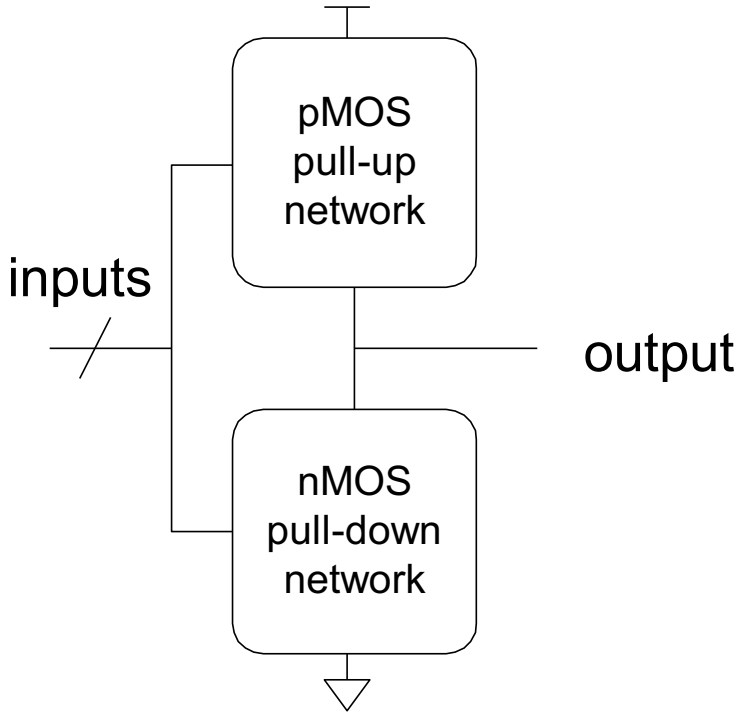
Simplified Design Rules

- **Conservative rules to get you started**



Complementary CMOS

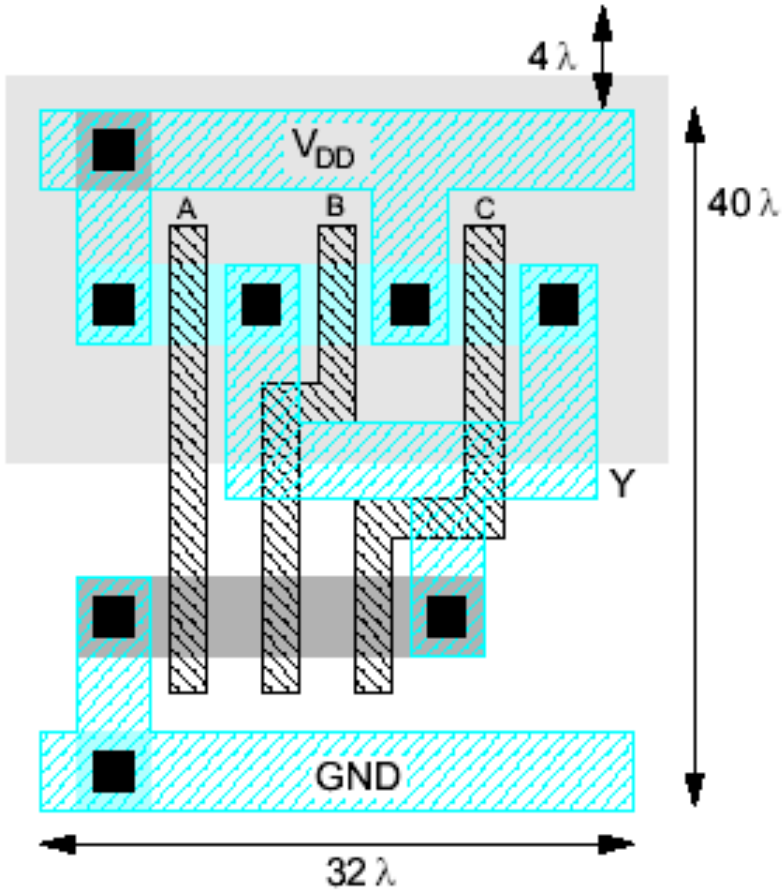
- **Complementary CMOS logic gates**
  - **nMOS *pull-down network***
  - **pMOS *pull-up network***
  - **a.k.a. static CMOS**



	Pull-up OFF	Pull-up ON
Pull-down OFF	Z (float)	1
Pull-down ON	0	X (crowbar)

Example: NAND3

- Horizontal N-diffusion and p-diffusion strips
- Vertical polysilicon gates
- Metal1 V_{DD} rail at top
- Metal1 GND rail at bottom
- 32 λ by 40 λ



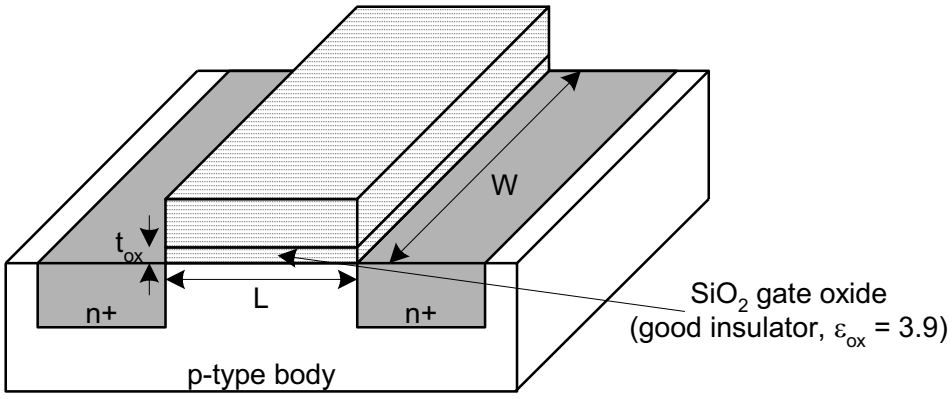
I-V Characteristics

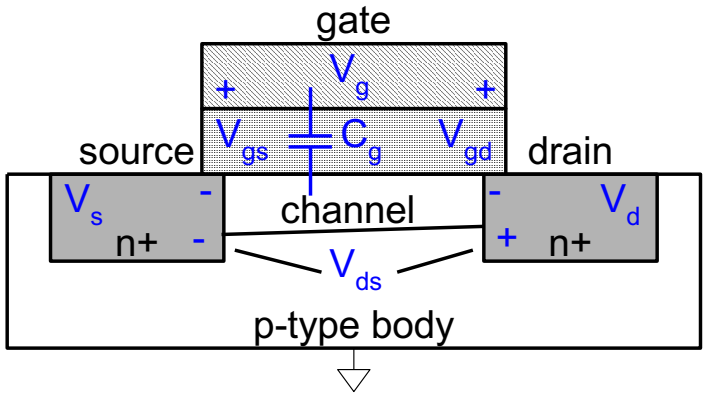
- **In Linear region, I_{ds} depends on**
  - **How much charge is in the channel?**
  - **How fast is the charge moving?**

Channel Charge

- **MOS structure looks like parallel plate capacitor while operating in inversion**
  - **Gate – oxide – channel**
- **$Q_{channel} = CV$**
- **$C = C_g = \varepsilon_{ox}WL/t_{ox} = C_{ox}WL$**
- **$V = V_{gc} - V_t = (V_{gs} - V_{ds}/2) - V_t$**

$$C_{ox} = \varepsilon_{ox} / t_{ox}$$





Carrier velocity

- Charge is carried by e-
- Carrier velocity *v* proportional to lateral E-field between source and drain
- $v = \mu E$ μ called mobility
- $E = V_{ds}/L$
- Time for carrier to cross channel:
  - $t = L / v$

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**

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

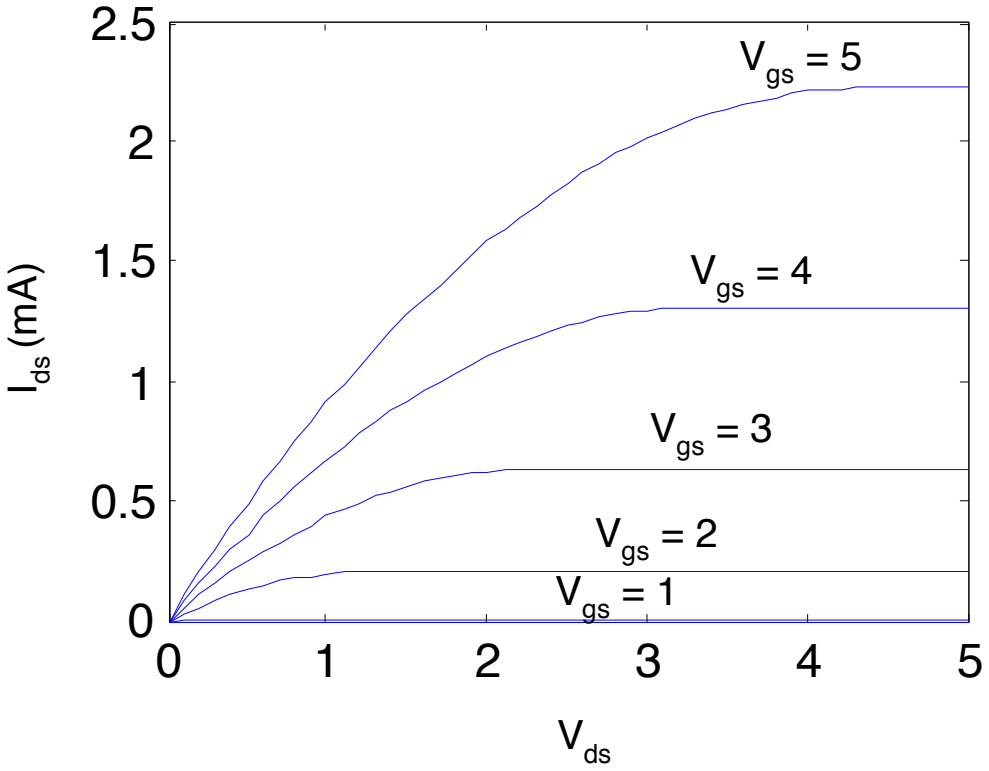
$$= \mu C_{\text{ox}} \frac{W}{L}\left(V_{gs} - V_t - V_{ds}/2\right)V_{ds}$$

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

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

Example

- **Example: a 0.6 μm process from AMI semiconductor**
  - t_{ox} = 100 Å
  - μ = 350 cm²/V*s
  - V_t = 0.7 V
- **Plot I_{ds} vs. V_{ds}**
  - V_{gs} = 0, 1, 2, 3, 4, 5
  - Use W/L = 4/2 λ

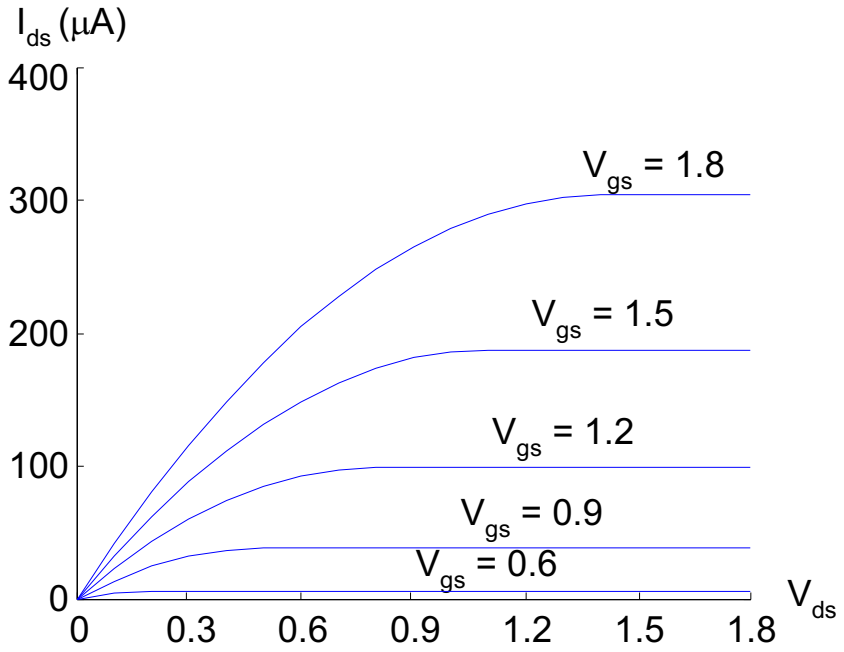


$$\beta = \mu C_{ox}\frac{W}{L} = (350)\left(\frac{3.9 \bullet 8.85 \cdot 10^{-14}}{100 \cdot 10^{-8}}\right)\left(\frac{W}{L}\right) = 120\frac{W}{L}\,\mu A/V^2$$

Ideal nMOS I-V Plot

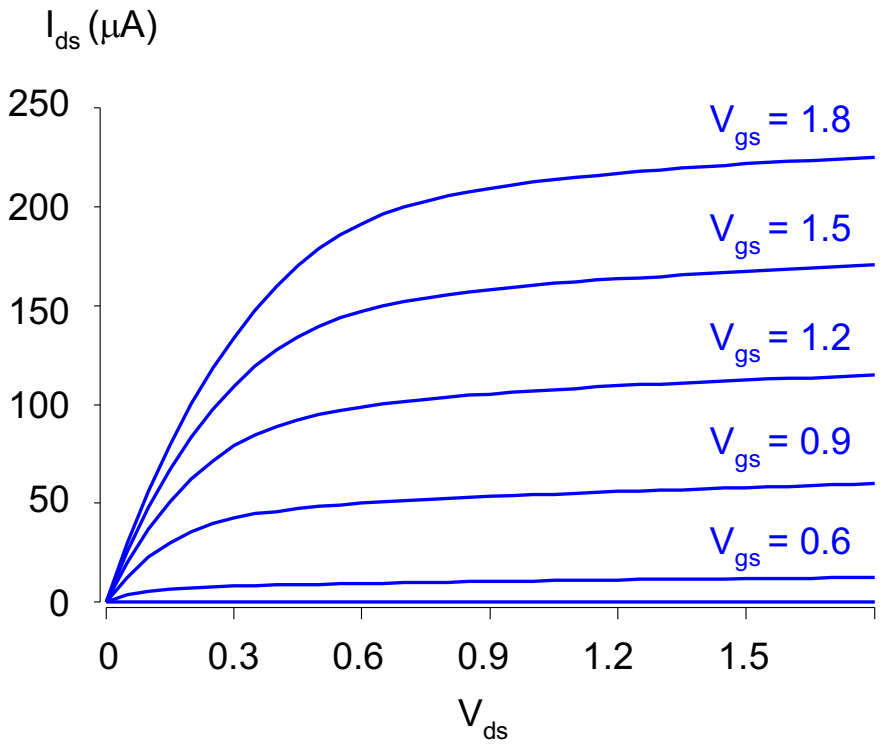
- **180 nm TSMC process**

- **Ideal Models**
  - $\beta = 155(W/L)\ \mu A/V^2$
  - $V_t = 0.4\ V$
  - $V_{DD} = 1.8\ V$



Simulated nMOS I-V Plot

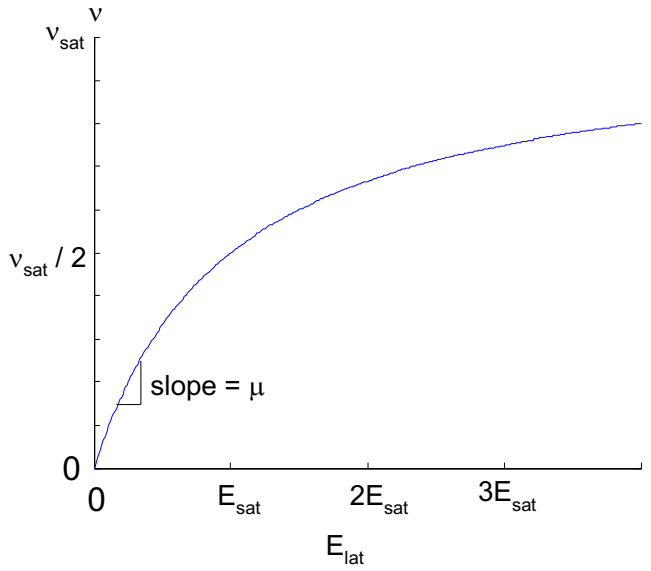
- 180 nm TSMC process
- BSIM 3v3 SPICE models
- What differs?
  - Less ON current
  - No square law
  - Current increases in saturation



Velocity Saturation

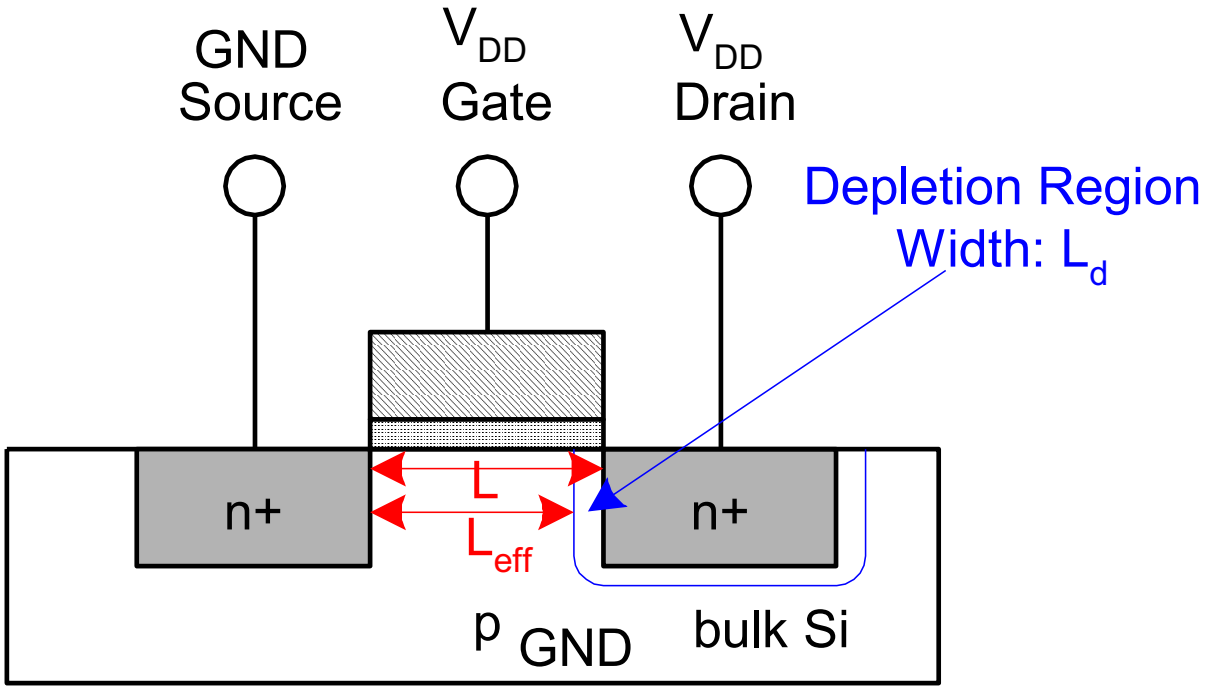
- **We assumed carrier velocity is proportional to E-field**
  - $v = \mu E_{lat} = \mu V_{ds}/L$
- **At high fields, this ceases to be true**
  - Carriers scatter off atoms
  - Velocity reaches v_{sat}
    - Electrons: 6-10 x 10^6 cm/s
    - Holes: 4-8 x 10^6 cm/s
  - Better model

$$v = \frac{\mu E_{lat}}{1 + \frac{E_{lat}}{E_{sat}}} \Rightarrow v_{sat} = \mu E_{sat}$$



Channel Length Modulation

- **Reverse-biased p-n junctions form a *depletion region***
  - Region between n and p with no carriers
  - Width of depletion L_d region grows with reverse bias
  - $L_{eff} = L - L_d$
- **Shorter L_{eff} gives more current**
  - I_{ds} increases with V_{ds}
  - Even in saturation

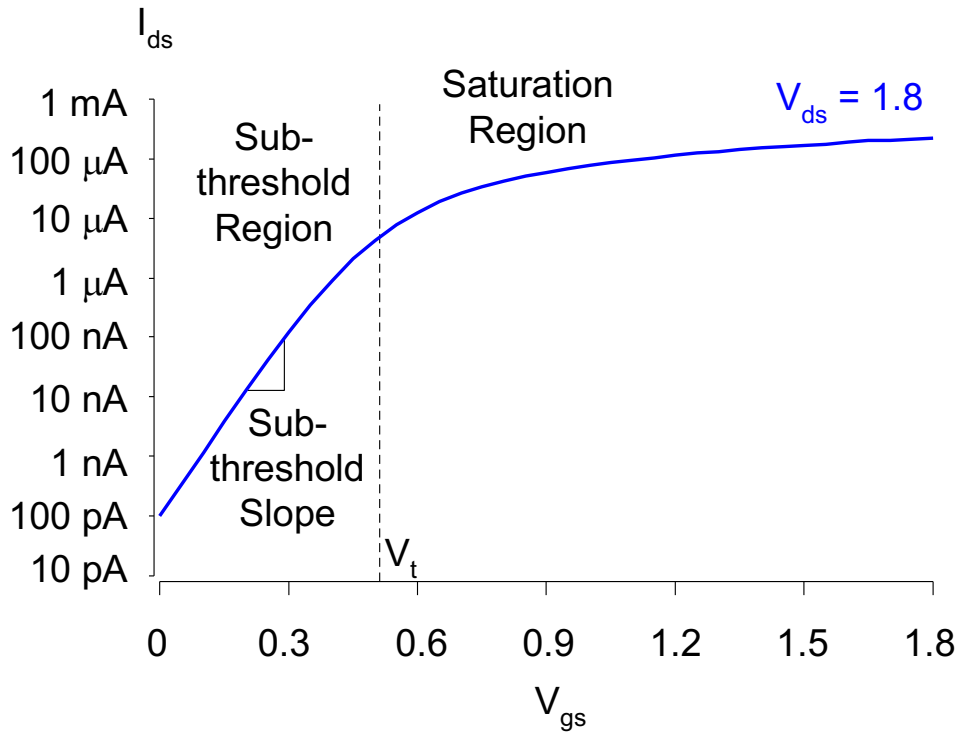


Body Effect

- V_t: gate voltage necessary to invert channel
- Increases if source voltage increases because source is connected to the channel
- Increase in V_t with V_s is called the *body effect*

OFF Transistor Behavior

- What about current in cutoff?
- Simulated results
- What differs?
  - Current doesn't go to 0 in cutoff



Leakage Sources

- **Subthreshold conduction**
  - Transistors can't abruptly turn ON or OFF
- **Junction leakage**
  - Reverse-biased PN junction diode current
- **Gate leakage**
  - Tunneling through ultra-thin gate dielectric

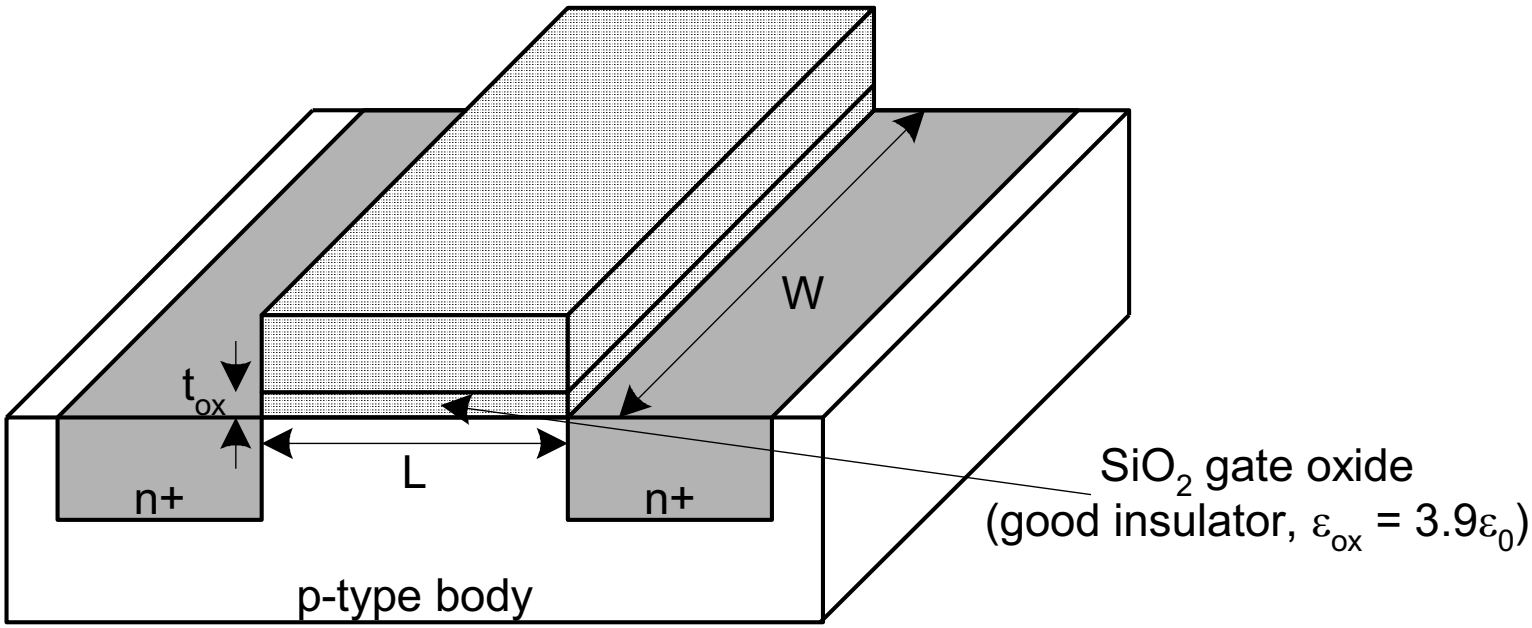
- **Subthreshold leakage is the biggest source in modern transistors**

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

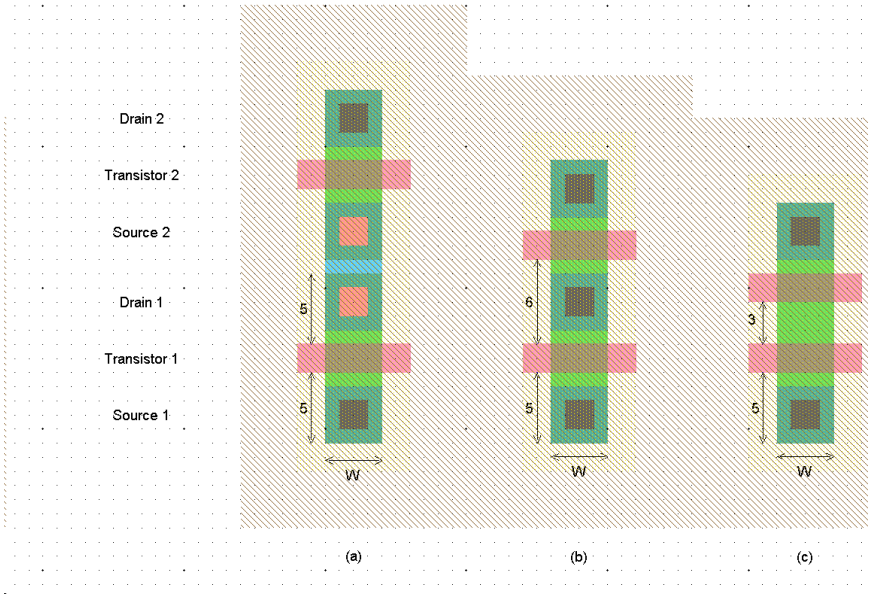
Gate Capacitance

- Approximate channel as connected to source
- $C_{gs} = \varepsilon_{ox}WL/t_{ox} = C_{ox}WL = C_{permicron}W$
- $C_{permicron}$ is typically about 2 fF/μm



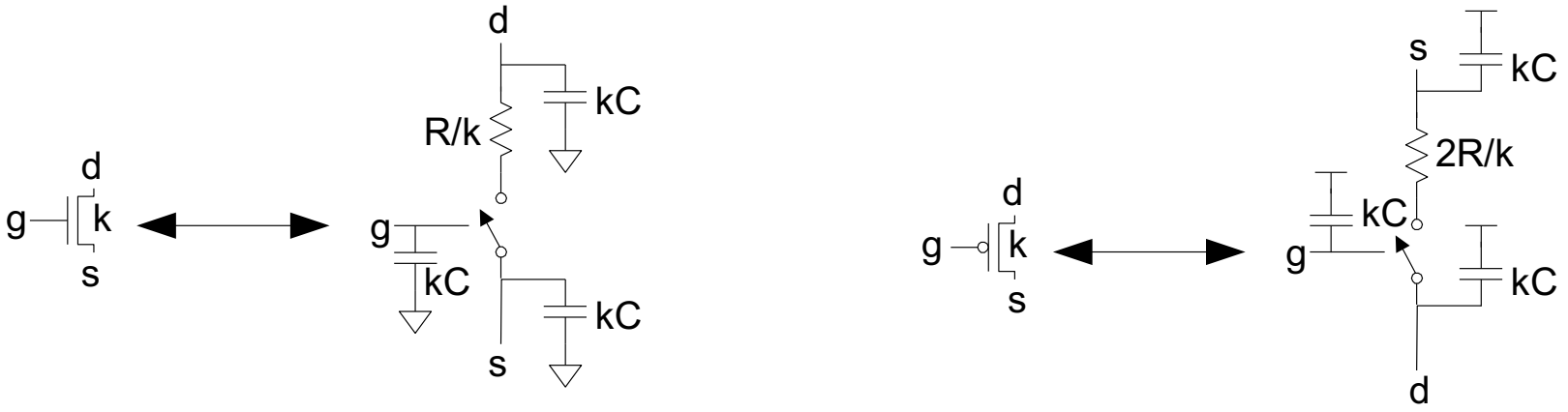
Diffusion Capacitance

- 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
  - ½ C_g for uncontacted
  - Varies with process



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

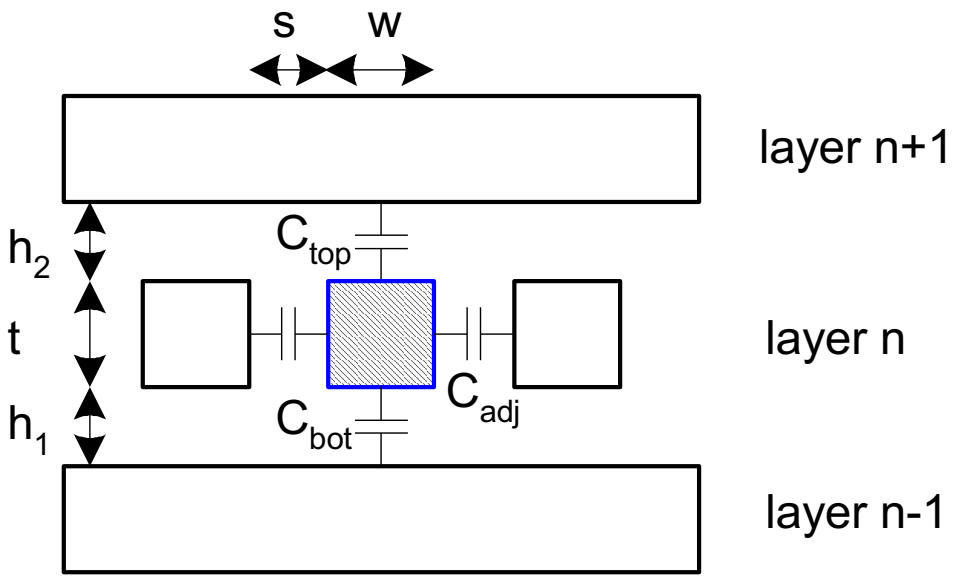


Interconnects

- **Chips are mostly made of wires called *interconnect***
  - **In stick diagram, wires set size**
  - **Transistors are little things under the wires**
  - **Many layers of wires**
- **Wires are as important as transistors**
  - **Speed**
  - **Power**
  - **Noise**
- **Alternating layers run orthogonally**

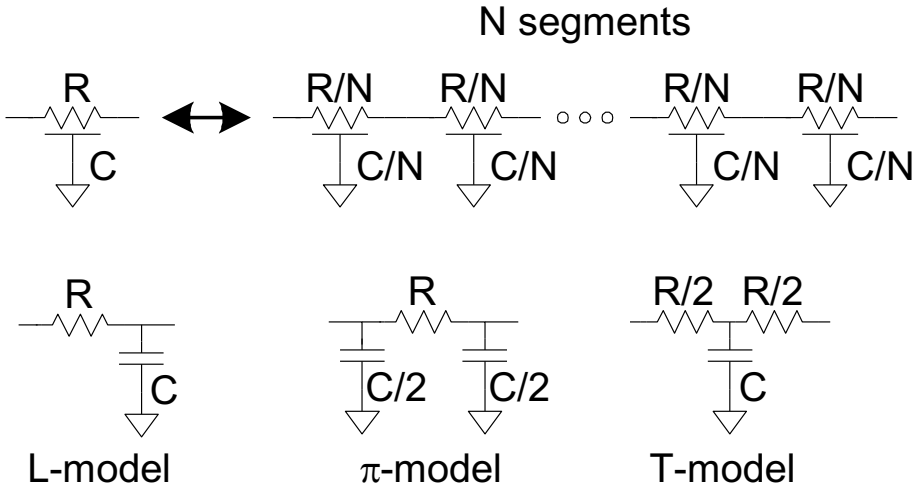
Wire Capacitance

- **Wire has capacitance per unit length**
  - To neighbors
  - To layers above and below
- **$C_{total} = C_{top} + C_{bot} + 2C_{adj}$**



Lumped Element Models

- **Wires are a distributed system**
  - **Approximate with lumped element models**



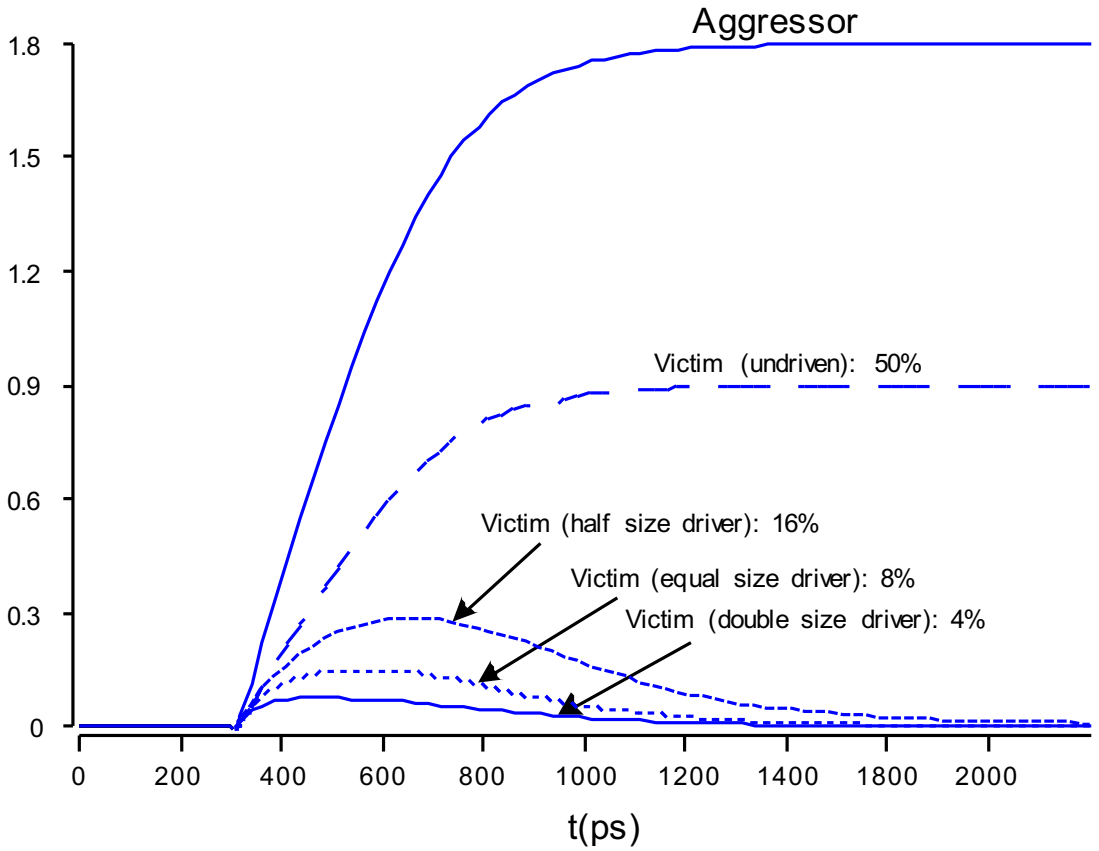
- **3-segment π-model is accurate to 3% in simulation**
- **L-model needs 100 segments for same accuracy!**
- **Use single segment π-model for Elmore delay**

Crosstalk

- A capacitor does not like to change its voltage instantaneously.
- A wire has high capacitance to its neighbor.
  - When the neighbor switches from 1-> 0 or 0->1, the wire tends to switch too.
  - Called capacitive *coupling* or *crosstalk*.
- Crosstalk effects
  - Noise on nonswitching wires
  - Increased delay on switching wires

Coupling Waveforms

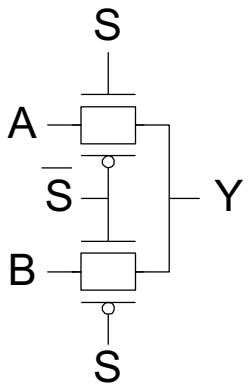
- **Simulated coupling for $C_{adj} = C_{victim}$**

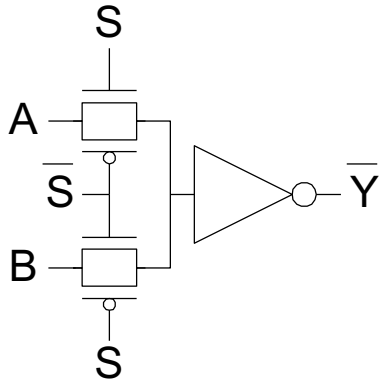


Pass Transistor Circuits

- Use pass transistors like switches to do logic
- Inputs drive diffusion terminals as well as gates

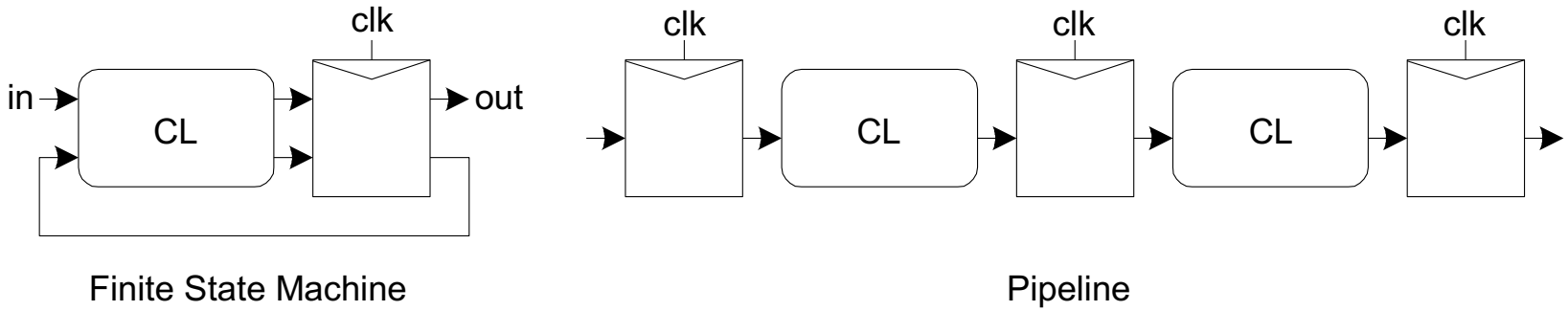
- CMOS + Transmission Gates:
  - 2-input multiplexer
  - Gates should be restoring





Sequencing

- ***Combinational logic***
  - output depends on current inputs
- ***Sequential logic***
  - output depends on current and previous inputs
  - Requires separating previous, current, future
  - Called *state* or *tokens*
  - Ex: FSM, pipeline



Finite State Machine

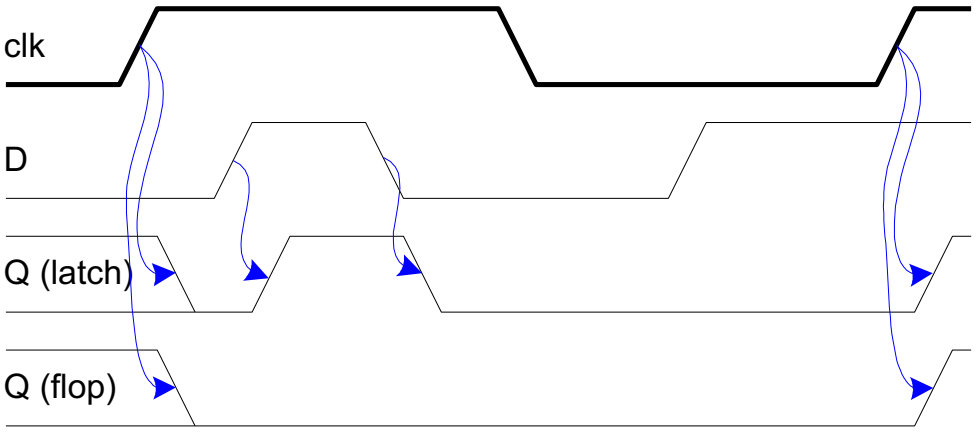
Pipeline

Sequencing Overhead

- Use flip-flops to delay fast tokens so they move through exactly one stage each cycle.
- Inevitably adds some delay to the slow tokens
- Makes circuit slower than just the logic delay
  - Called sequencing overhead
- Some people call this clocking overhead
  - But it applies to asynchronous circuits too
  - Inevitable side effect of maintaining sequence

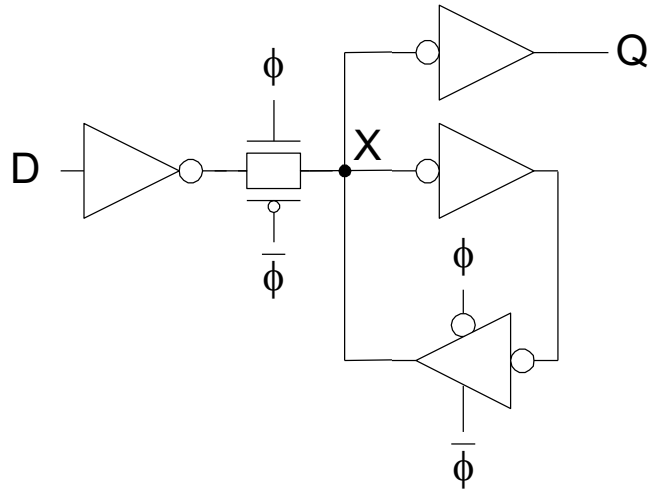
Sequencing Elements

- **Latch: Level sensitive**
  - a.k.a. transparent latch, D latch
- **Flip-flop: edge triggered**
  - A.k.a. master-slave flip-flop, D flip-flop, D register
- **Timing Diagrams**
  - Transparent
  - Opaque
  - Edge-trigger



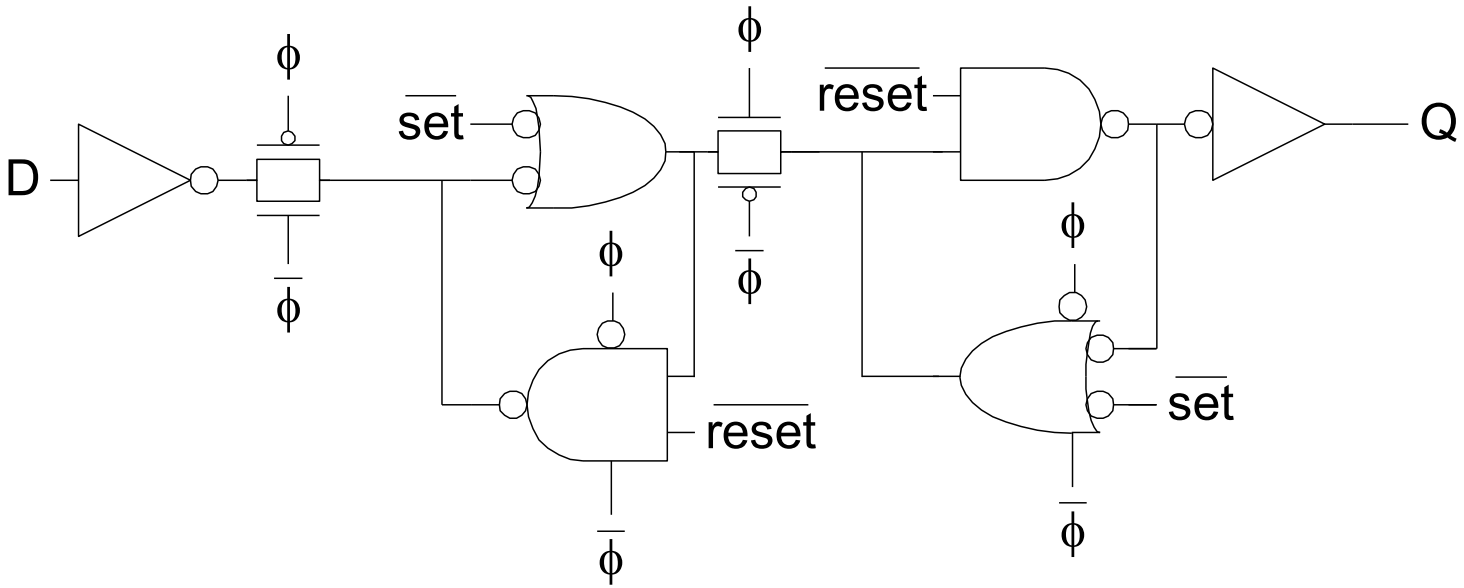
Latch Design

- **Buffered output**
  - + No backdriving
- **Widely used in standard cells**
  - + Very robust (most important)
  - - Rather large
  - - Rather slow (1.5 – 2 FO4 delays)
  - - High clock loading



Set / Reset

- Set forces output high when enabled
- Flip-flop with asynchronous set and reset

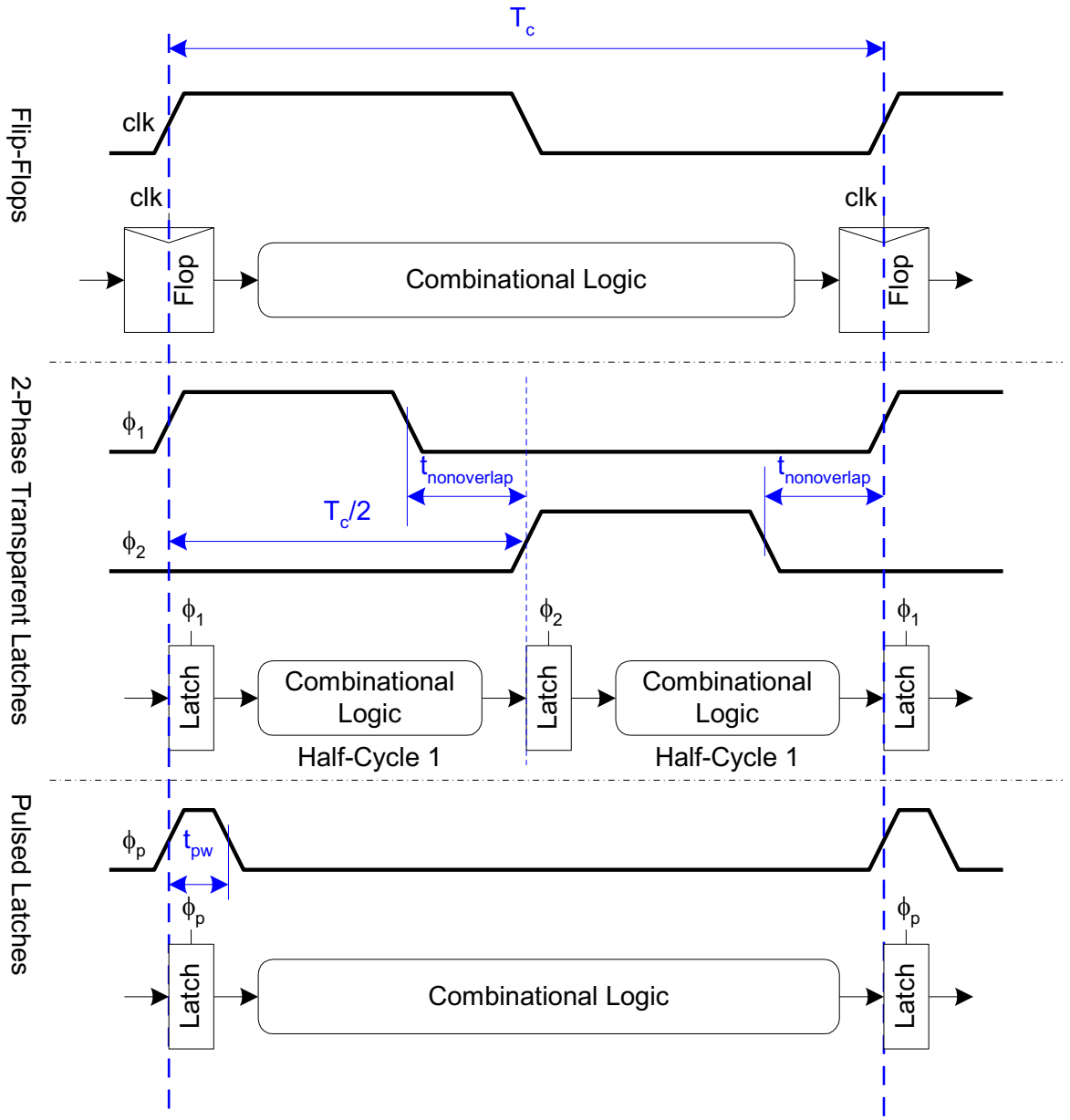


Sequencing Methods

Flip-flops

2-Phase Latches

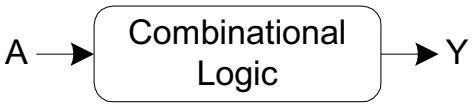
Pulsed Latches

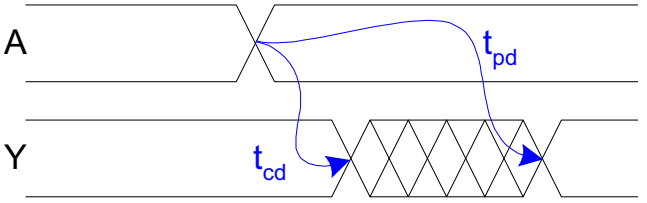


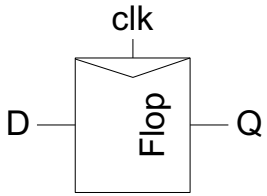
Timing Diagrams

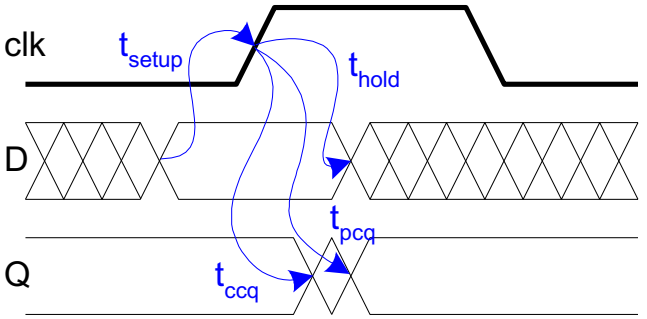
Contamination and Propagation Delays

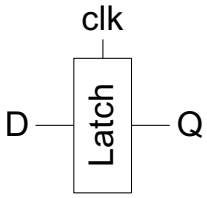
t_{pd}	Logic Propagation Delay
t_{cd}	Logic Contamination Delay
t_{pcq}	Latch/Flop Clk-Q Prop Delay
t_{ccq}	Latch/Flop Clk-Q Cont. Delay
t_{pdq}	Latch D-Q Prop Delay
t_{cdq}	Latch D-Q Cont. Delay
t_{setup}	Latch/Flop Setup Time
t_{hold}	Latch/Flop Hold Time

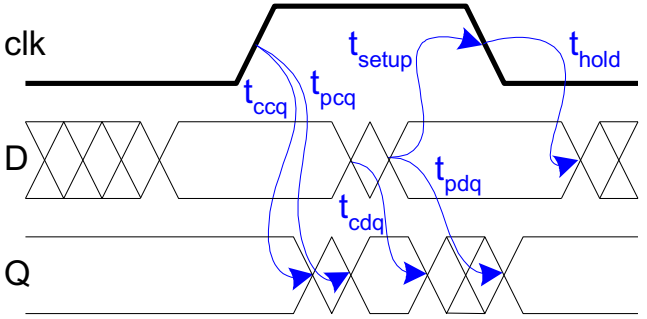




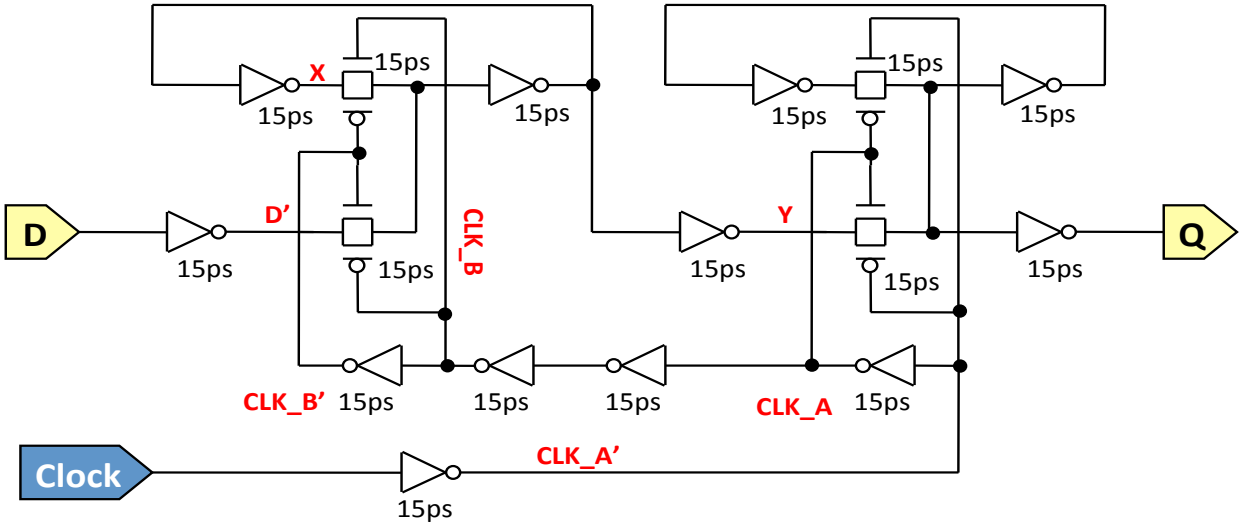








Master-Slave Flip-Flop



What is the SETUP time?

(15ps + 15ps + 15ps + 15ps) – (15ps + 15ps) = 30 ps
Input path from D to NMOS Clock Inverters to Slave stage

NOTE: The two inverter delay between the clocks to Master and Slave stages requires that the data is set up to the slave stage (node Y) earlier than to node X in the master stage so that there is not a Clock->Q push-out.

What is the CLOCK – Q Delay

15ps + 15ps + 15ps + 15ps = 60 ps
inv inv pass inv

If you use 0ps for the setup time, the Clock-Q delay will be 90ps.

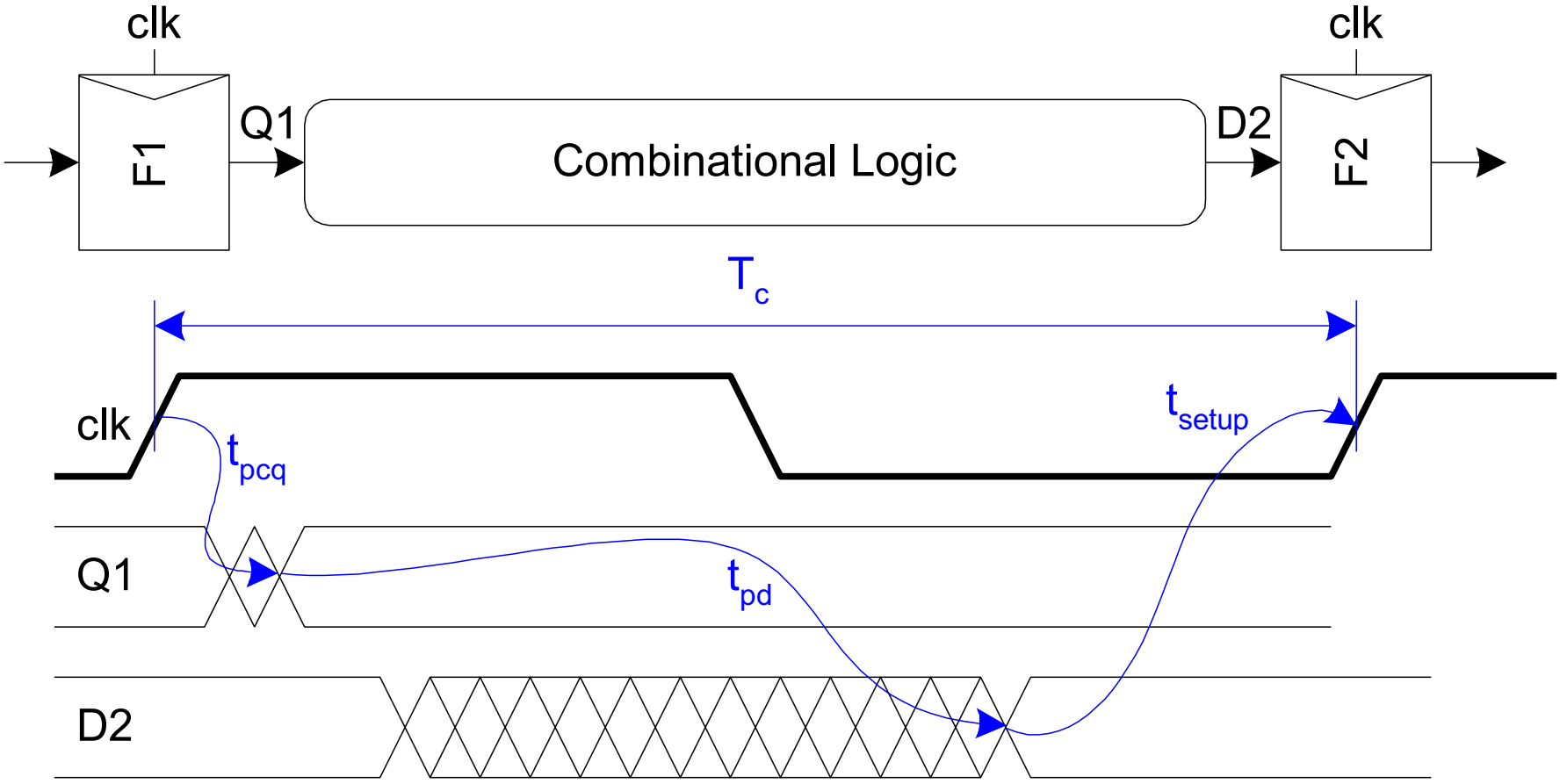
What is the HOLD time?

(15ps + 15ps + 15ps + 15ps +15ps) – (15ps) = 60 ps
Clock inverters inv
on D input

Max-Delay: Flip-Flops

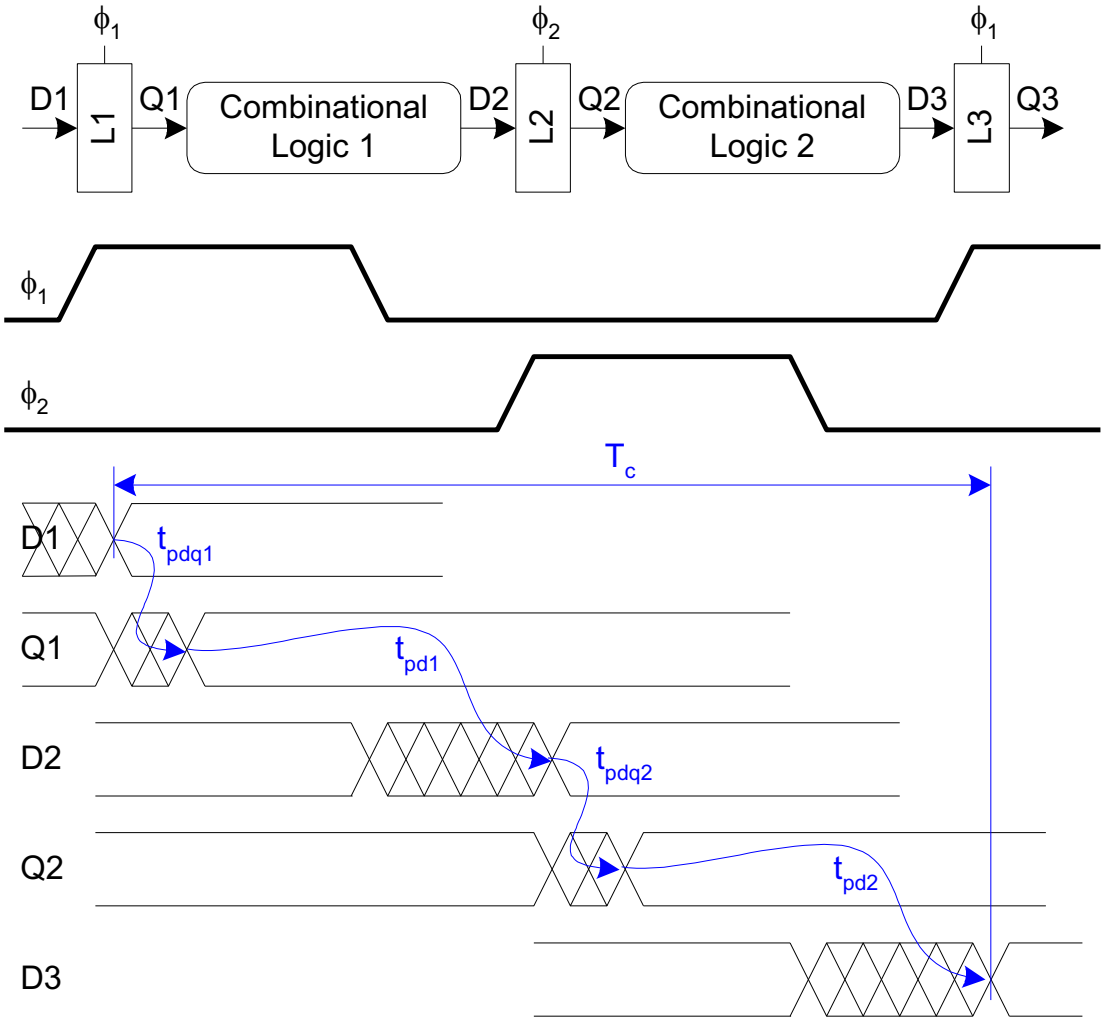
$$T_c \geq t_{pd+} (t_{setup} + t_{pcq})$$

sequencing overhead



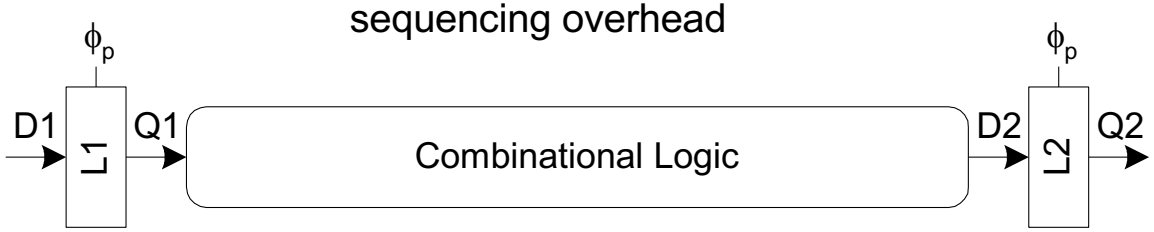
Max Delay: 2-Phase Latches

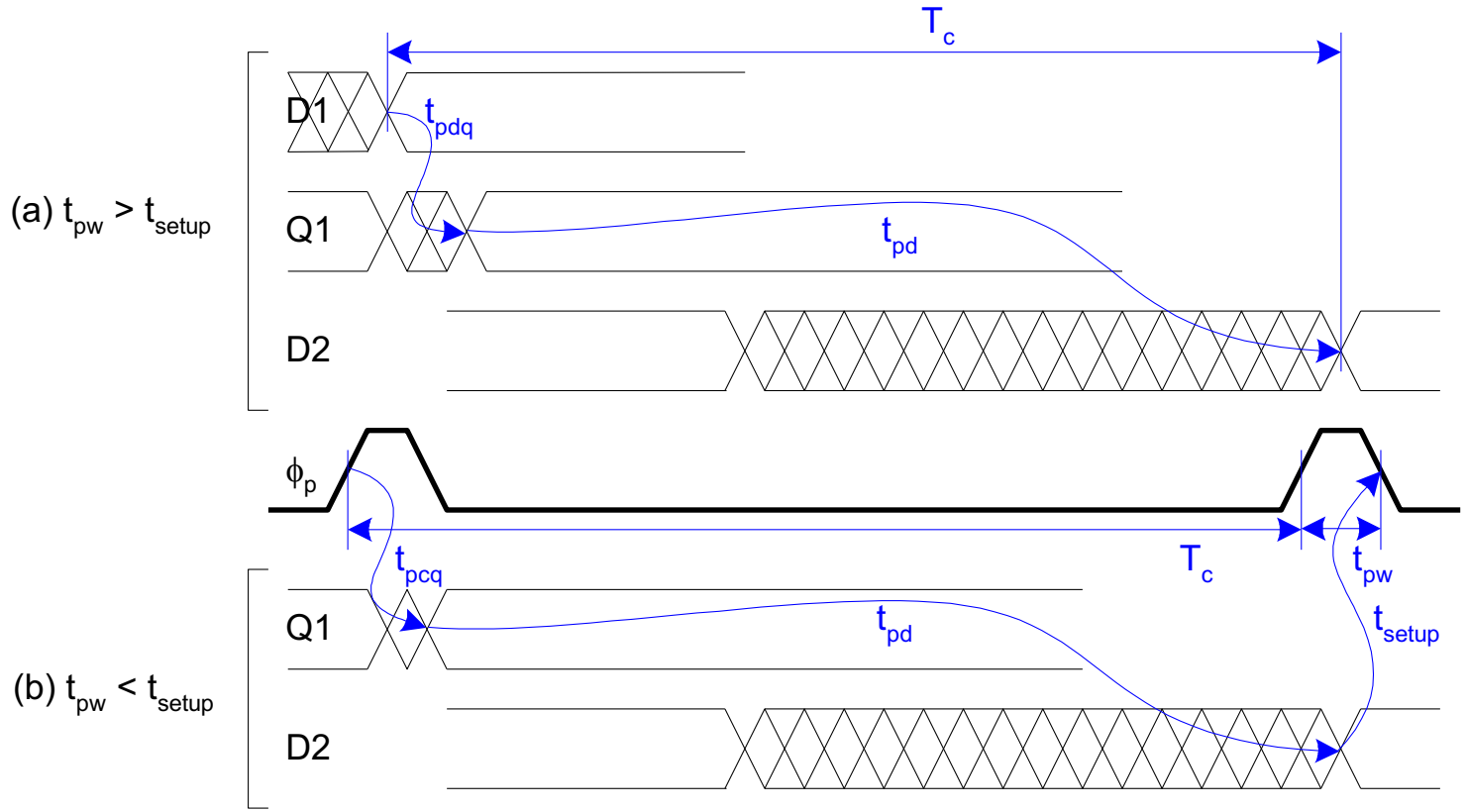
$$T_c \geq t_{pd1} + t_{pd2} + \underbrace{t_{pdq1} + t_{pdq2}}_{\text{sequencing overhead}}$$



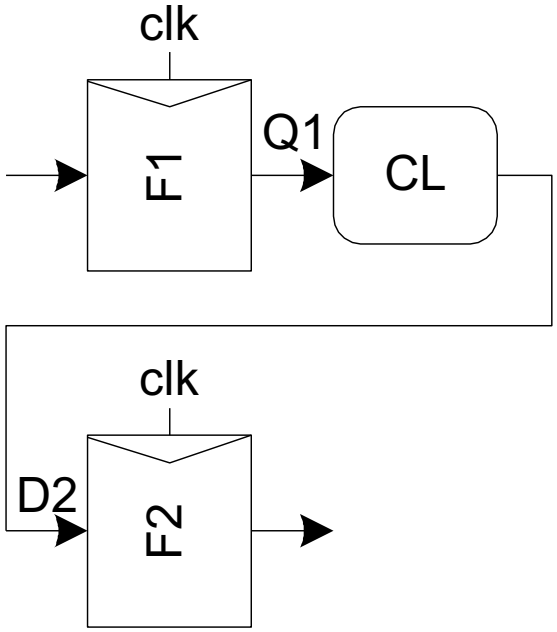
Max Delay: Pulsed Latches

$$t_{pd} + \underbrace{\max(t_{pdq},\ t_{pcq} + t_{setup} - t_{pw})}_{\text{sequencing overhead}} \le T_c$$

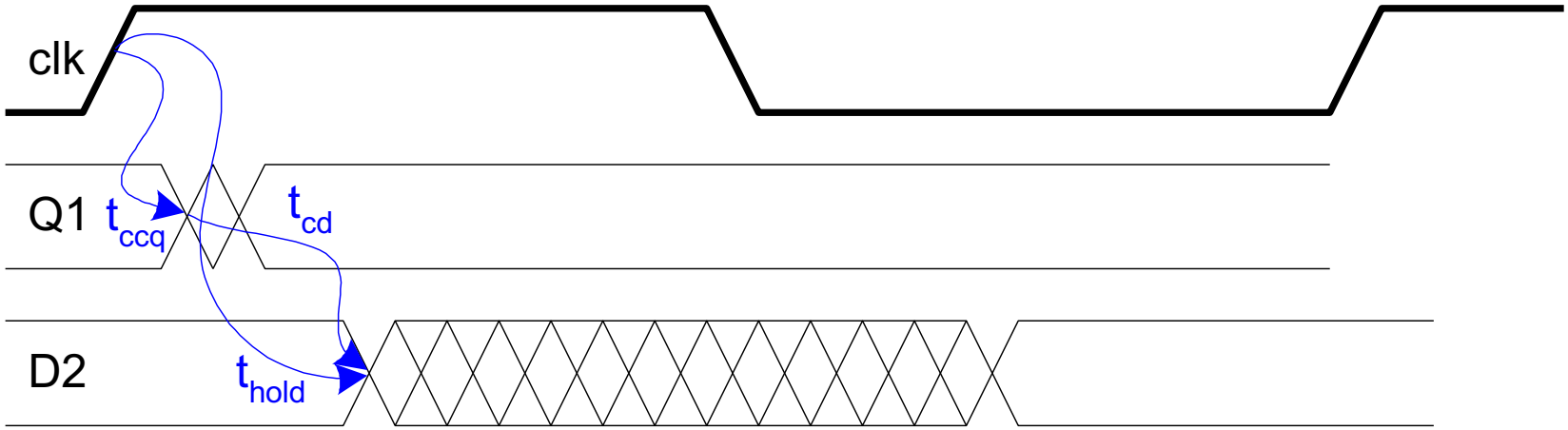




Min-Delay: Flip-Flops

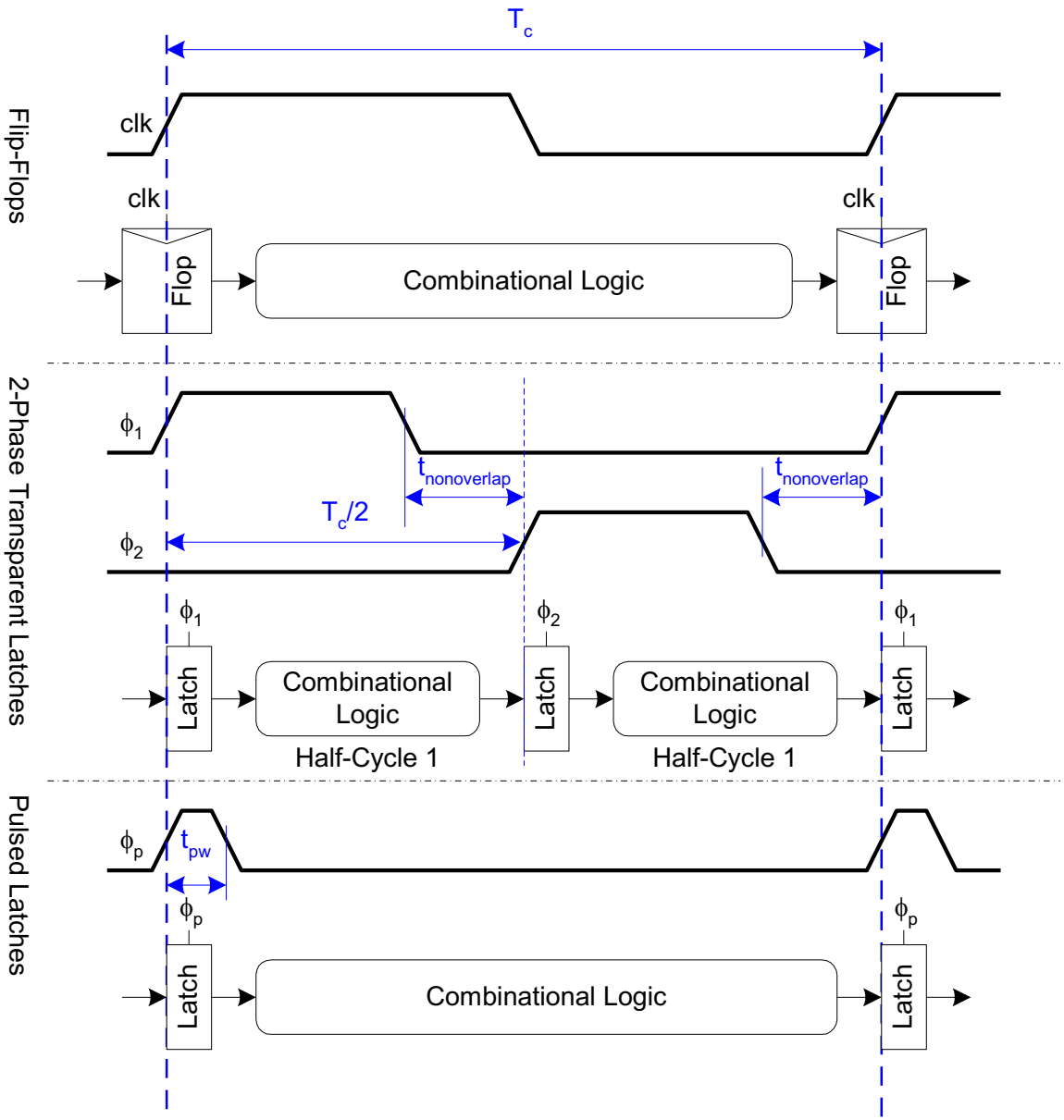


$$t_{cd} \geq t_{\text{hold}} - t_{ccq}$$



Sequencing Methods

- Flip-flops
- 2-Phase Latches
- Pulsed Latches



Flip-Flop Summary

- **Flip-Flops:**
  - Very easy to use, supported by all tools
- **2-Phase Transparent Latches:**
  - Lots of skew tolerance and time borrowing
- **Pulsed Latches:**
  - Fast, some skew tol & borrow, hold time risk

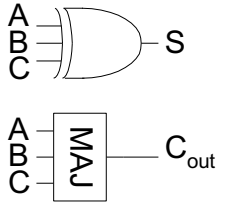
	Sequencing overhead ($T_c - t_{pd}$)	Minimum logic delay t_{cd}	Time borrowing t_{borrow}
Flip-Flops	$t_{pcq} + t_{setup} + t_{skew}$	$t_{hold} - t_{ccq} + t_{skew}$	0
Two-Phase Transparent Latches	$2t_{pdq}$	$t_{hold} - t_{ccq} - t_{nonoverlap} + t_{skew}$ in each half-cycle	$\frac{T_c}{2} - \left(t_{setup} + t_{nonoverlap} + t_{skew}\right)$
Pulsed Latches	$\max\left(t_{pdq}, t_{pcq} + t_{setup} - t_{pw} + t_{skew}\right)$	$t_{hold} - t_{ccq} + t_{pw} + t_{skew}$	$t_{pw} - \left(t_{setup} + t_{skew}\right)$

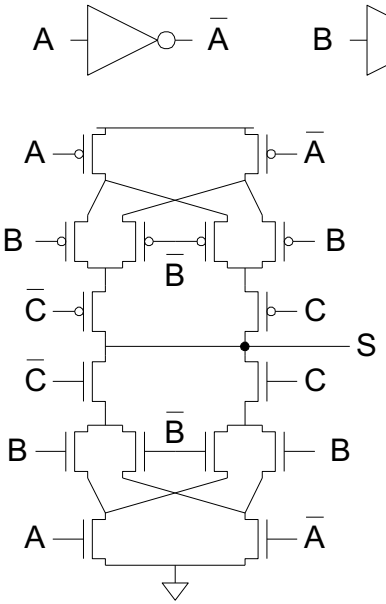
Full Adder Design I

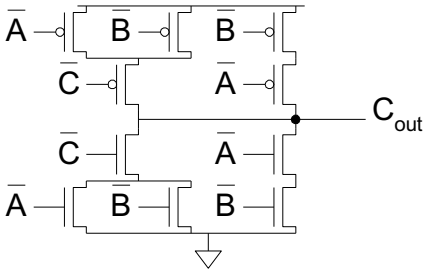
- **Brute force implementation from eqns**

$$S = A \oplus B \oplus C$$

$$C_{out} = MAJ(A, B, C)$$

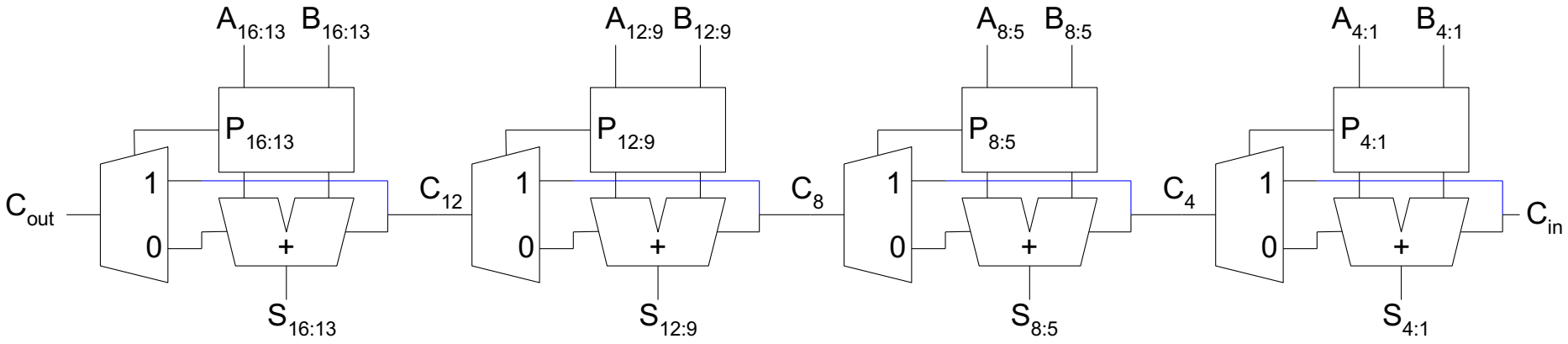






Carry-Skip Adder

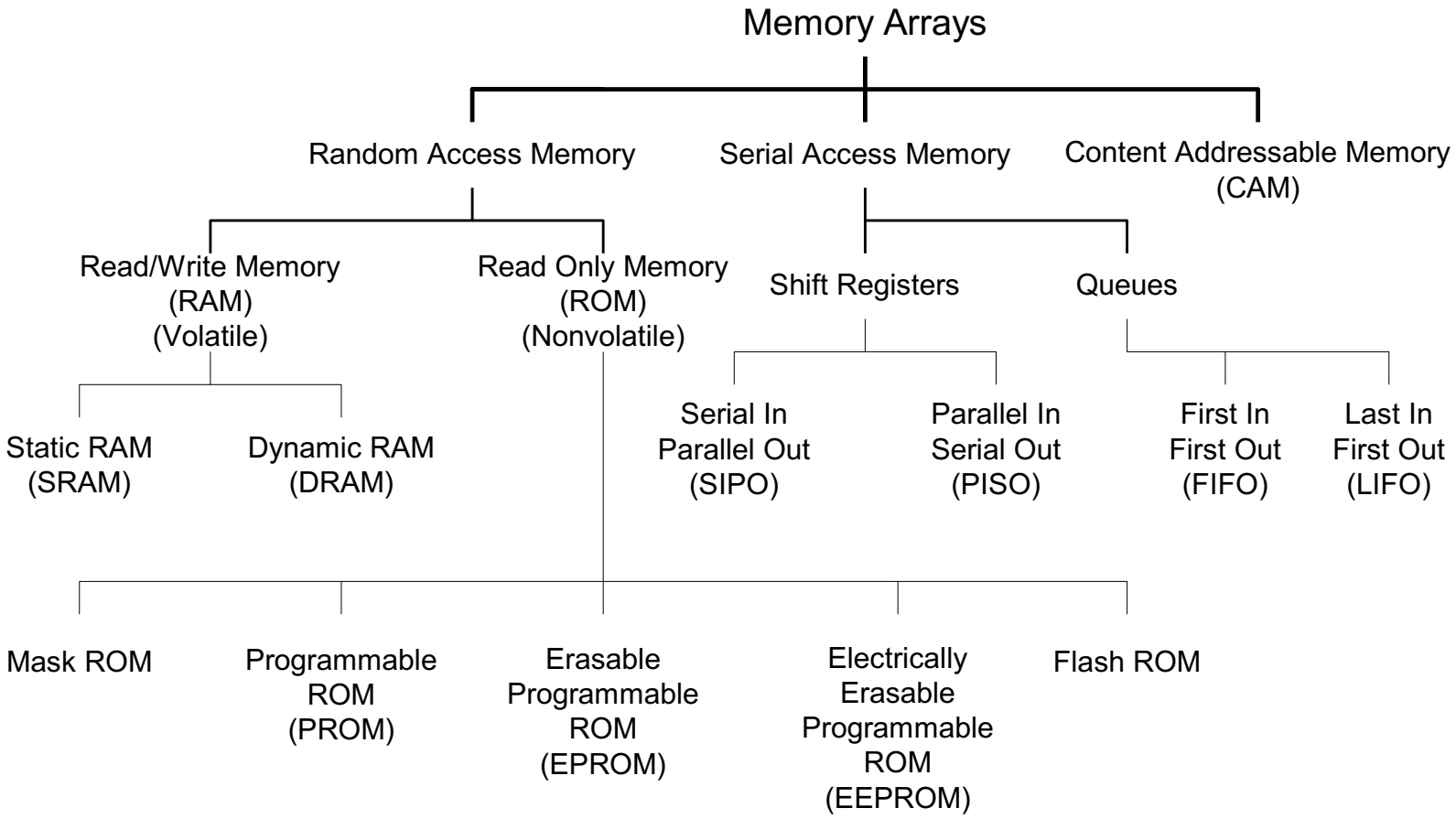
- Carry-ripple is slow through all N stages
- Carry-skip allows carry to skip over groups of n bits
  - Decision based on n-bit propagate signal



Tree Adder

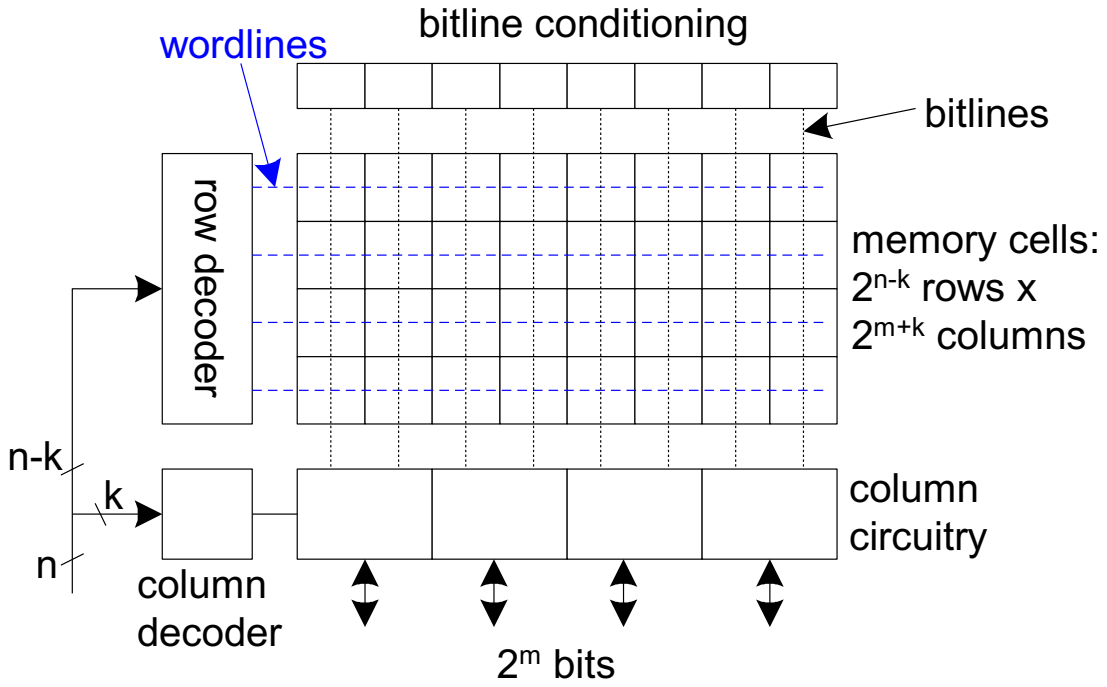
- If lookahead is good, lookahead across lookahead!
  - Recursive lookahead gives O(log N) delay
- Many variations on tree adders

Memory Arrays



Array Architecture

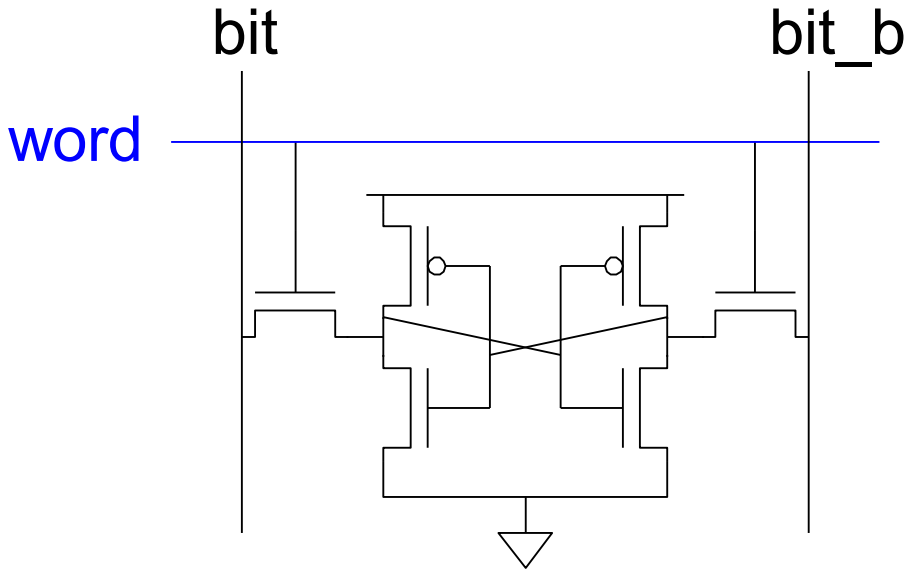
- **2^n *words* of 2^m *bits* each**
- **If n >> m, fold by 2^k into fewer *rows* of more *columns***



- **Good regularity – easy to design**
- **Very high density if good cells are used**

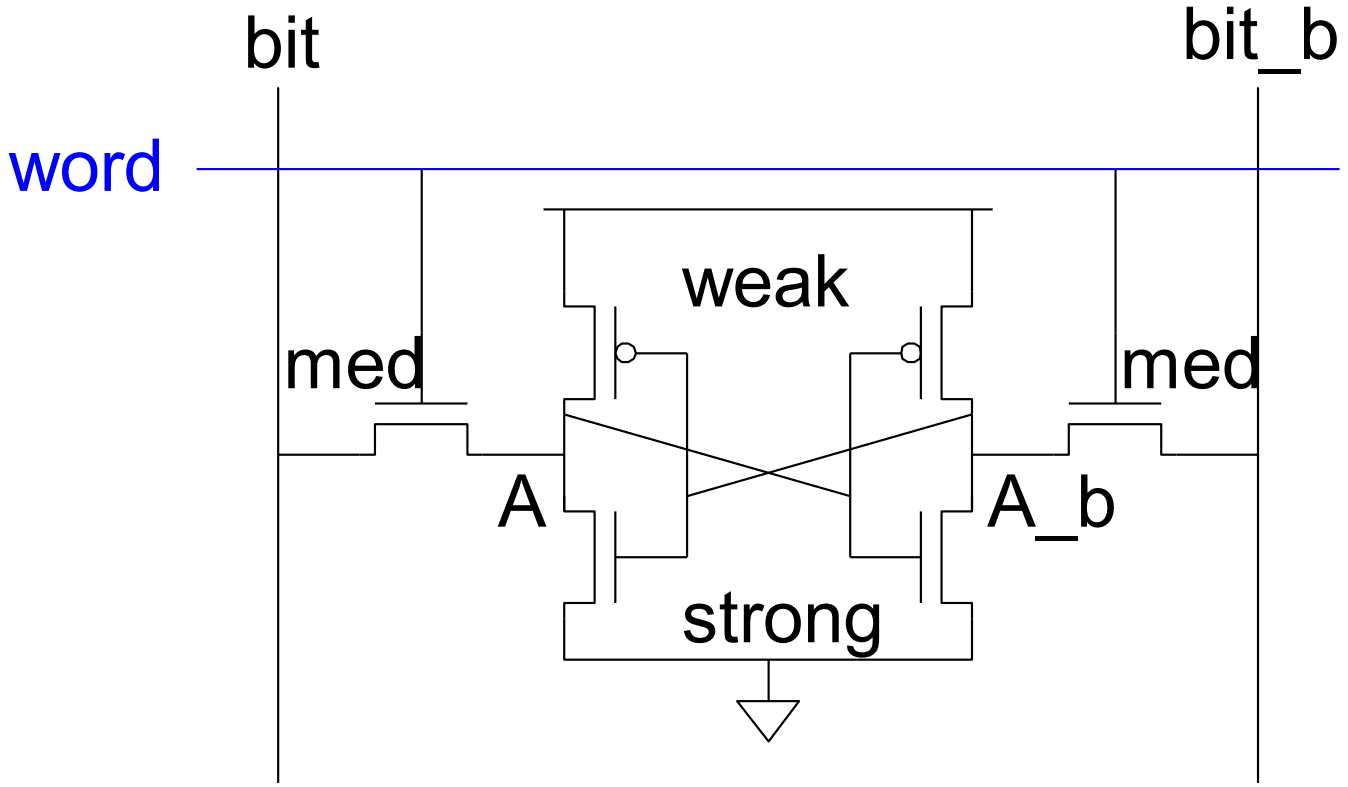
6T SRAM Cell

- **Cell size accounts for most of array size**
  - Reduce cell size at expense of complexity
- **6T SRAM Cell**
  - Used in most commercial chips
  - Data stored in cross-coupled inverters
- **Read:**
  - Precharge bit, bit_b
  - Raise wordline
- **Write:**
  - Drive data onto bit, bit_b
  - Raise wordline



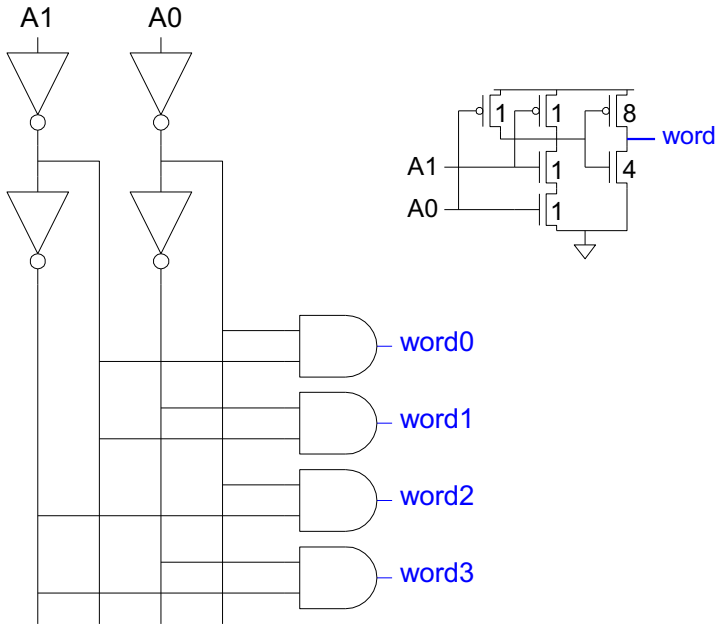
SRAM Sizing

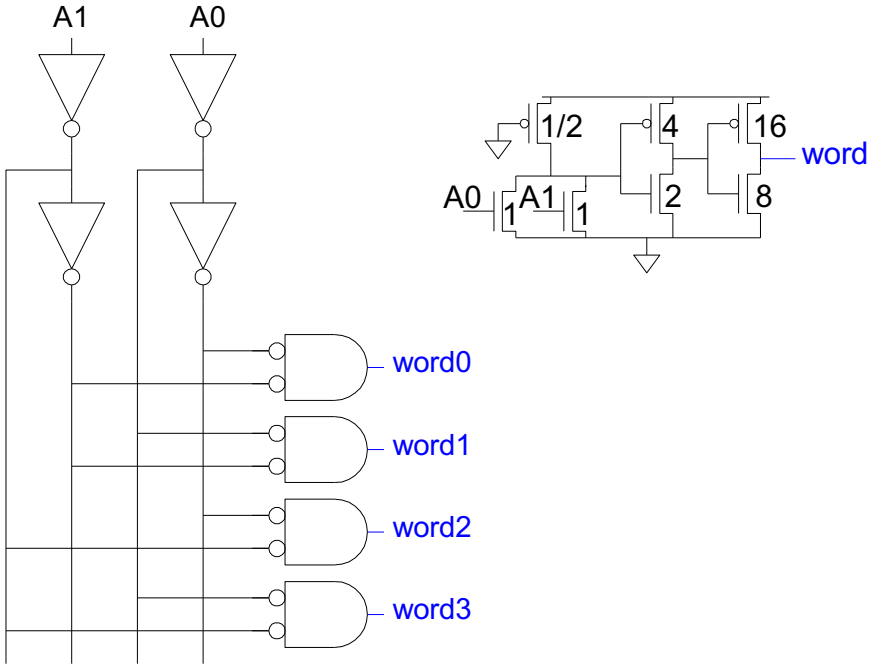
- High bitlines must not overpower inverters during reads
- But low bitlines must write new value into cell



Decoders

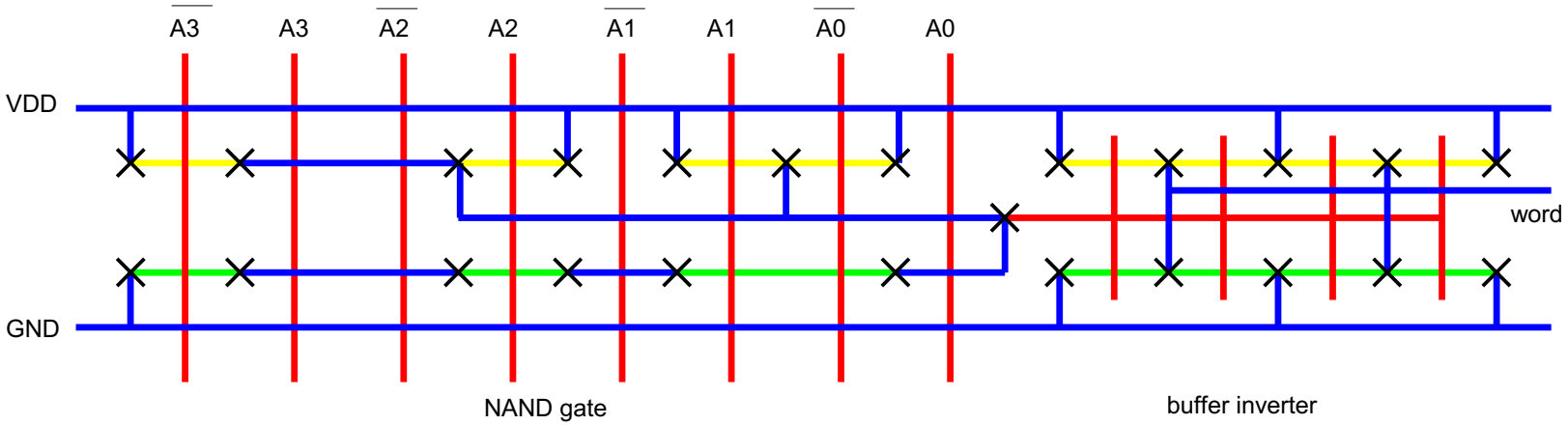
- **n:2^n decoder consists of 2^n n-input AND gates**
  - One needed for each row of memory
  - Build AND from NAND or NOR gates
- **Static CMOS** **Pseudo-nMOS**





Decoder Layout

- **Decoders must be pitch-matched to SRAM cell**
  - **Requires very skinny gates**



Sense Amplifiers

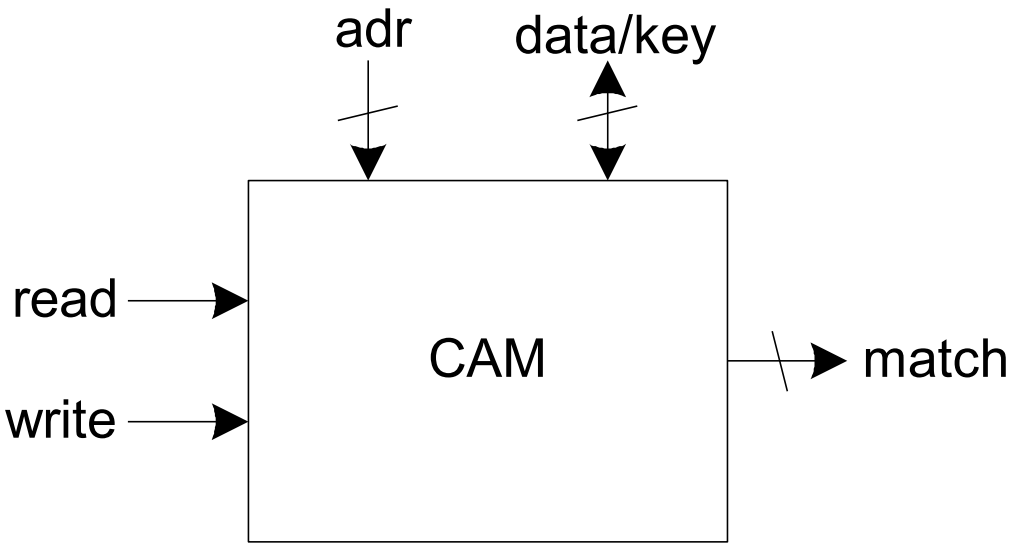
- **Bitlines have many cells attached**
  - **Ex: 32-kbit SRAM has 256 rows x 128 cols**
  - **128 cells on each bitline**

- **$t_{pd} \propto (C/I)\,\Delta V$**
  - **Even with shared diffusion contacts, 64C of diffusion capacitance (big C)**
  - **Discharged slowly through small transistors (small I)**

- ***Sense amplifiers* are triggered on small voltage swing (reduce ΔV)**

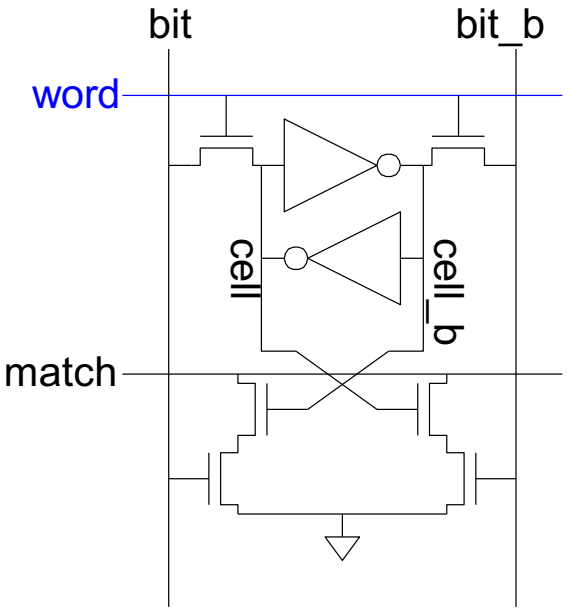
CAMs

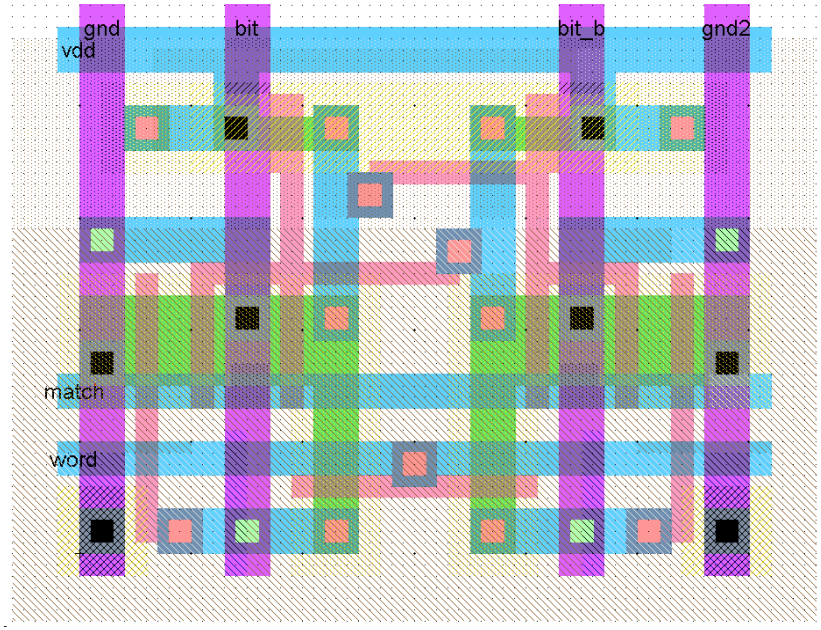
- **Extension of ordinary memory (e.g. SRAM)**
  - **Read and write memory as usual**
  - **Also *match* to see which words contain a *key***



10T CAM Cell

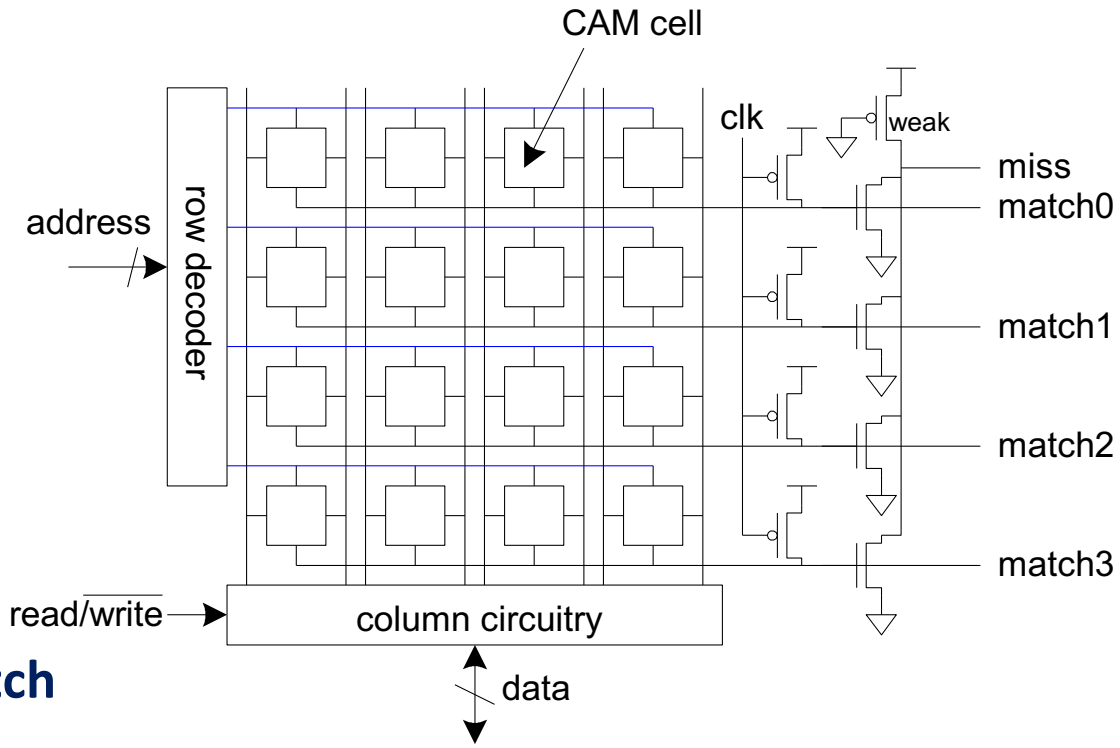
- **Add four match transistors to 6T SRAM**
  - **56 x 43 λ unit cell**





CAM Cell Operation

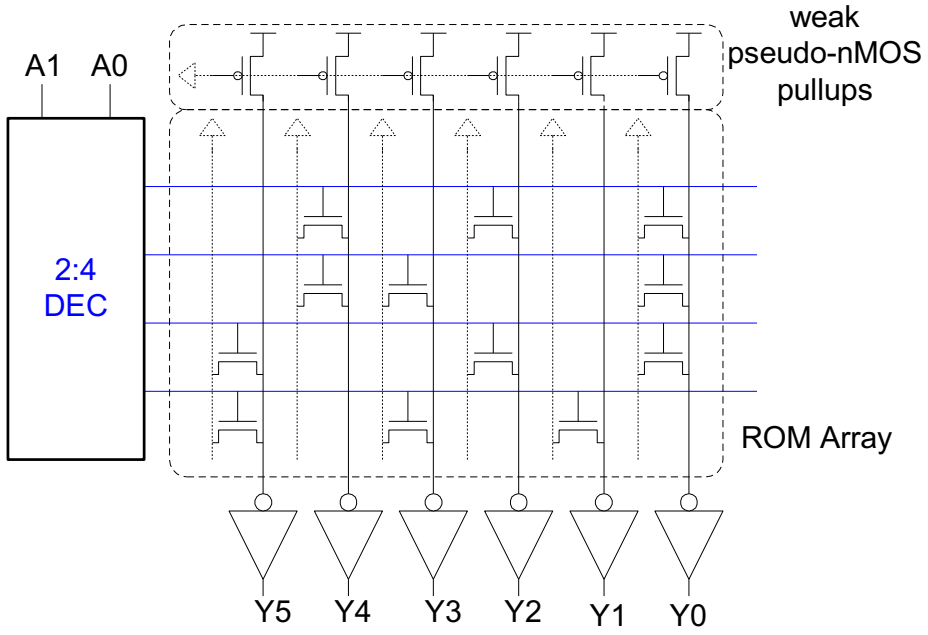
- **Read and write like ordinary SRAM**
- **For matching:**
  - **Leave wordline low**
  - **Precharge matchlines**
  - **Place key on bitlines**
  - **Matchlines evaluate**
- **Miss line**
  - **Pseudo-nMOS NOR of match lines**
  - **Goes high if no words match**

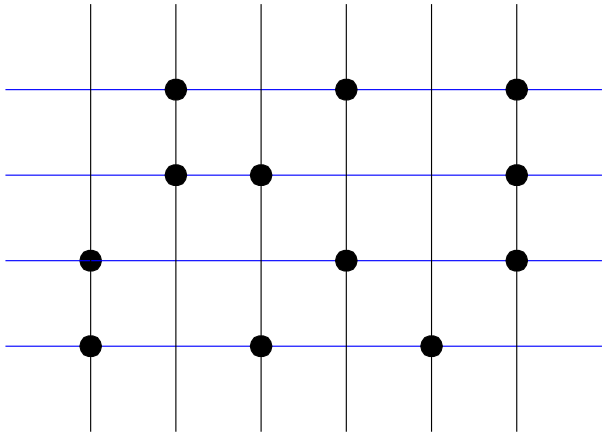


ROM Example

- **4-word x 6-bit ROM**
  - **Represented with dot diagram**
  - **Dots indicate 1's in ROM**

```
Word 0: 010101
Word 1: 011001
Word 2: 100101
Word 3: 101010
```



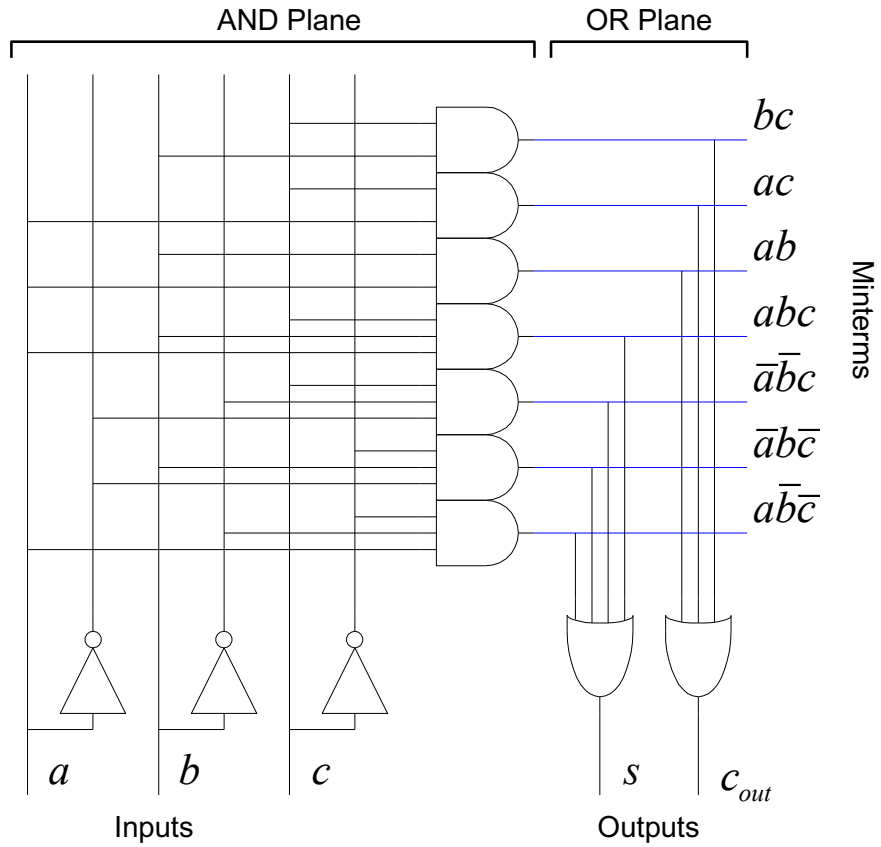


Looks like 6 4-input pseudo-nMOS NORs

PLAs

- A *Programmable Logic Array* performs any function in sum-of-products form.
- *Literals*: inputs & complements
- *Products / Minterms*: AND of literals
- *Outputs*: OR of Minterms

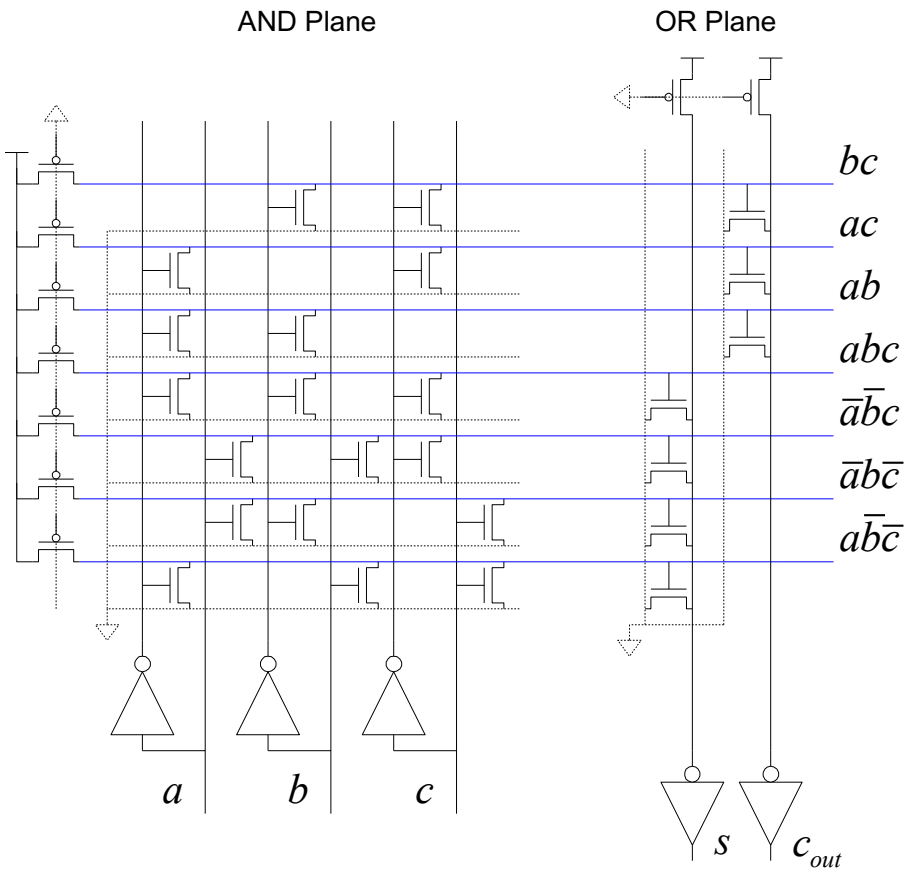
- Example: Full Adder

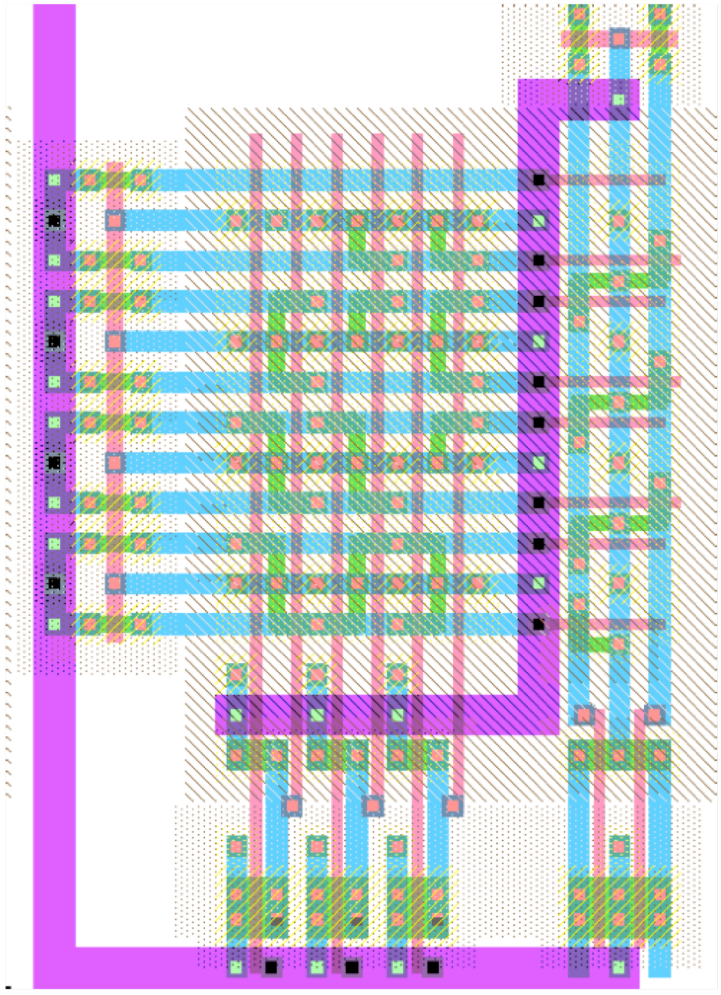


$$s = a\bar{b}\bar{c} + \bar{a}b\bar{c} + \bar{a}\bar{b}c + abc$$

$$c_{\text{out}} = ab + bc + ac$$

PLA Schematic & Layout



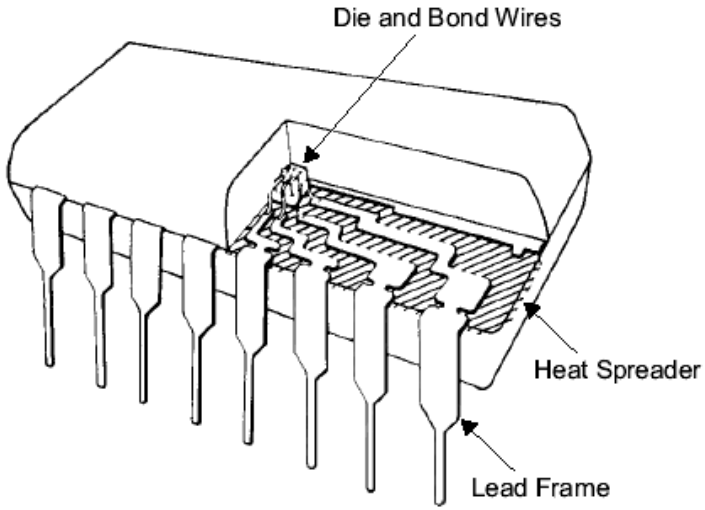


Low Power Design

- **Reduce dynamic power**
  - α: clock gating, sleep mode
  - C: small transistors (esp. on clock), short wires
  - V_{DD}: lowest suitable voltage
  - f: lowest suitable frequency
- **Reduce static power**
  - Selectively use ratioed circuits
  - Selectively use low V_t devices
  - Leakage reduction:
    stacked devices, body bias, low temperature

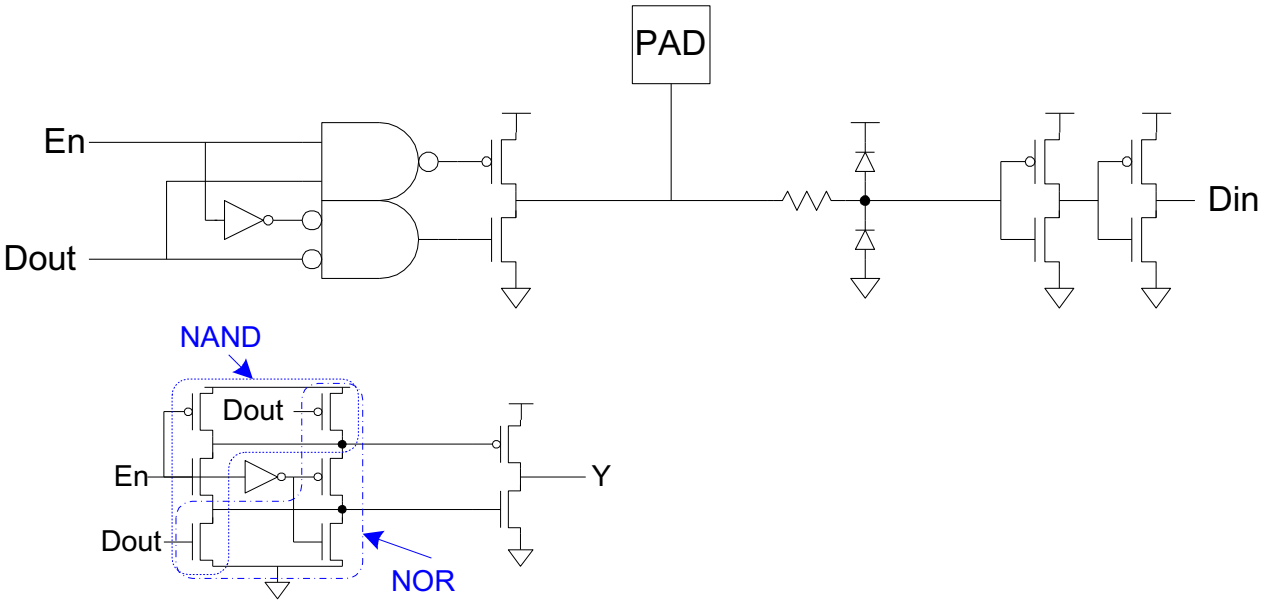
Chip-to-Package Bonding

- **Traditionally, chip is surrounded by *pad frame***
  - Metal pads on 100 – 200 μm pitch
  - Gold *bond wires* attach pads to package
  - *Lead frame* distributes signals in package
  - Metal *heat spreader* helps with cooling



Bidirectional Pads

- Combine input and output pad
- Need tristate driver on output
  - Use enable signal to set direction
  - Optimized tristate avoids huge series transistors



Device Scaling

Table 4.15 Influence of scaling on MOS device characteristics

Parameter	Sensitivity	Constant Field	Lateral
Scaling Parameters			
Length: L		$1/S$	$1/S$
Width: W		$1/S$	1
Gate oxide thickness: t_{ox}		$1/S$	1
Supply voltage: V_{DD}		$1/S$	1
Threshold voltage: V_{tn}, V_{tp}		$1/S$	1
Substrate doping: N_A		S	1
Device Characteristics			
β	$\frac{W}{L}\frac{1}{t_{ox}}$	S	S
Current: I_{ds}	$\beta(V_{DD}-V_t)^2$	$1/S$	S
Resistance: R	$\frac{V_{DD}}{I_{ds}}$	1	$1/S$
Gate capacitance: C	$\frac{WL}{t_{ox}}$	$1/S$	$1/S$
Gate delay: τ	RC	$1/S$	$1/S^2$
Clock frequency: f	$1/\tau$	S	S^2
Dynamic power dissipation (per gate): P	CV^2f	$1/S^2$	S
Chip area: A		$1/S^2$	1
Power density	P/A	1	S
Current density	I_{ds}/A	S	S

Interconnect Delay

Table 4.16 Influence of scaling on interconnect characteristics

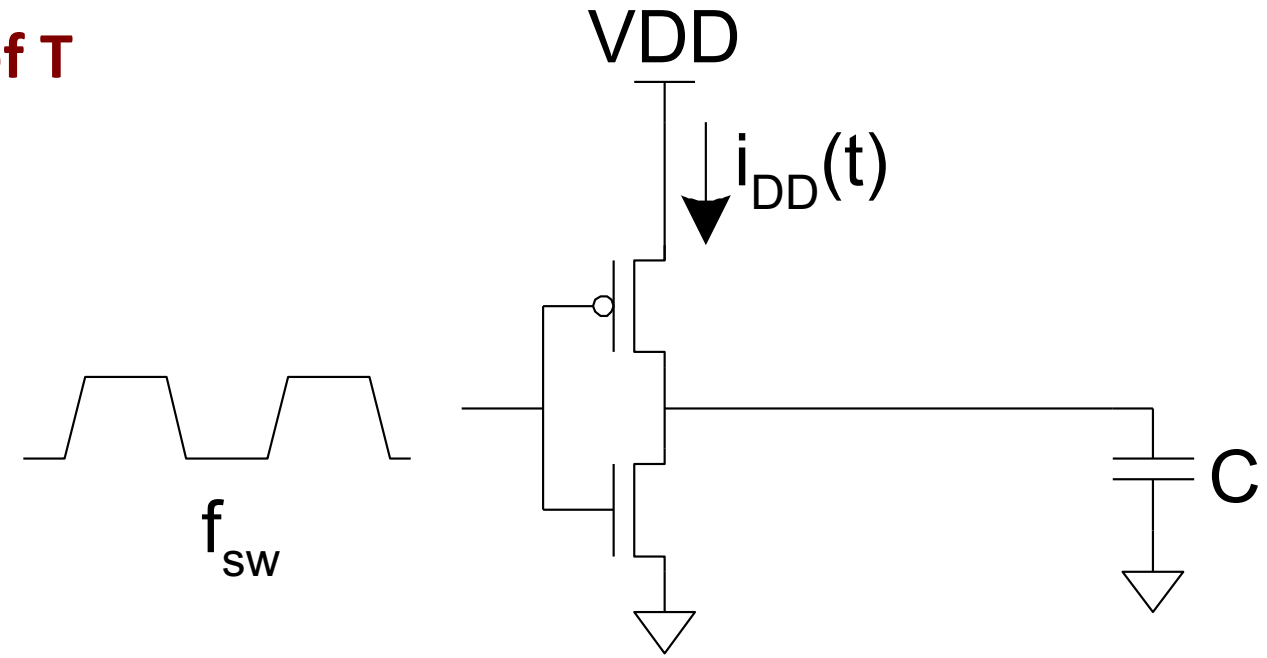
Parameter	Sensitivity	Reduced Thickness	Constant Thickness
Scaling Parameters			
Width: w		$1/S$	
Spacing: s		$1/S$	
Thickness: t		$1/S$	1
Interlayer oxide height: h		$1/S$	
Local/Scaled Interconnect Characteristics			
Length: l		$1/S$	
Unrepeated wire RC delay	$l^2 t_{wu}$	1	between $1/S$, 1
Repeated wire delay	$l t_{wr}$	$\sqrt{1/S}$	between $1/S$, $\sqrt{1/S}$
Global Interconnect Characteristics			
Length: l		D_c	
Unrepeated wire RC delay	$l^2 t_{wu}$	$S^2 D_c^2$	between SD_c^2, $S^2 D_c^2$
Repeated wire delay	$l t_{wr}$	$D_c\sqrt{S}$	between D_c, $D_c\sqrt{S}$

Energy and Power

- **Energy is drawn from a voltage source**
- **Instantaneous Power:** $P(t) = i_{DD}(t)V_{DD}$
- **Energy:** $E = \int_0^T P(t)dt = \int_0^T i_{DD}(t)V_{DD}dt$
- **Average Power:** $P_{\text{avg}} = \frac{E}{T} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$

Dynamic Power

- Dynamic power required to charge and discharge load capacitances when transistors switch
- One cycle involves a rising and falling output
- On rising output, charge Q = CV_{DD} is required
- On falling output, charge is dumped to GND
- This repeats $T*f_{sw}$ times over an interval of T



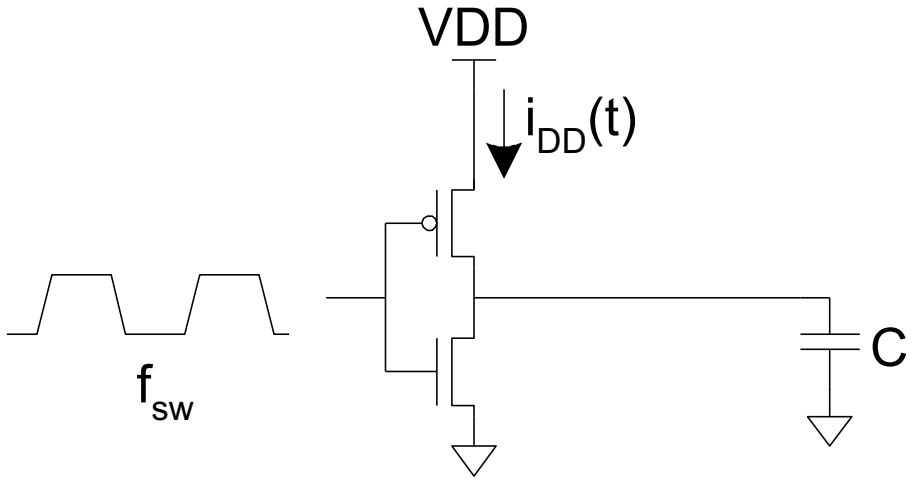
Dynamic Power (Cont.)

$$P_{\text{dynamic}} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$$

$$= \frac{V_{DD}}{T}\int_0^T i_{DD}(t)dt$$

$$= \frac{V_{DD}}{T}\left[Tf_{\text{sw}}CV_{DD}\right]$$

$$= CV_{DD}{}^2 f_{\text{sw}}$$



Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = ½$
  - Dynamic gates:
    - Switch either 0 or 2 times per cycle, $\alpha = 1$
  - Static gates:
    - Depends on the type of gate and logic network, but typically $\alpha = 0.1 – 0.2$
- Dynamic power: $P_{dyn} = \alpha * C * V_{dd} * \Delta V * freq$

Activity Factor Estimation

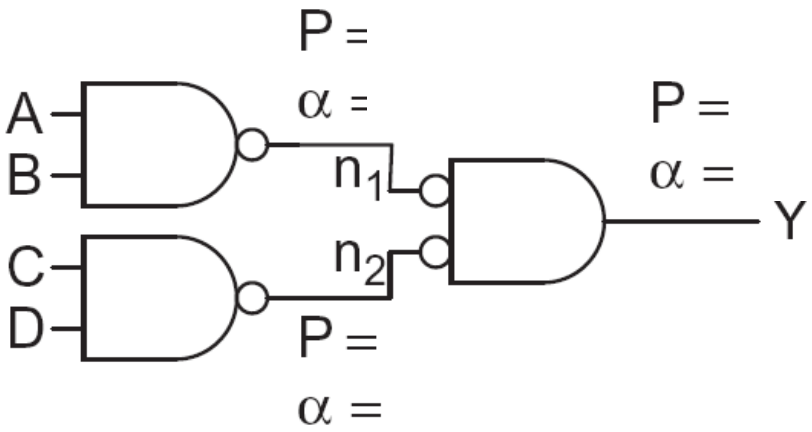
- Let P_i = Prob(node i = 1)
  - $\overline{P}_i = 1-P_i$
- $\alpha_i = \overline{P}_i * P_i$
- Completely random data has P = 0.5 and α = 0.25
- Data is often not completely random
  - e.g. upper bits of 64-bit words representing bank account balances are usually 0
- Data propagating through ANDs and ORs has lower activity factor
  - Depends on design, but typically $\alpha \approx 0.1$

Switching Probability

Gate	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1 - \overline{P}_A \overline{P}_B$
NAND2	$1 - P_A P_B$
NOR2	$\overline{P}_A \overline{P}_B$
XOR2	$P_A \overline{P}_B + \overline{P}_A P_B$

Switching Probability Example

- A 4-input AND is built out of two levels of gates
- Estimate the activity factor at each node if the inputs have P = 0.5

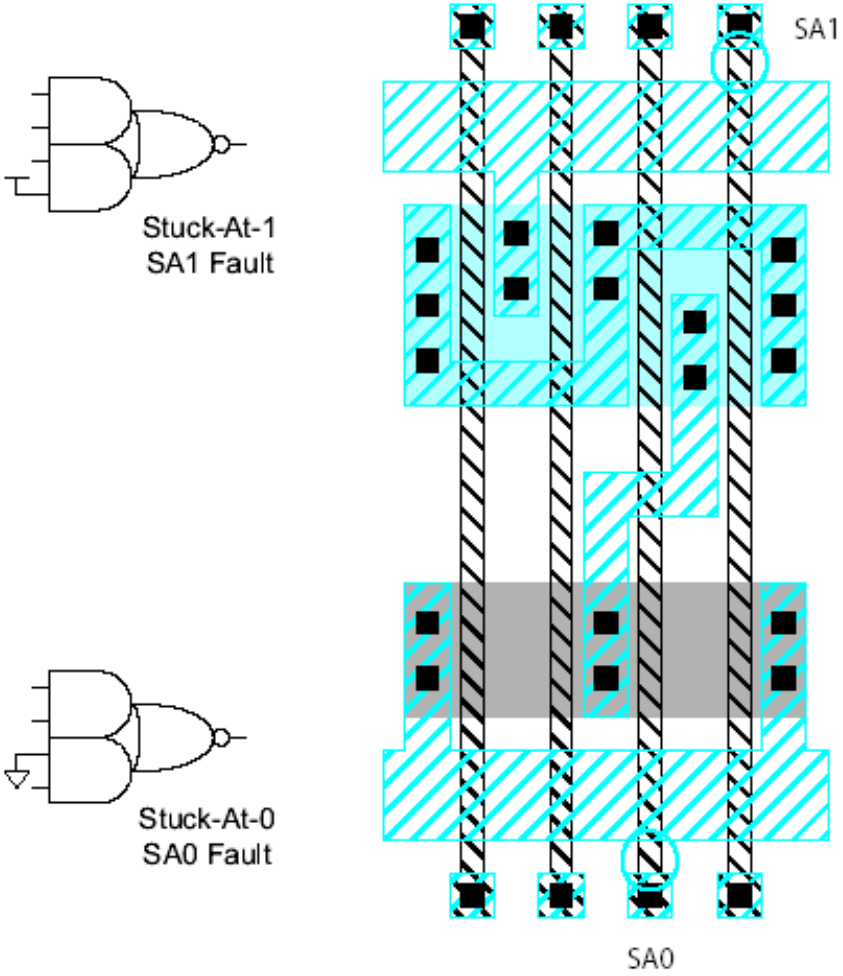


Stuck-At Faults

- **How does a chip fail?**
  - Need "fault model"
  - Usually failures are shorts between two conductors or opens in a conductor
  - This can cause very complicated behavior

- **A simpler model: *Stuck-At***
  - Assume all failures cause nodes to be "stuck-at" 0 or 1, i.e. shorted to GND or V_{DD}
  - Not quite true, but works well in practice

Examples



Observability & Controllability

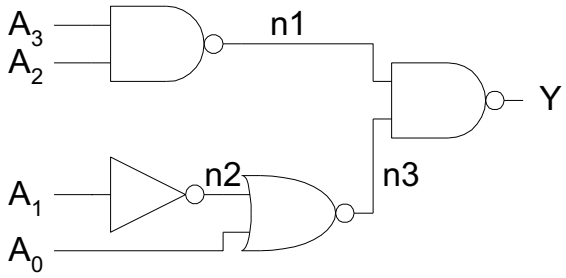
- ***Observability*: ease of observing a node by watching external output pins of the chip**
- ***Controllability*: ease of forcing a node to 0 or 1 by driving input pins of the chip**
- **Combinational logic is usually easy to observe and control**
- **Finite state machines can be very difficult, requiring many cycles to enter desired state**
  - **Especially if state transition diagram is not known to the test engineer**

Test Pattern Generation

- **Manufacturing test ideally would check every node in the circuit to prove it is not stuck.**
- **Apply the smallest sequence of test vectors necessary to prove each node is not stuck.**
- **Good observability and controllability reduces number of test vectors required for manufacturing test.**
  - **Reduces the cost of testing**
  - **Motivates design-for-test**

Test Example

	SA1	SA0
A_3	{0110}	{1110}
A_2	{1010}	{1110}
A_1	{0100}	{0110}
A_0	{0110}	{0111}
n1	{1110}	{0110}
n2	{0110}	{0100}
n3	{0101}	{0110}
Y	{0110}	{1110}



- Minimum set: {0100, 0101, 0110, 0111, 1010, 1110}