EE 466/586
VLSI Design

Partha Pande
School of EECS
Washington State University
pande@eecs.wsu.edu

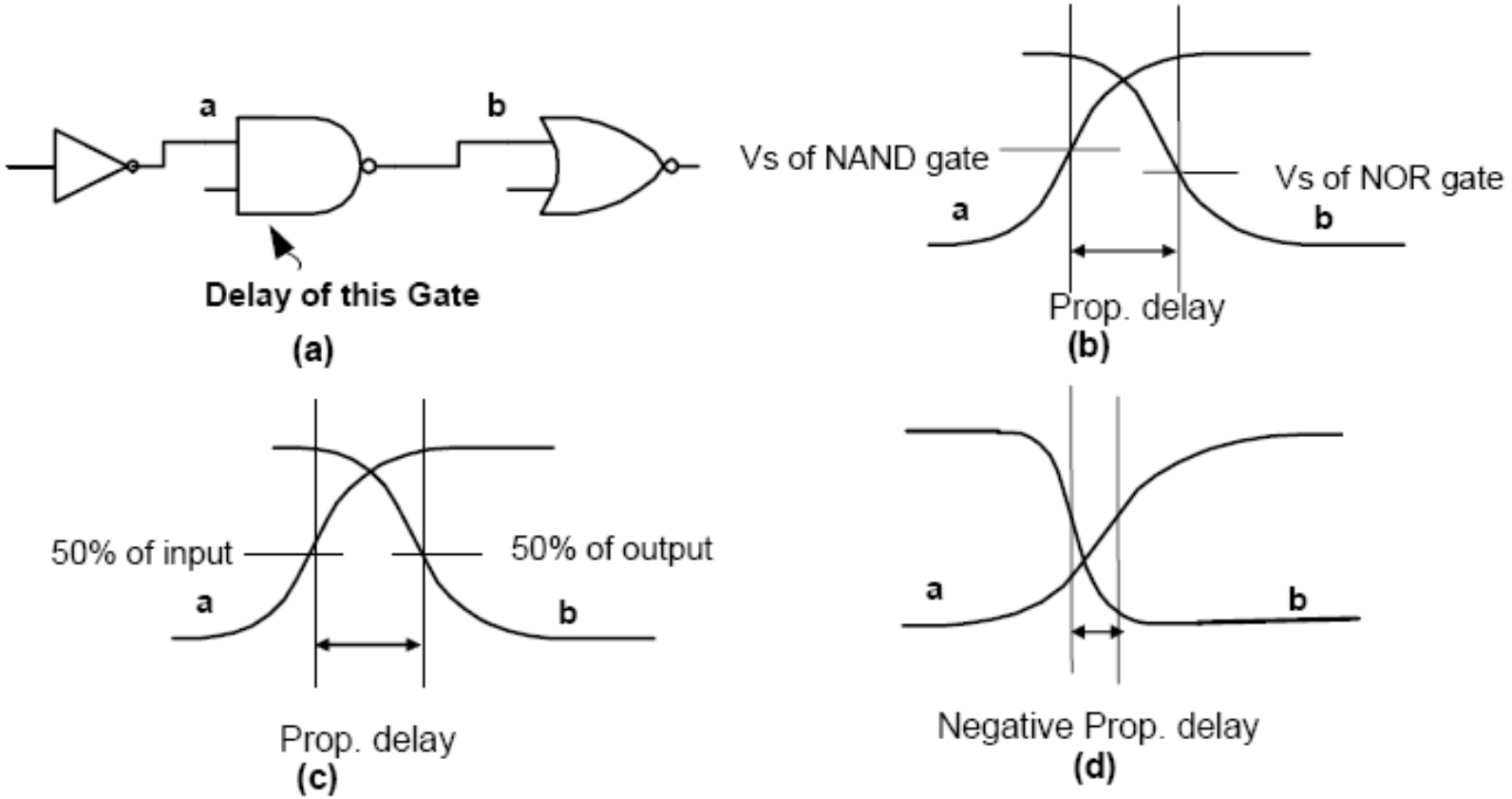
Lecture 9

Propagation delay

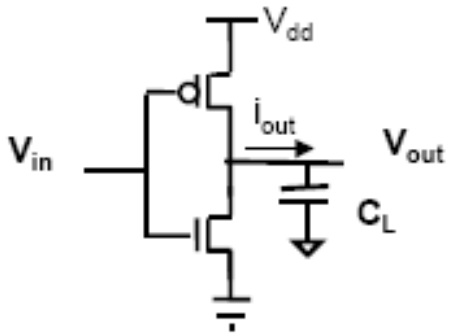
Power and delay Tradeoffs

- Follow board notes

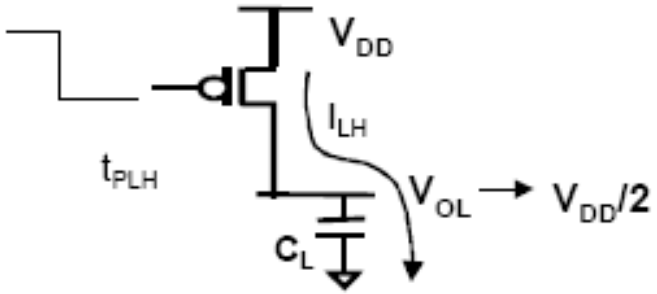
Propagation Delay



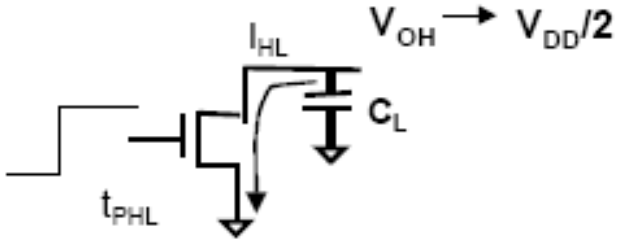
Switching Time



(a)



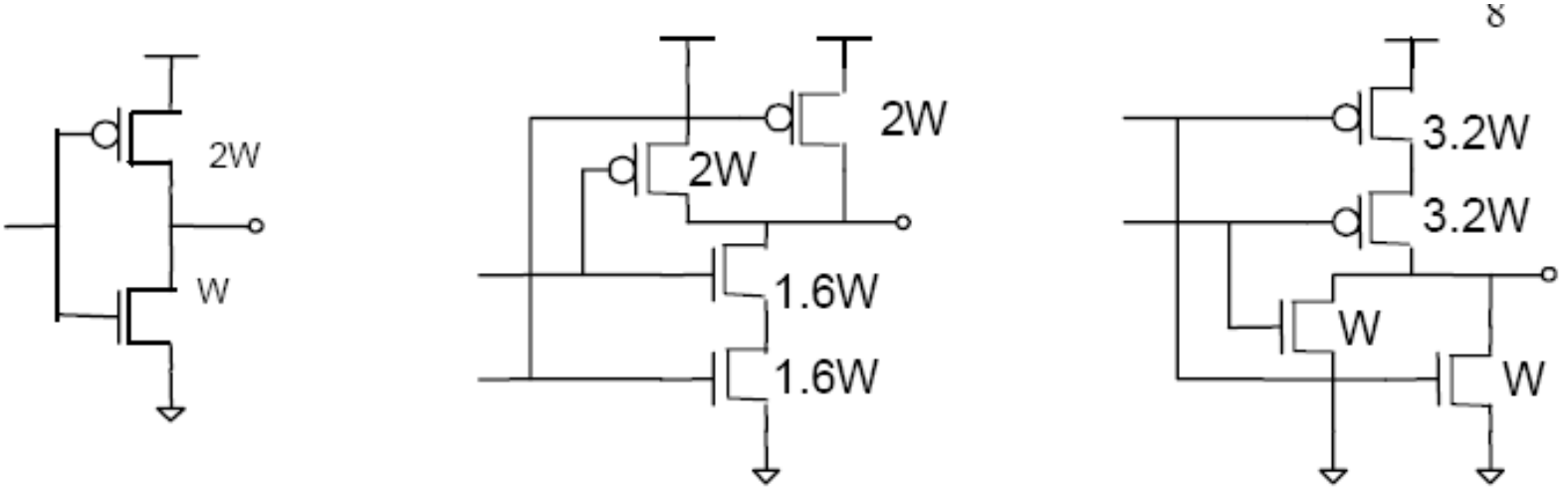
(b)



(c)

- *Follow board notes*

Gate sizing-Velocity Saturation

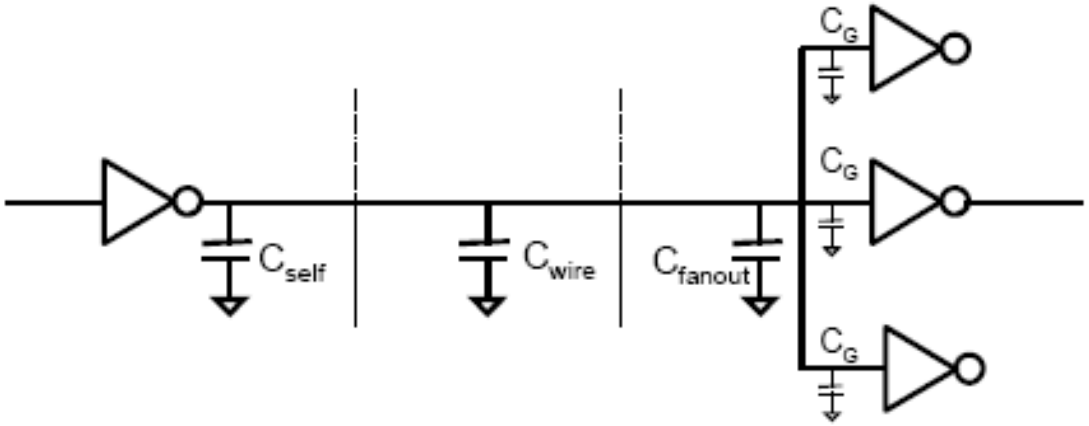


- Size of the stacked devices will be smaller

Gate sizing-Velocity Saturation

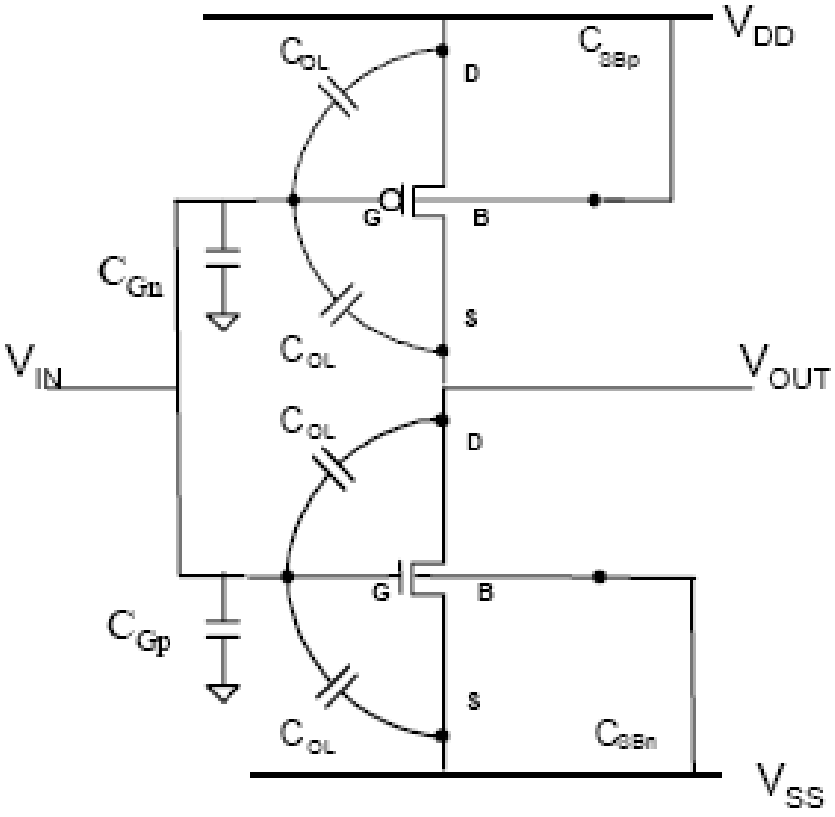
- A single device of size W takes longer to discharge the load capacitance than two series stacked 2W devices
- The single W device is in saturation for the entire transition of the output from V_{DD} to $V_{DD}/2$
- The two series devices operate in the different regions
- Transistor lower in the stack will be in linear, with $V_{GS1}=V_{DD}$
- Upper transistor operates in saturation, $V_{GS2}=V_{DD}-V_{DS1}$

Load Capacitance



$$C_{load} = C_{self} + C_{wire} + C_{fanout}$$

Fanout Gate Capacitance



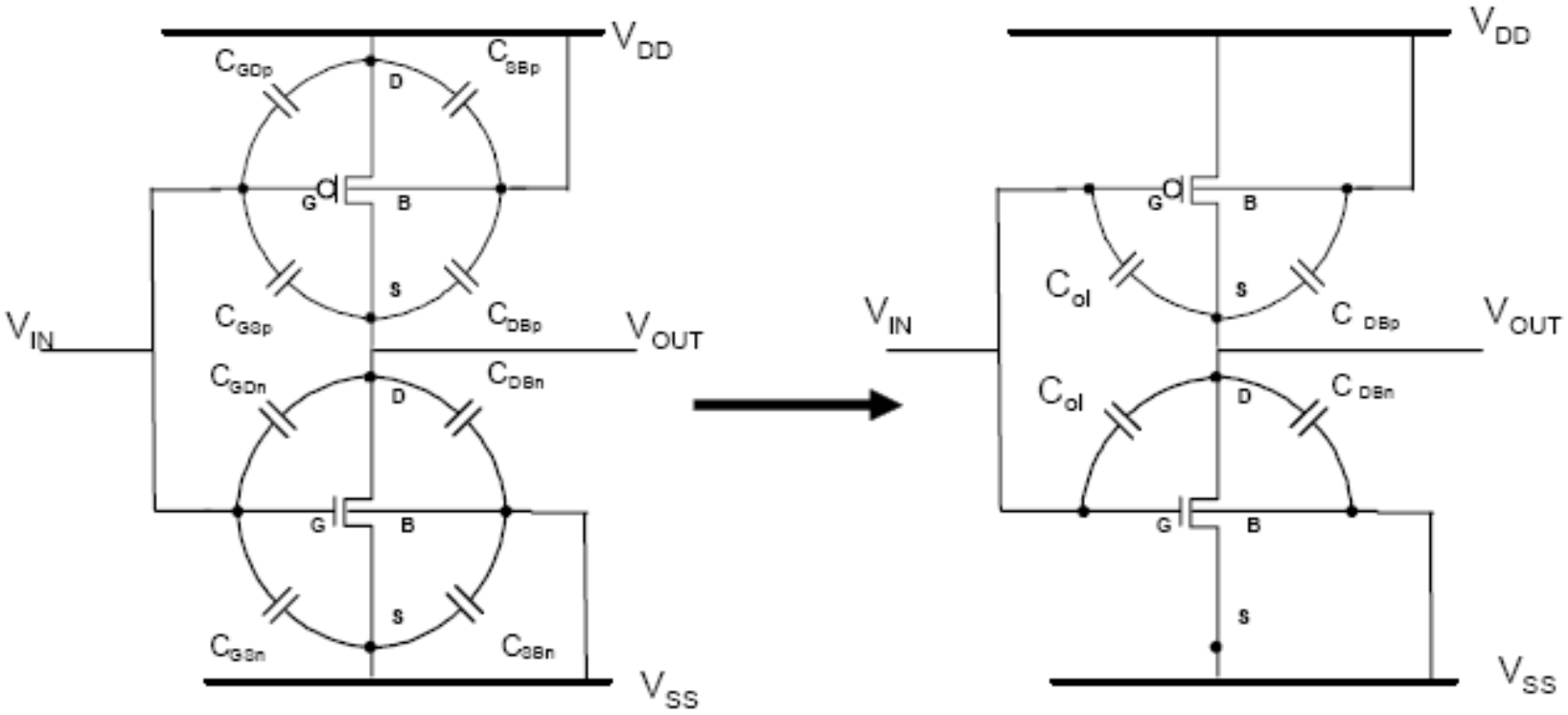
$$C_{fanout} = \sum C_G$$

$$C_G = C_{Gn} + 2C_{OL} + C_{Gp} + 2C_{OL}$$

$$= C_{ox}LW_n + 2C_{ol}W_n + C_{ox}LW_p + 2C_{ol}W_p$$

$$= (C_{ox}L + 2C_{ol})(W_n + W_p)$$

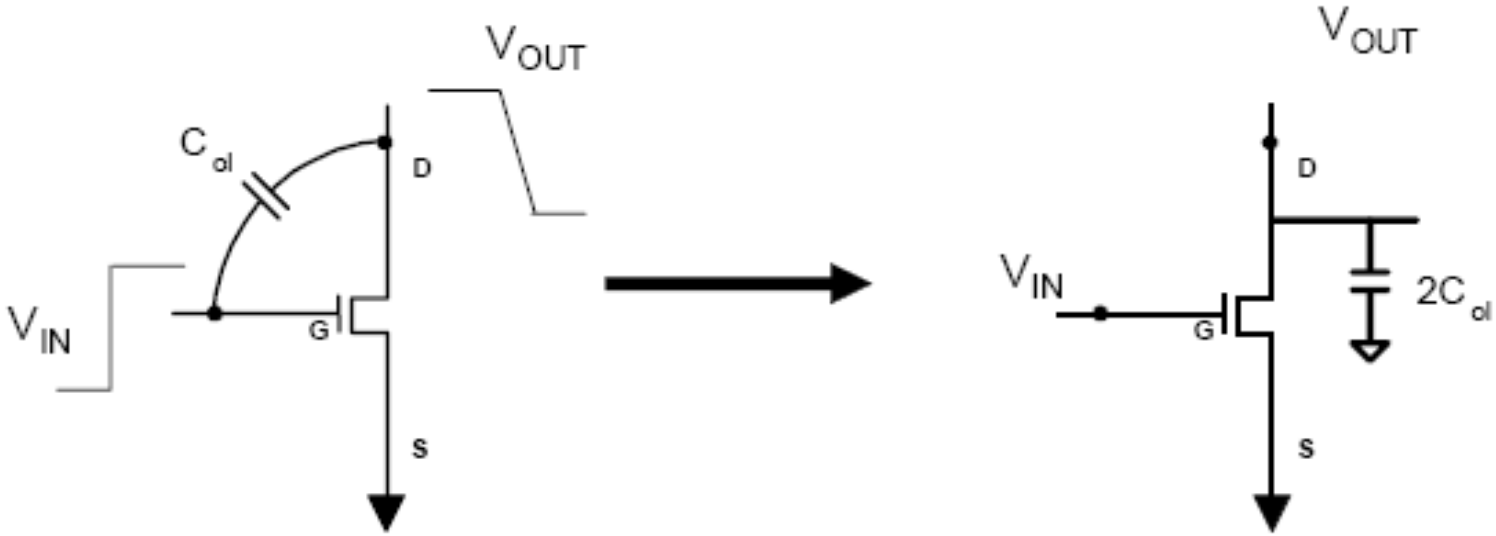
Self-Capacitance



Self-Capacitance

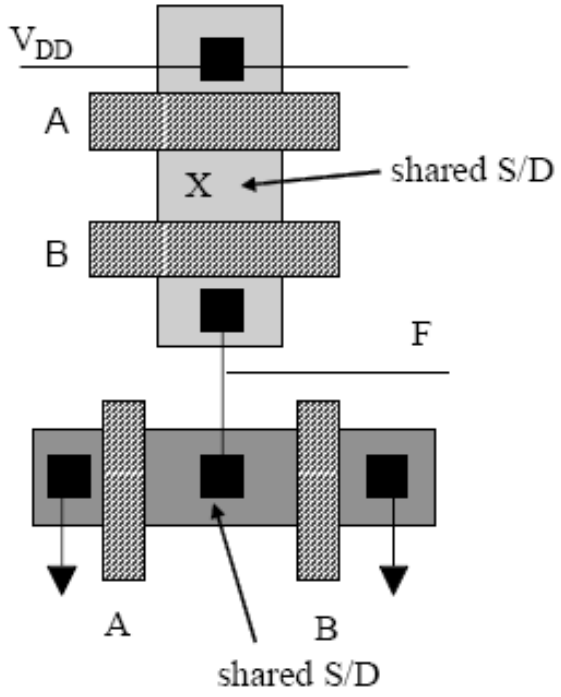
- Four main capacitances: C_{GS}, C_{GD}, C_{DB}, C_{SB}.
- Eliminate: C_{GSn}, C_{GSp}, C_{SBn}, and C_{SBp}.
- Apply a step input
  - One transistor is always off and the other is in saturation
  - In either region C_{GD} is negligible
  - We are left with only the overlap capacitances from gate-to-source and gate-to-drain, and one junction capacitance per device, C_{DBn} and C_{DBp}.

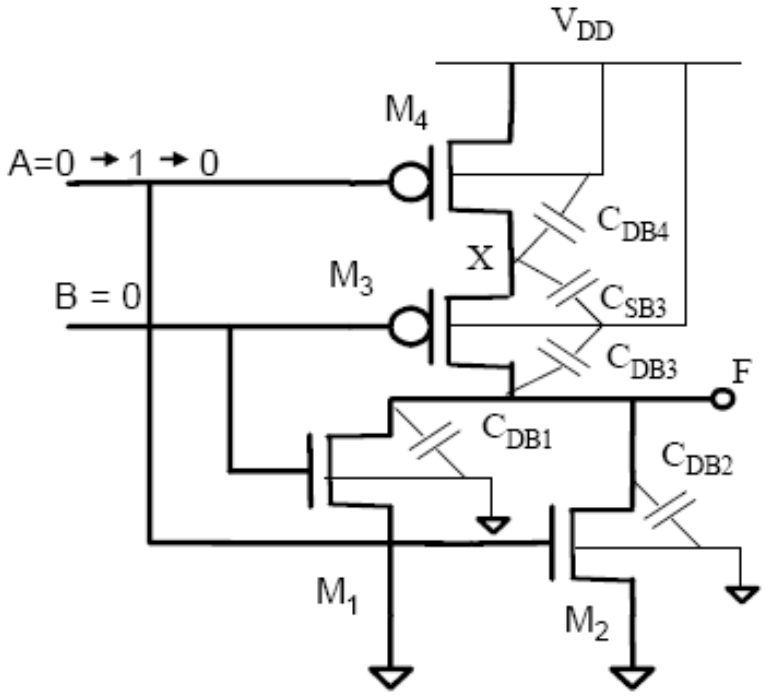
Overlap Capacitance



- *The input makes a transition from 0 to V_{DD} while the output makes a transition from V_{DD} to 0.*
  - *The overlap capacitance experiences a voltage swing of $2V_{DD}$.*
  - *Model this by assuming that the swing is only V_{DD} and then double the size of the capacitance and place it at the output.*

Effect of shared Source and Drain





Shared Source and Drain

- Let's examine the pull-down case
- In the worst-case, input A switches from low to high while B remains low since the capacitances at both the output node and the internal node X must also be discharged.

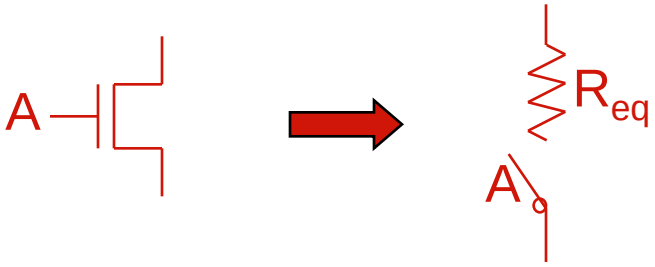
$$C_{self} = \underbrace{C_{DB1} + C_{DB2}}_{n^+ shared \quad S/D} + C_{DB3} + \underbrace{C_{SB3} + C_{DB4}}_{p^+ shared \quad S/D}$$

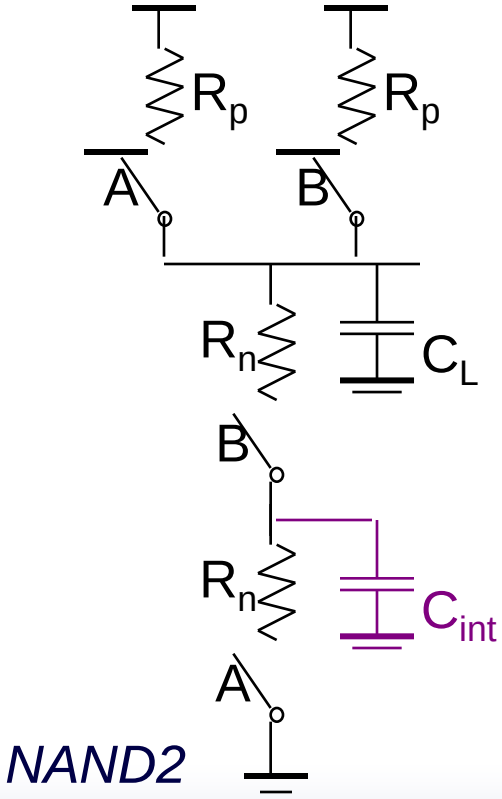
$$= C_{DB12} + C_{DB3} + C_{SDB34}$$

Note: If A stays low but B goes high, the only output capacitances to be charged are C_{DB12}+C_{DB3}.

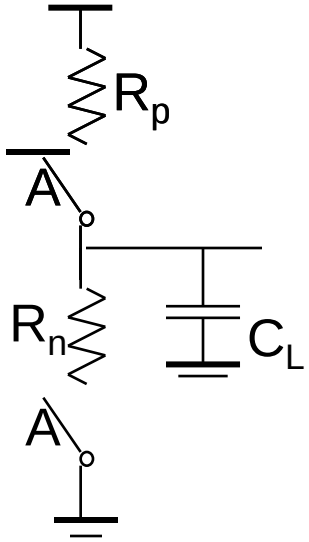
If we assume that input A switches from high to low while input B remains low, then the total capacitance to be charged is given by the above equation

Switch Delay Model

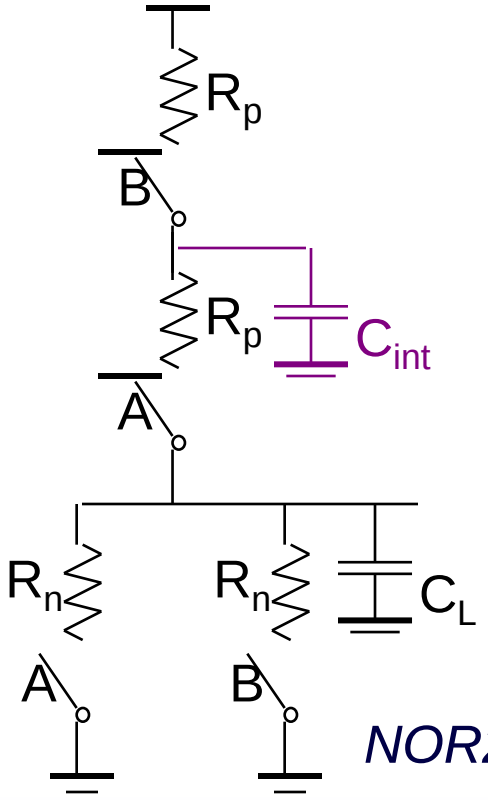




NAND2

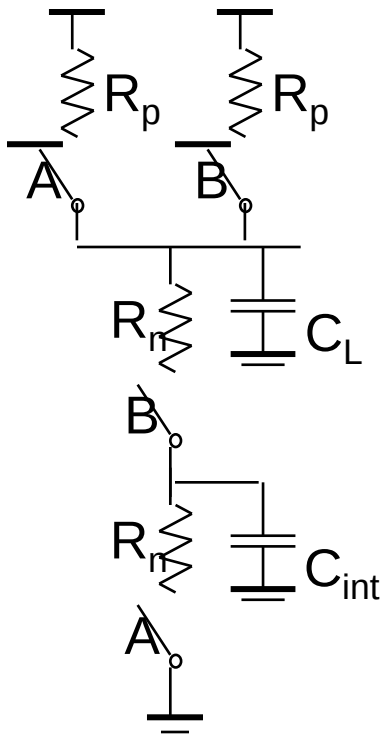


INV



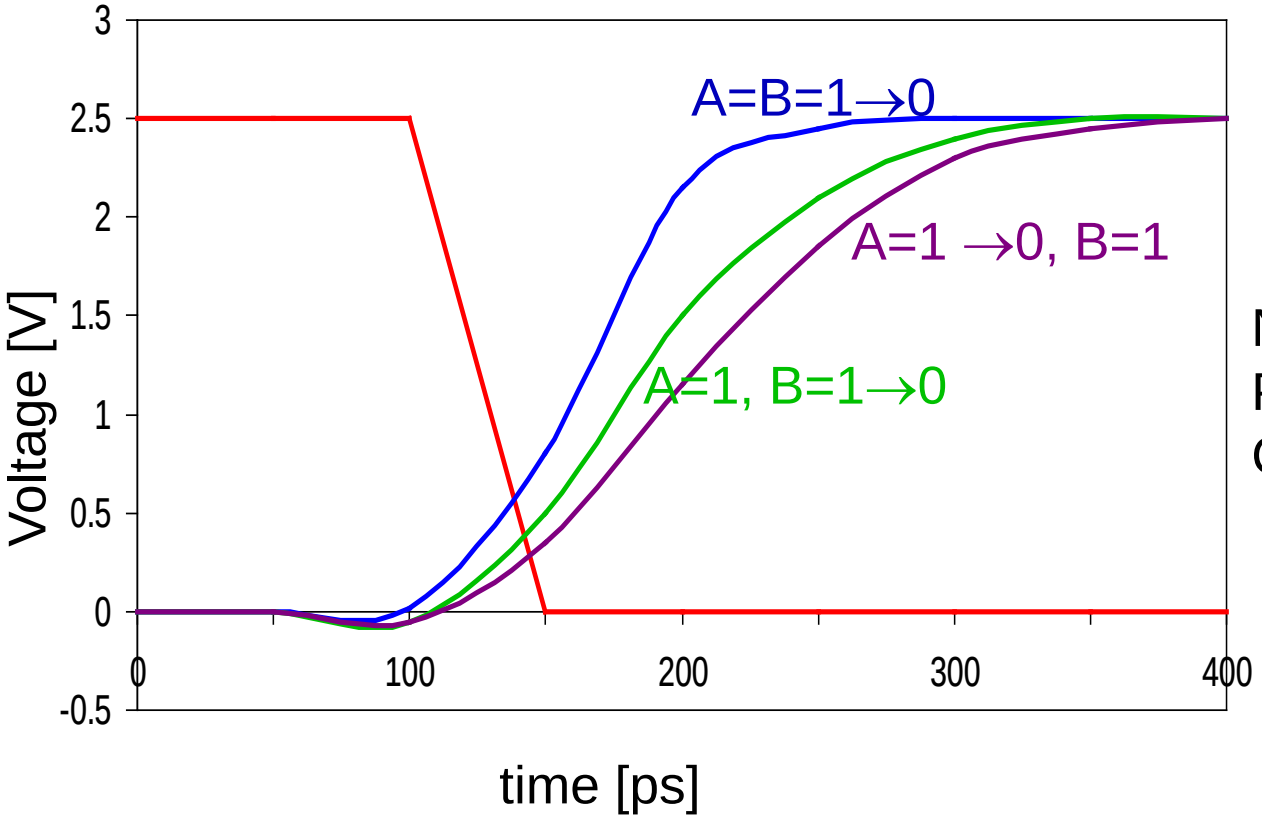
NOR2

Input Pattern Effects on Delay



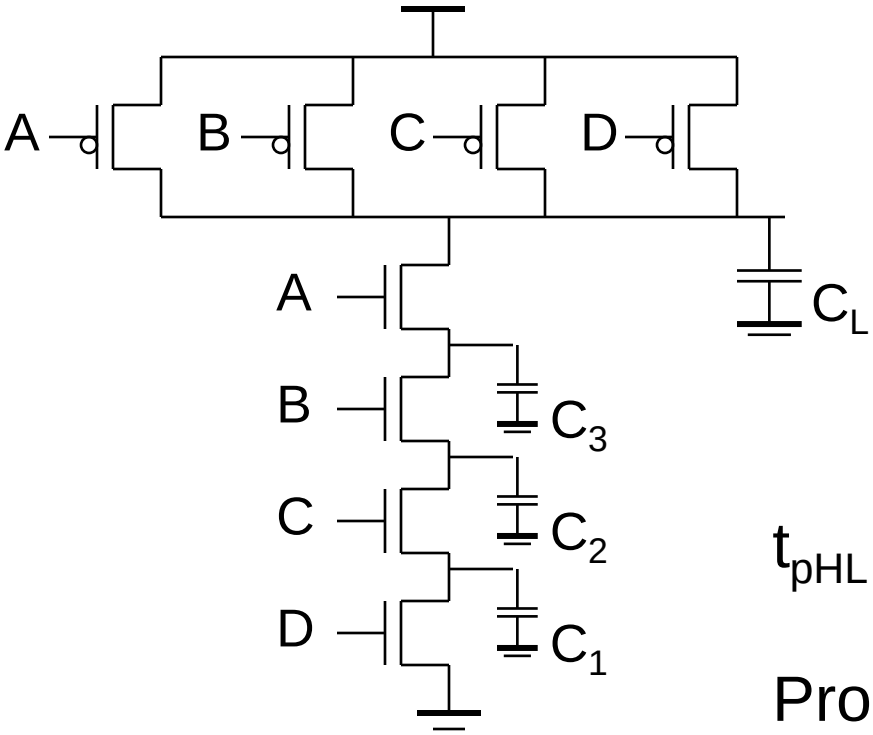
- Delay is dependent on the pattern of inputs
- Low to high transition
  - both inputs go low
    - delay is 0.69 $R_p/2$ C_L
  - one input goes low
    - delay is 0.69 R_p C_L
- High to low transition
  - both inputs go high
    - delay is 0.69 $2R_n$ C_L

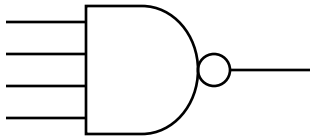
Delay Dependence on Input Patterns



NMOS = 0.5μm/0.25 μm
PMOS = 0.75μm/0.25 μm
C_L = 100 fF

Fan-In Considerations





Distributed RC model
(Elmore delay)

$$t_{pHL} = 0.69\, R_{eqn}(C_1+2C_2+3C_3+4C_L)$$

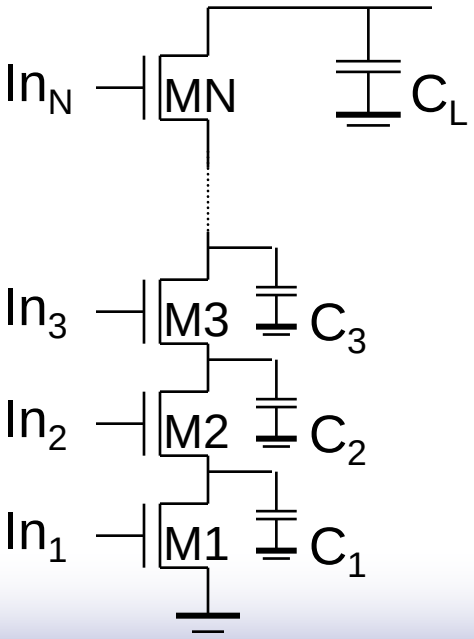
Propagation delay deteriorates rapidly as a function of fan-in – quadratically in the worst case.

t_p as a Function of Fan-In and Fan-Out

- Fan-in: quadratic due to increasing resistance and capacitance
- Fan-out: each additional fan-out gate adds two gate capacitances to C_L

Fast Complex Gates: Design Technique 1

- Transistor sizing
  - as long as fan-out capacitance dominates
- Progressive sizing



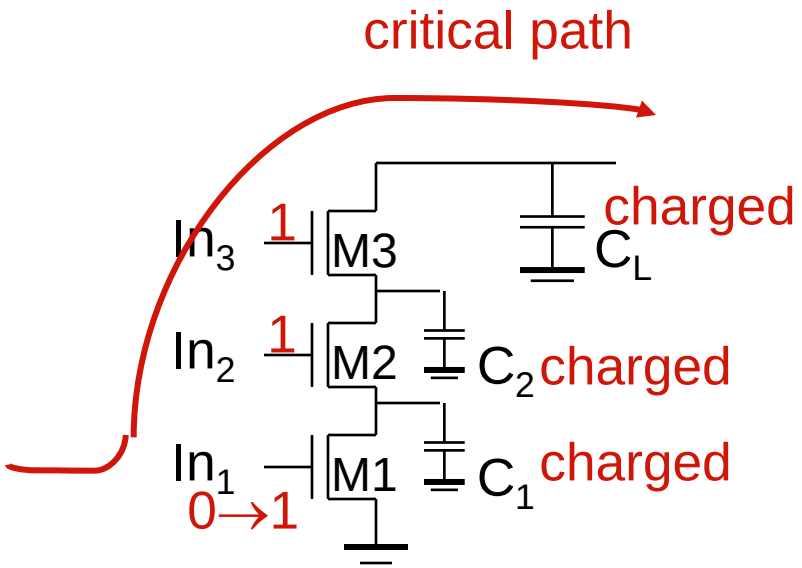
Distributed RC line

M1 > M2 > M3 > … > MN
(the FET closest to the output is the smallest)

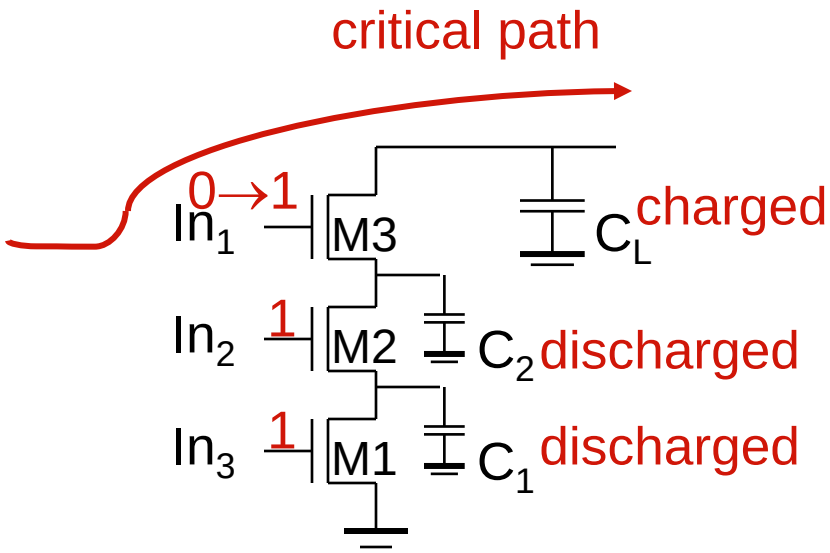
Can reduce delay by more than 20%; decreasing gains as technology shrinks

Fast Complex Gates: Design Technique 2

- Transistor ordering



delay determined by time to discharge C_L, C_1 and C_2



delay determined by time to discharge C_L

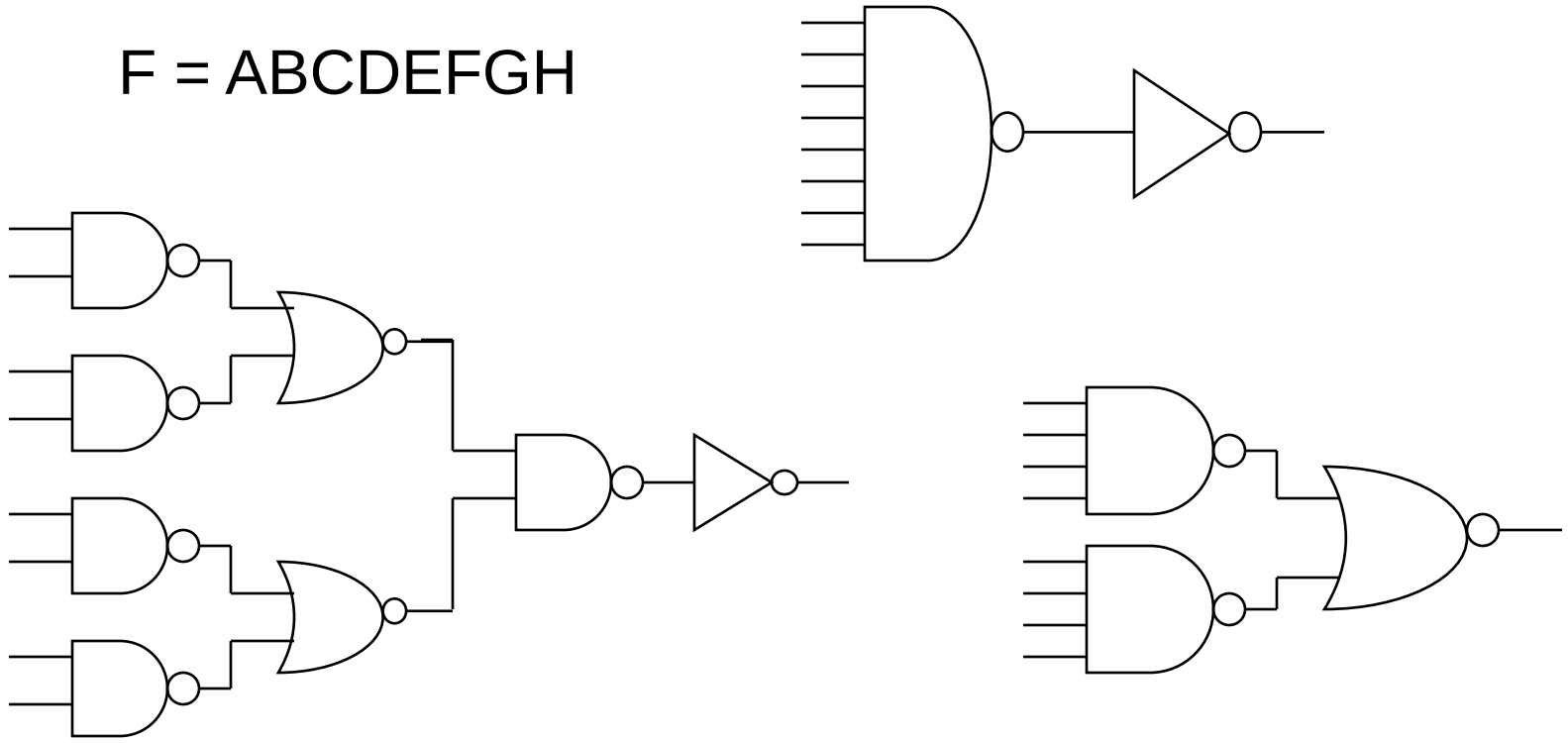
Propagation Delay Depends on the order of arrivals of the inputs

- In a series stack, the delay increases as the late arriving input is further from the output.
- Reorder the inputs such that the earliest signals arrive lower in the stack and the latest signals arrive near the top of the stack.
- Each device is progressively larger as we move from the output to ground
- Each one must discharge a progressively larger capacitance.

Fast Complex Gates: Design Technique 3

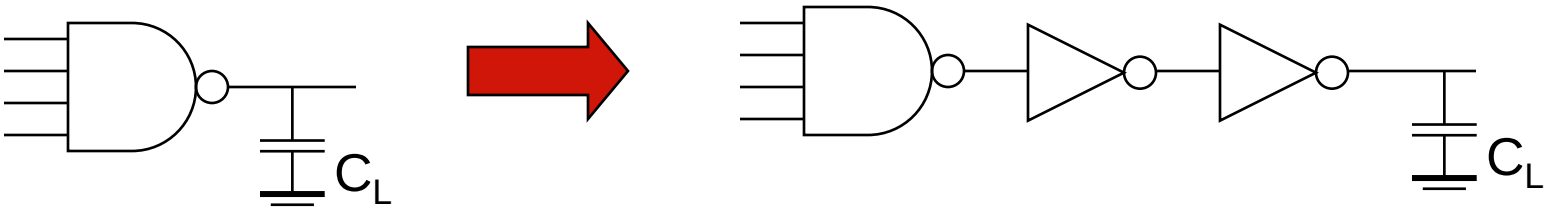
- Alternative logic structures

F = ABCDEFGH



Fast Complex Gates: Design Technique 4

- Isolating fan-in from fan-out using buffer insertion



Fast Complex Gates: Design Technique 5

- Reducing the voltage swing
  - linear reduction in delay
  - also reduces power consumption
- But the following gate is much slower!
- Or requires use of “sense amplifiers” on the receiving end to restore the signal level (memory design)