CSE477
VLSI Digital Circuits
Fall 2002

Lecture 11: Designing for Speed

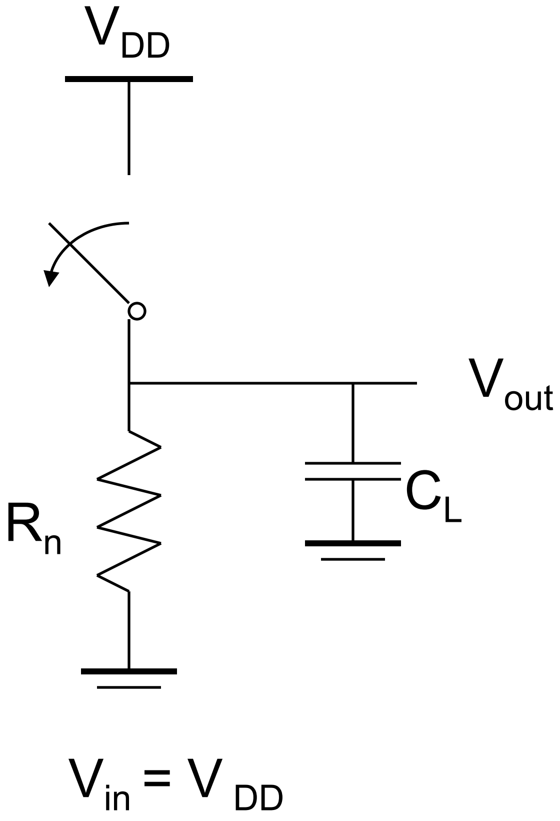
Mary Jane Irwin (www.cse.psu.edu/~mji)
www.cse.psu.edu/~cg477

[Adapted from Rabaey's *Digital Integrated Circuits*, ©2002, J. Rabaey et al.]

Cray was a legend in computers … said that he liked to hire inexperienced engineers right out of school, because they do not usually know what's supposed to be impossible.

The Soul of a New Machine, Kidder, pg. 77

Review: CMOS Inverter: Dynamic



$t_{pHL} = f(R_n, C_L)$

$t_{pHL} = 0.69\ R_{eqn}\ C_L$

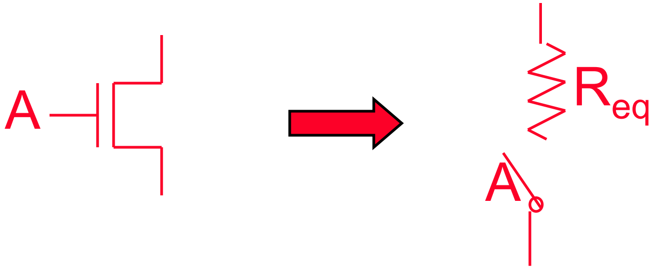
$t_{pHL} = 0.69\ (3/4\ (C_L\ V_{DD}) / I_{DSATn})$

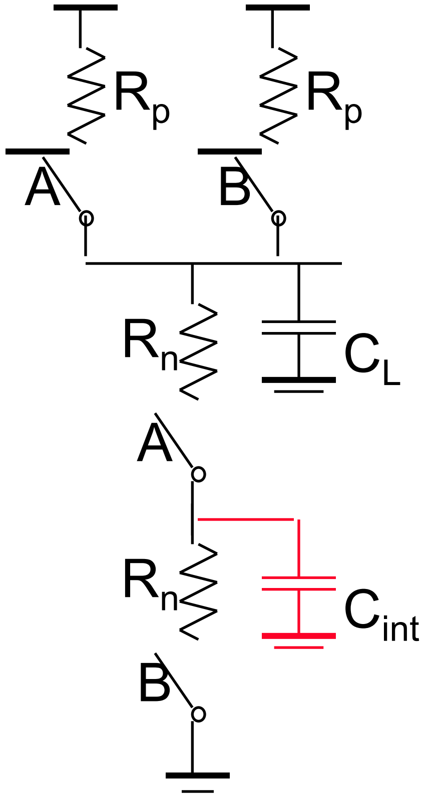
$= 0.52\ C_L\ /\ (W/L_n\ k'_n\ V_{DSATn})$

Review: Designing Inverters for Performance

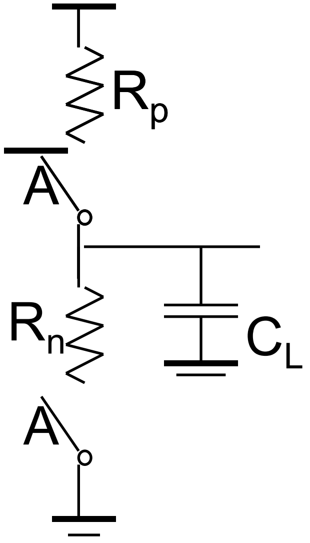
- Reduce C_L
  - internal diffusion capacitance of the gate itself
  - interconnect capacitance
  - fanout
- Increase W/L ratio of the transistor
  - the most powerful and effective performance optimization tool in the hands of the designer
  - watch out for self-loading!
- Increase V_{DD}
  - only minimal improvement in performance at the cost of increased energy dissipation
- Slope engineering - keeping signal rise and fall times smaller than or equal to the gate propagation delays and of approximately equal values
  - good for performance
  - good for power consumption

Switch Delay Model

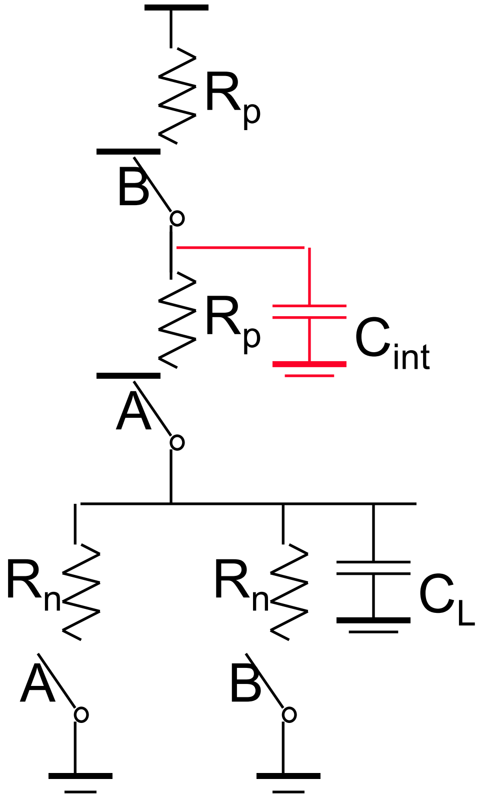




NAND

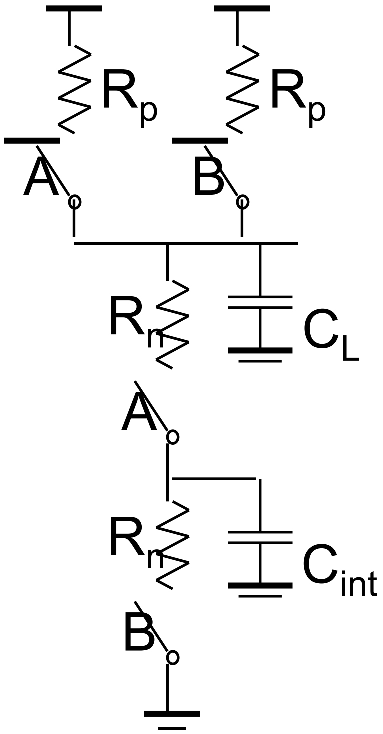


INVERTER



NOR

Input Pattern Effects on Delay



- Delay is dependent on the pattern of inputs

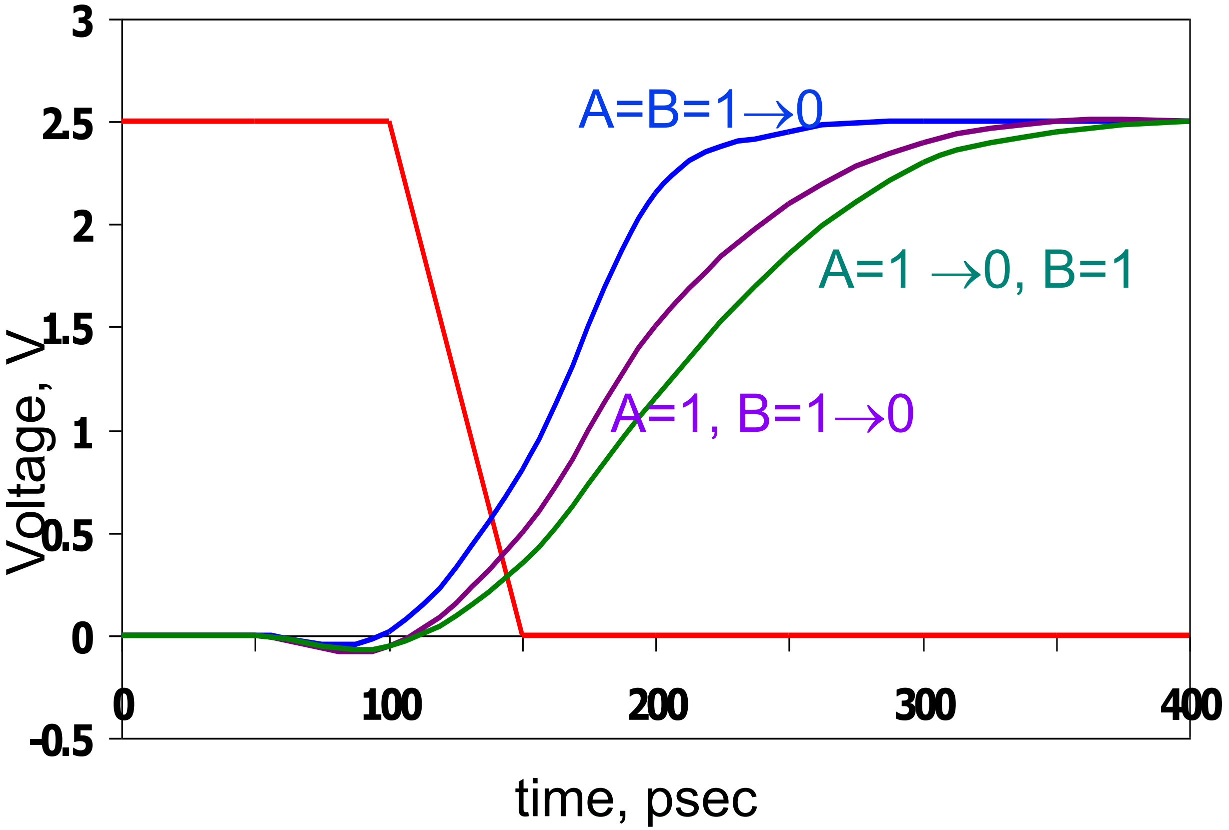
- Low to high transition
  - both inputs go low
    - delay is 0.69 $R_p/2$ C_L since two p-resistors are on in parallel
  - one input goes low
    - delay is 0.69 R_p C_L

- High to low transition
  - both inputs go high
    - delay is 0.69 $2R_n$ C_L

- Adding transistors in series (without sizing) slows down the circuit

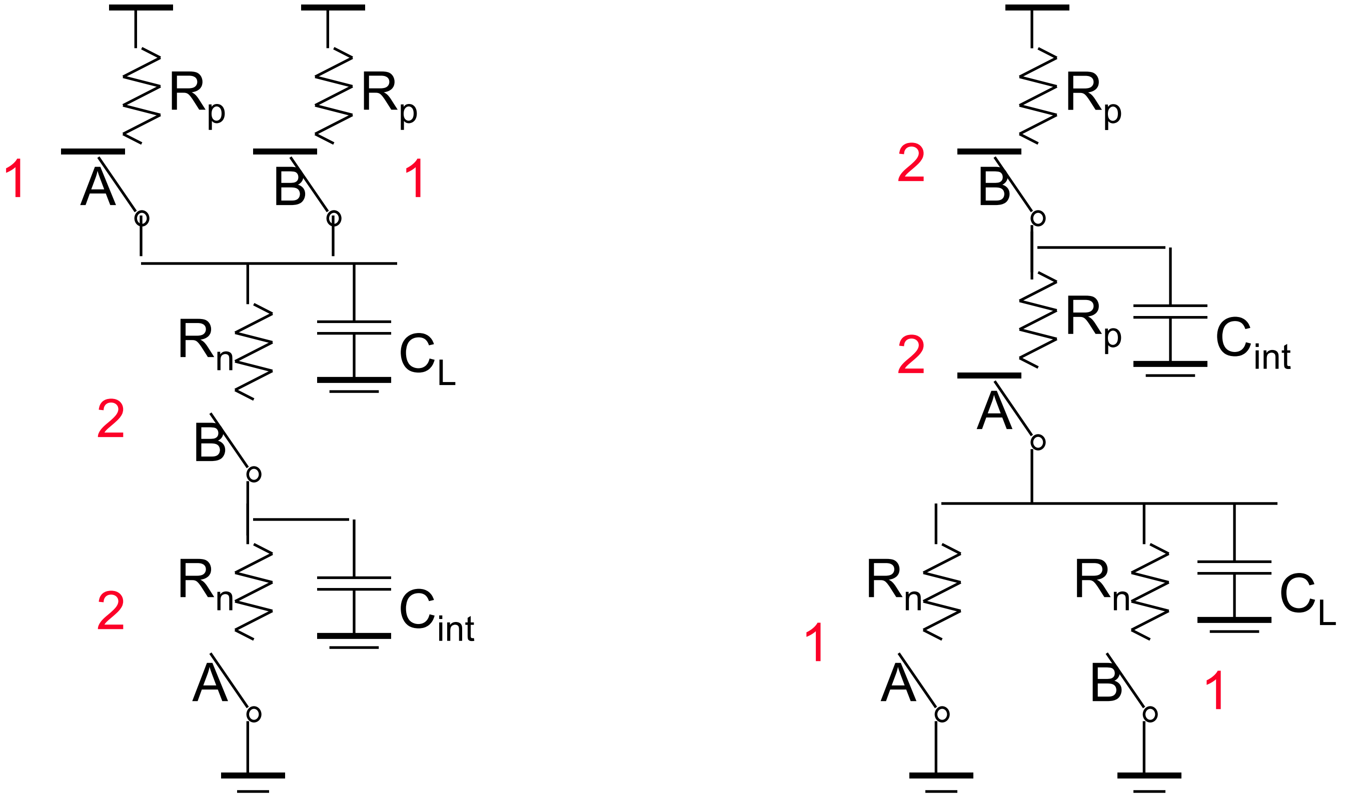
Delay Dependence on Input Patterns

2-input NAND with
NMOS = 0.5μm/0.25 μm
PMOS = 0.75μm/0.25 μm
C_L = 10 fF

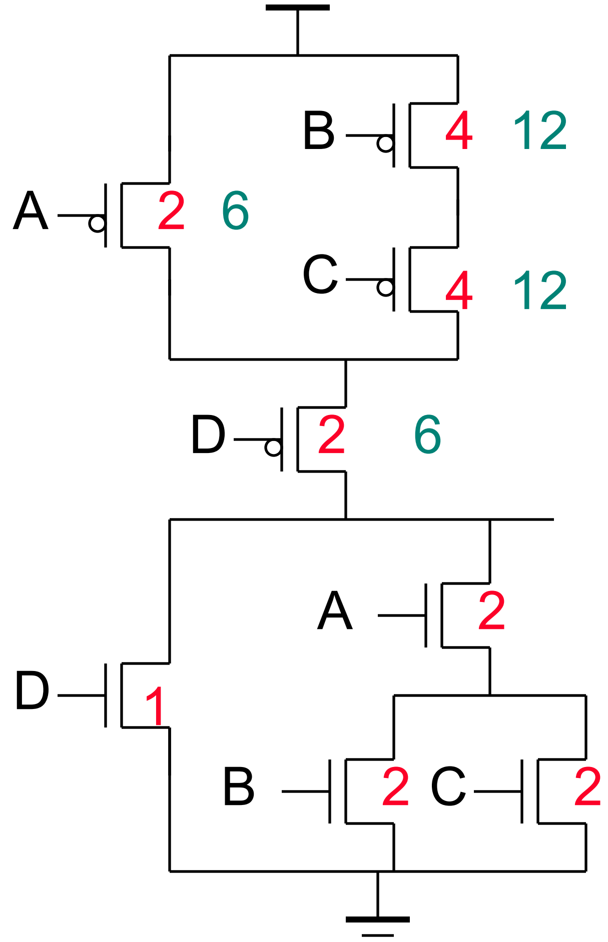


Input Data Pattern	Delay (psec)
A=B=0→1	69
A=1, B=0→1	62
A= 0→1, B=1	50
A=B=1→0	35
A=1, B=1→0	76
A= 1→0, B=1	57

Transistor Sizing

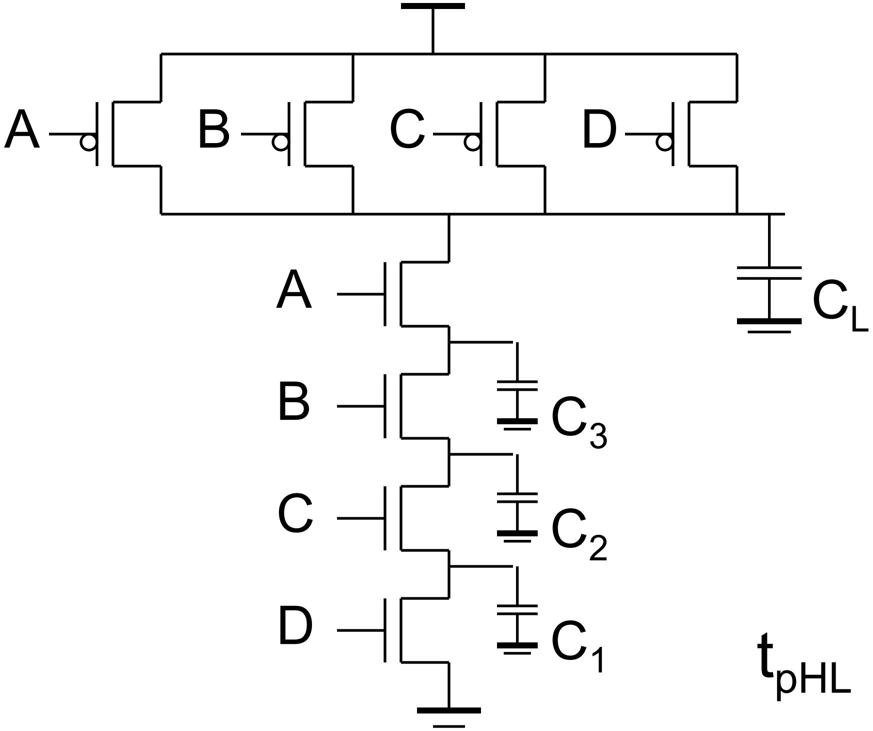


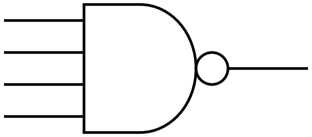
Transistor Sizing a Complex CMOS Gate



OUT = !(D + A • (B + C))

Fan-In Considerations



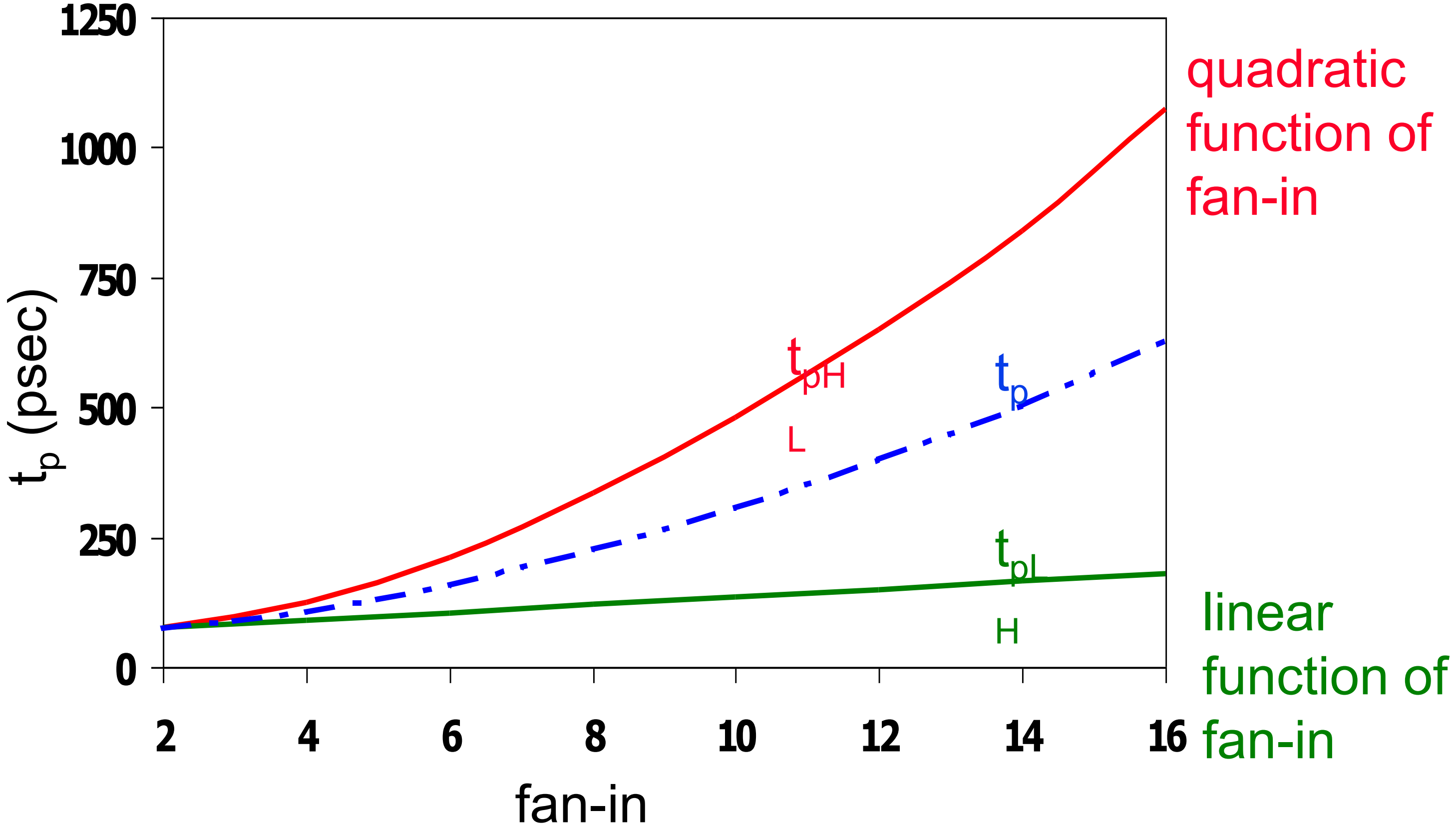


Distributed RC model
(Elmore delay)

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

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

t_p as a Function of Fan-In



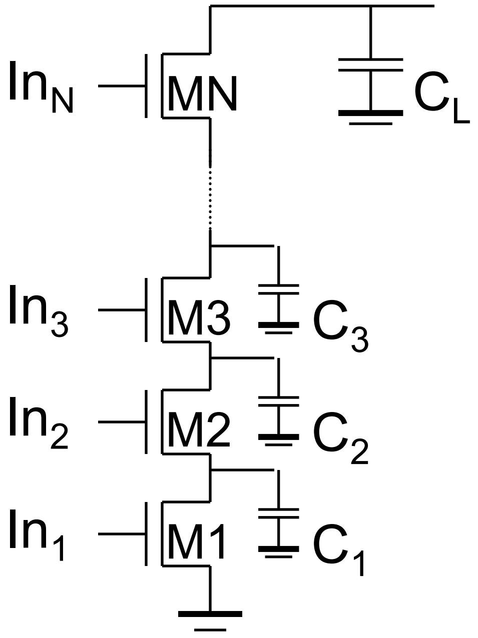
- Gates with a fan-in greater than 4 should be avoided.

Fast Complex Gates: Design Technique 1

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

- Progressive sizing

Distributed RC line



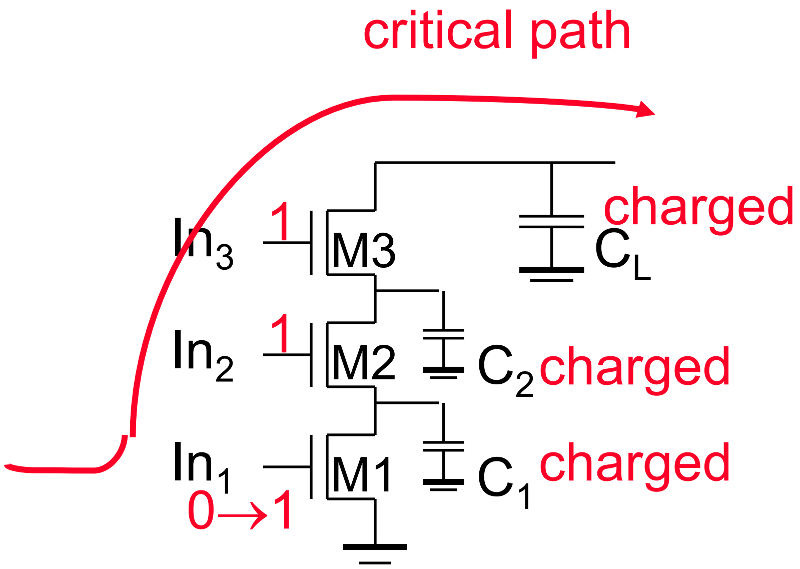
M1 > M2 > M3 > … > MN

(the fet closest to the output should be the smallest)

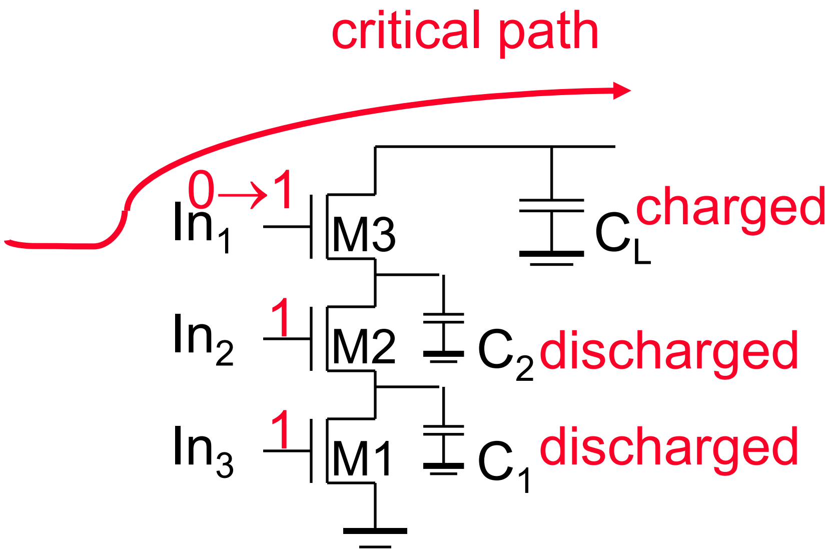
Can reduce delay by more than 20%; decreasing gains as technology shrinks

Fast Complex Gates: Design Technique 2

- Input re-ordering
  - when not all inputs arrive at the same time

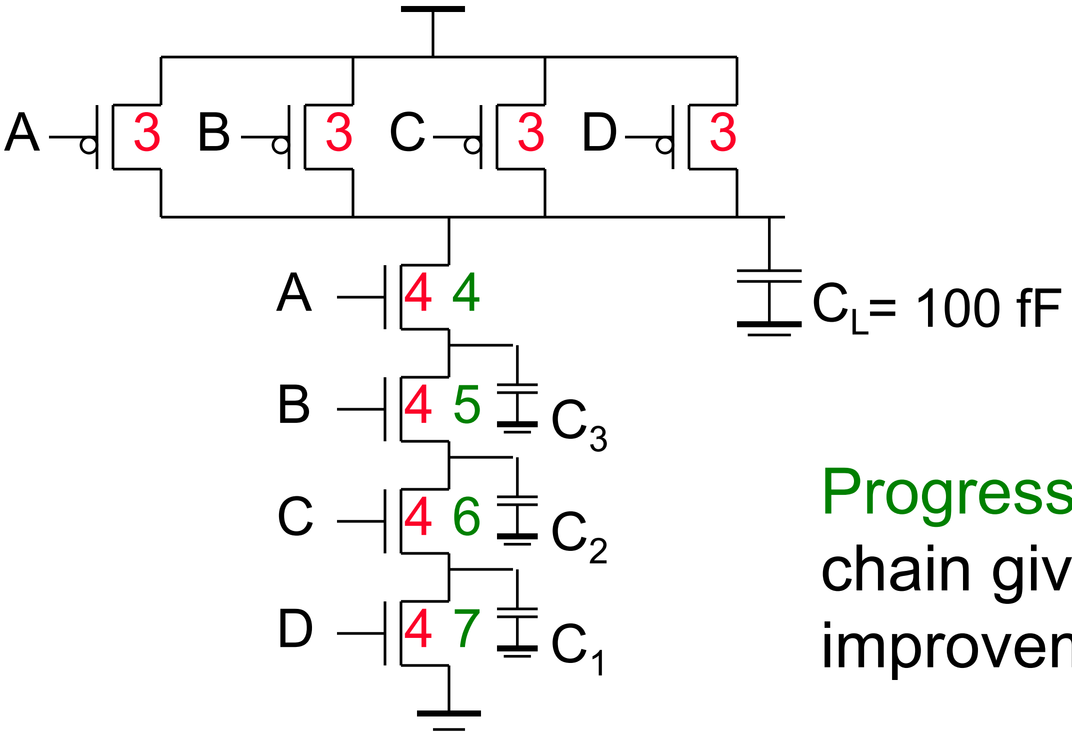


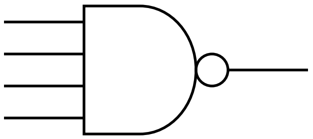
delay determined by time to discharge C_L, C_1 and C_2



delay determined by time to discharge C_L

Sizing and Ordering Effects





Progressive sizing in pull-down chain gives up to a 23% improvement.

Input ordering saves 5%

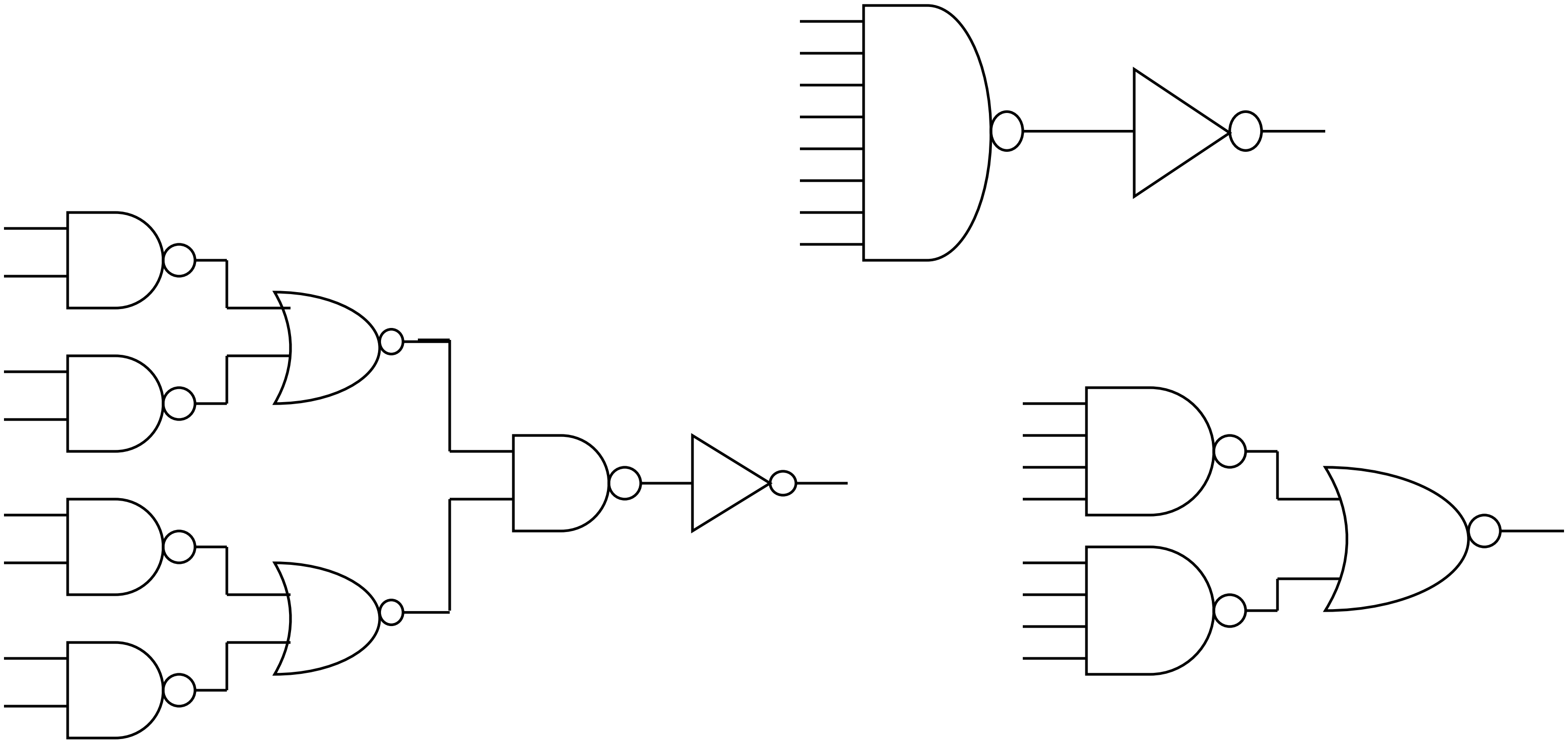
critical path A – 23%

critical path D – 17%

Fast Complex Gates: Design Technique 3

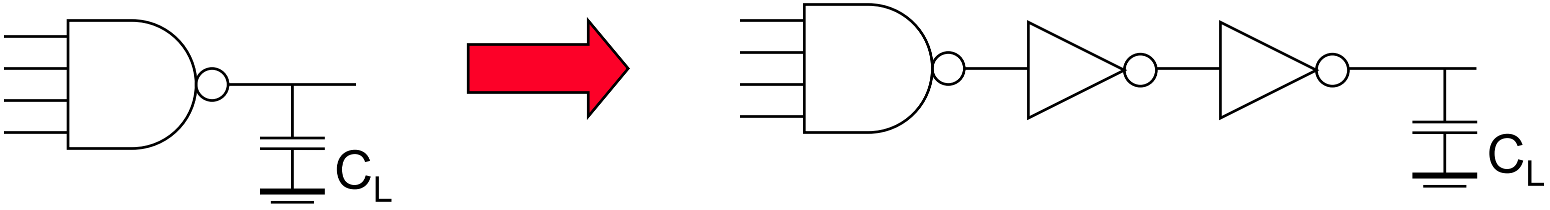
- Alternative logic structures

F = ABCDEFGH



Fast Complex Gates: Design Technique 4

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



- Real lesson is that optimizing the propagation delay of a gate in isolation is misguided.

Fast Networks: Design Technique 5 - Logical Effort

- The optimum fan-out for a chain of N inverters driving a load C_L is

$$f = \sqrt[N]{(C_L/C_{in})}$$

  - so, if we can, keep the fan-out per stage around 4.

- Can the same approach (logical effort) be used for any combinational circuit?
  - For a complex gate, we expand the inverter equation

$$t_p = t_{p0}\,(1 + C_{ext}/\ \gamma C_g) = t_{p0}\,(1 + f/\gamma)$$

to

$$\boxed{t_p = t_{p0}\,(p\ +\ g\ f/\gamma)}$$

- t_{p0} is the intrinsic delay of an inverter
- f is the effective fan-out (C_{ext}/C_g) – also called the electrical effort
- p is the ratio of the instrinsic (unloaded) delay of the complex gate and a simple inverter (a function of the gate topology and layout style)
- g is the logical effort

Intrinsic Delay Term, p

- The more involved the structure of the complex gate, the higher the intrinsic delay compared to an inverter

Gate Type	p
Inverter	1
n-input NAND	n
n-input NOR	n
n-way mux	2n
XOR, XNOR	$n\ 2^{n-1}$

Ignoring second order effects such as internal node capacitances

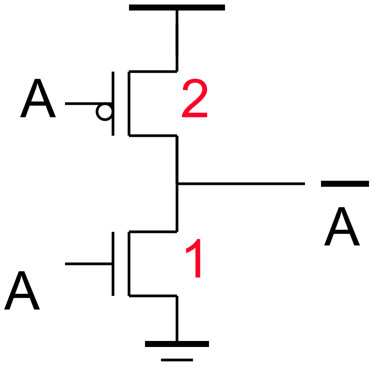
Logical Effort Term, g

- g represents the fact that, for a given load, complex gates have to work harder than an inverter to produce a similar (speed) response
  - the logical effort of a gate tells how much worse it is at producing an output current than an inverter (how much more input capacitance a gate presents to deliver it same output current)

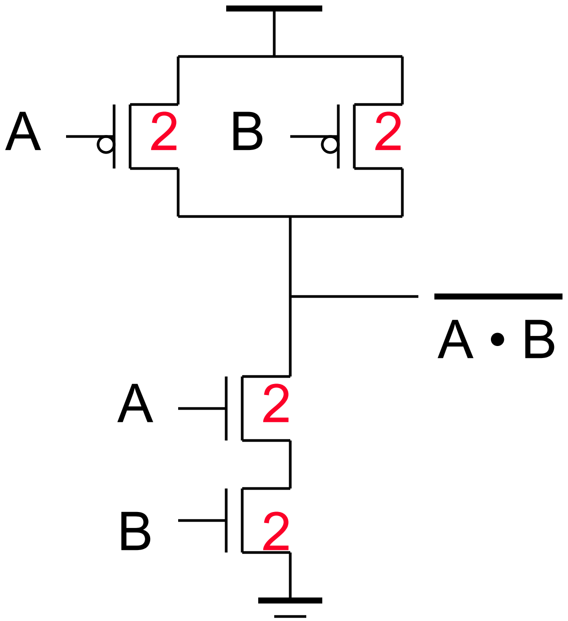
Gate Type	**g (for 1 to 4 input gates)**			
	1	**2**	**3**	**4**
Inverter	1			
NAND		4/3	5/3	(n+2)/3
NOR		5/3	7/3	(2n+1)/3
mux		2	2	2
XOR		4	12	

Example of Logical Effort

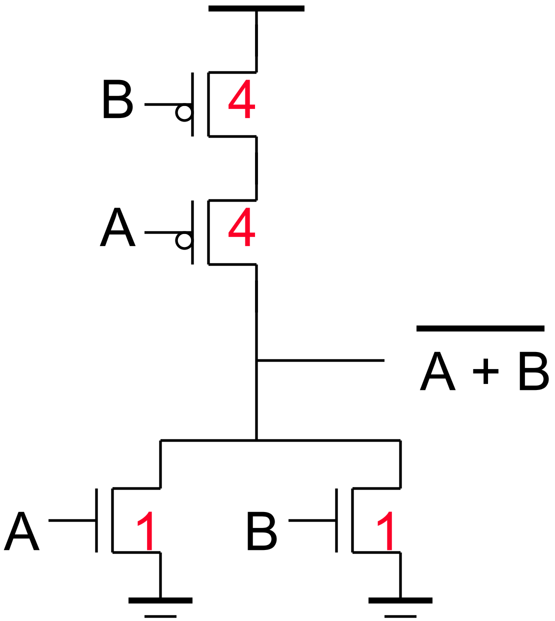
- Assuming a pmos/nmos ratio of 2, the input capacitance of a minimum-sized inverter is three times the gate capacitance of a minimum-sized nmos (C_{unit})



$C_{unit} = 3$

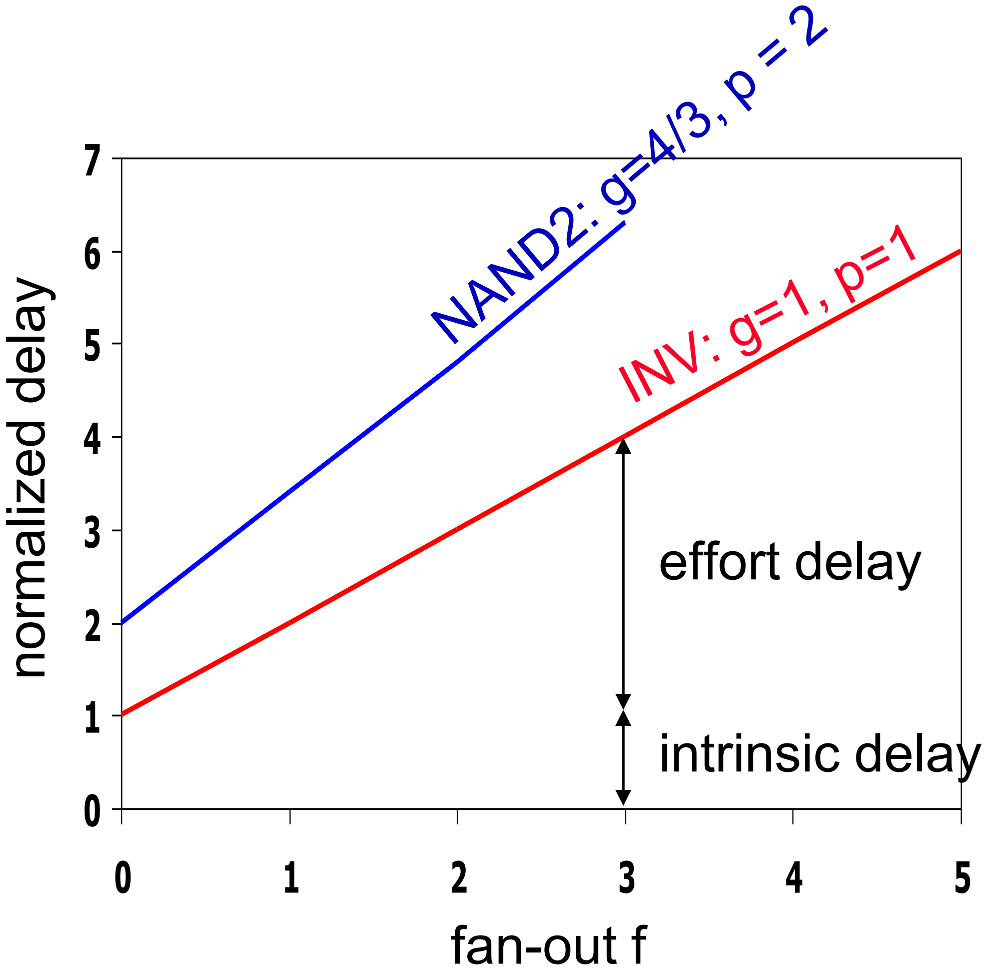


$C_{unit} = 4$



$C_{unit} = 5$

Delay as a Function of Fan-Out



- ❑ The slope of the line is the logical effort of the gate
- ❑ The y-axis intercept is the intrinsic delay
- ❑ Can adjust the delay by adjusting the effective fan-out (by sizing) or by choosing a gate with a different logical effort
- ❑ Gate effort: h = fg

Path Delay of Complex Logic Gate Network

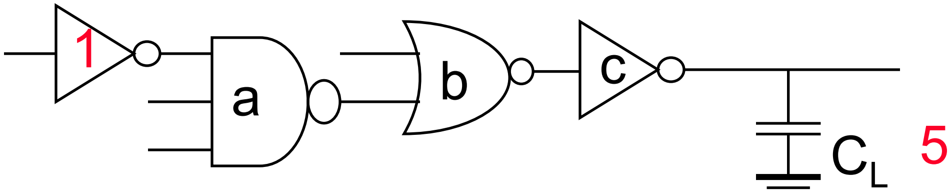
- Total path delay through a combinational logic block

$$t_p = \sum t_{p,j} = t_{p0} \sum(p_j + (f_j\, g_j)/\gamma\ )$$

- So, the minimum delay through the path determines that each stage should bear the same gate effort

$$f_1 g_1 = f_2 g_2 = . . . = f_N g_N$$

- Consider optimizing the delay through the logic network



how do we determine a, b, and c sizes?

Path Delay Equation Derivation

- The path logical effort, $G = \prod g_i$
- And the path effective fan-out (path electrical effort) is $F = C_L/g_1$
- The branching effort accounts for fan-out to other gates in the network

$$b = (C_{\text{on-path}} + C_{\text{off-path}})/C_{\text{on-path}}$$

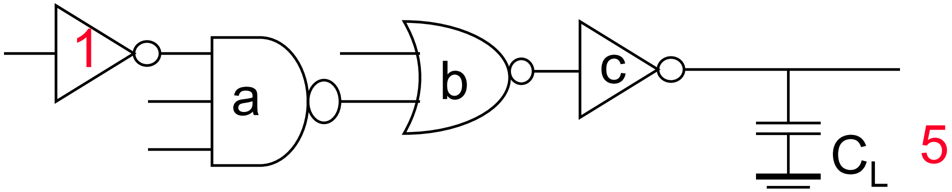
- The path branching effort is then $B = \prod b_i$
- And the total path effort is then $H = GFB$
- So, the minimum delay through the path is

$$D = t_{p0} \left(\sum p_j + (N \sqrt[N]{H}) / \gamma \right)$$

Path Delay of Complex Logic Gates, con't

- For gate i in the chain, its size is determined by

$$s_i = (g_1\ s_1)/g_i \prod_{j=1}^{i-1} (f_j/b_j)$$



- For this network
  - $F = C_L/C_{g1} = 5$
  - $G = 1 \times 5/3 \times 5/3 \times 1 = 25/9$
  - $B = 1$ (no branching)
  - $H = GFB = 125/9$, so the optimal stage effort is $\sqrt[4]{H} = 1.93$
    - Fan-out factors are $f_1=1.93$, $f_2=1.93 \times 3/5 = 1.16$, $f_3 = 1.16$, $f_4 = 1.93$
  - So the gate sizes are $a = f_1 g_1/g_2 = 1.16$, $b = f_1 f_2 g_1/g_3 = 1.34$ and $c = f_1 f_2 f_3 g_1/g_4 = 2.60$

Fast Complex Gates: Design Technique 6

- Reducing the voltage swing

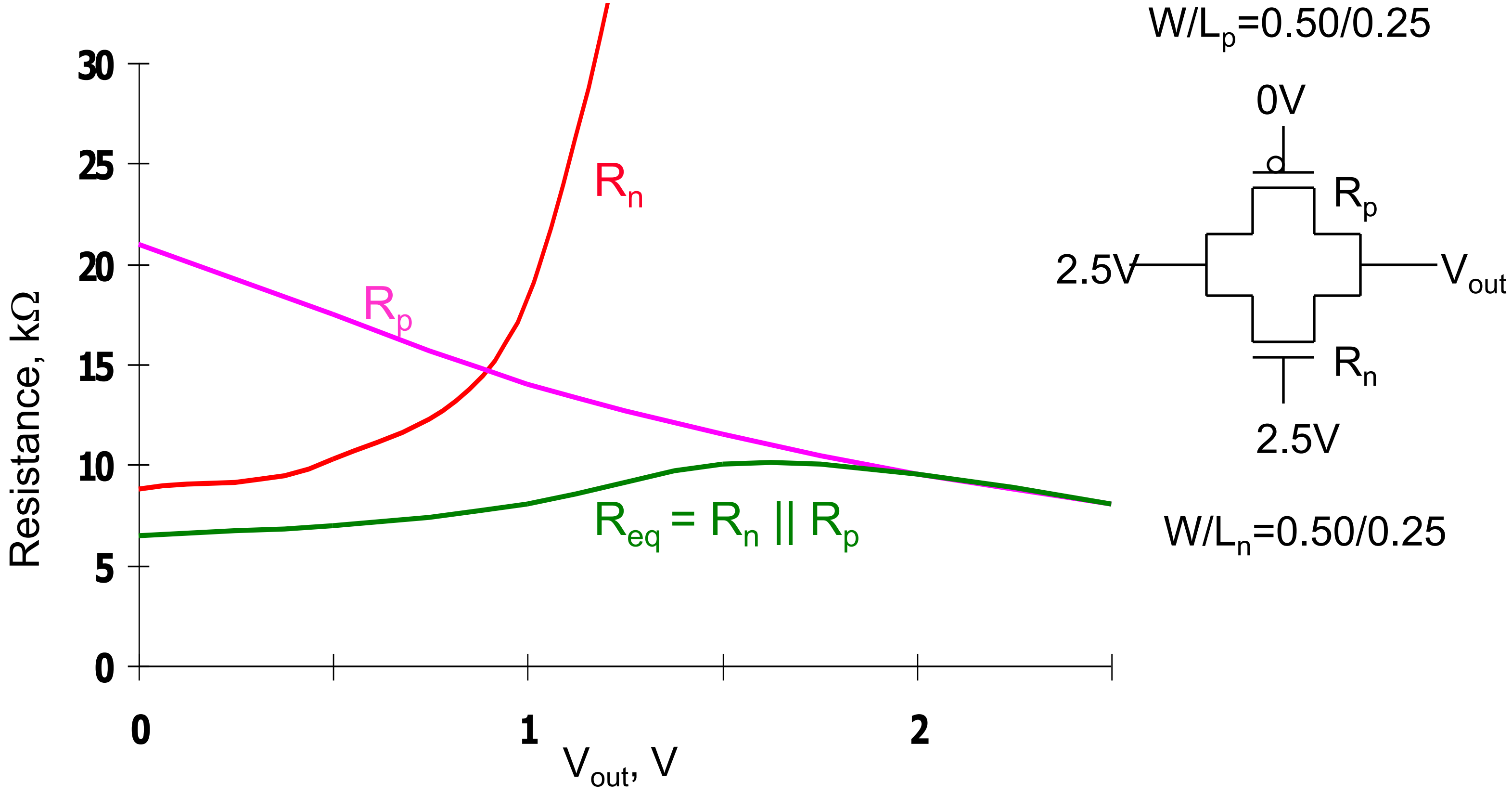
$$t_{pHL} = 0.69\ (3/4\ (C_L\ V_{DD})/\ I_{DSATn}\ )$$

$$= 0.69\ (3/4\ (C_L\ V_{swing})/\ I_{DSATn}\ )$$

  - linear reduction in delay
  - also reduces power consumption
  - requires use of “sense amplifiers” on the receiving end to restore the signal level (will look at their design when covering memory design)

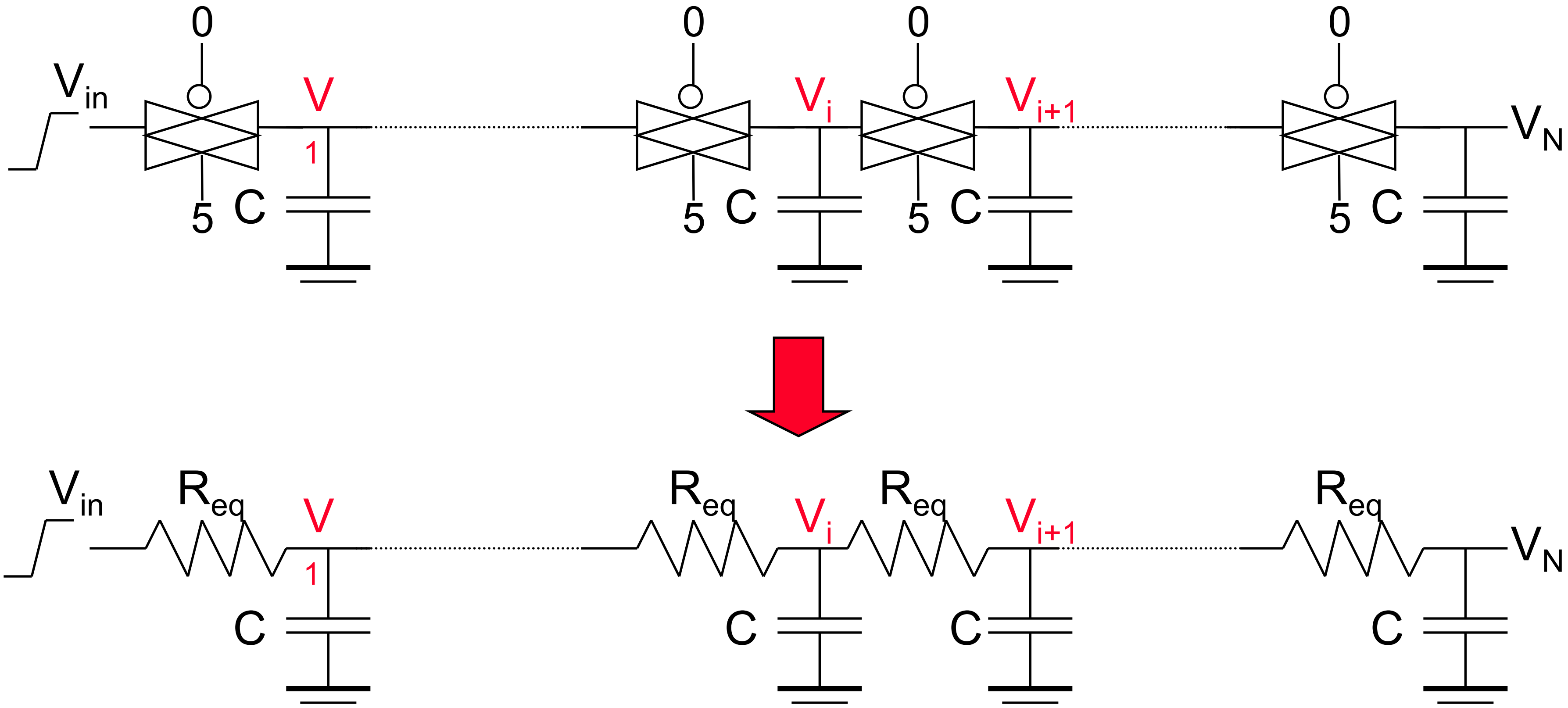
TG Logic Performance

- Effective resistance of the TG is modeled as a parallel connection of R_p (= $(V_{DD} – V_{out})/(-I_{Dp})$) and R_n (=$V_{DD} – V_{out})/I_{Dn}$)



- So, the assumption that the TG switch has a constant resistive value, R_{eq}, is acceptable

Delay of a TG Chain

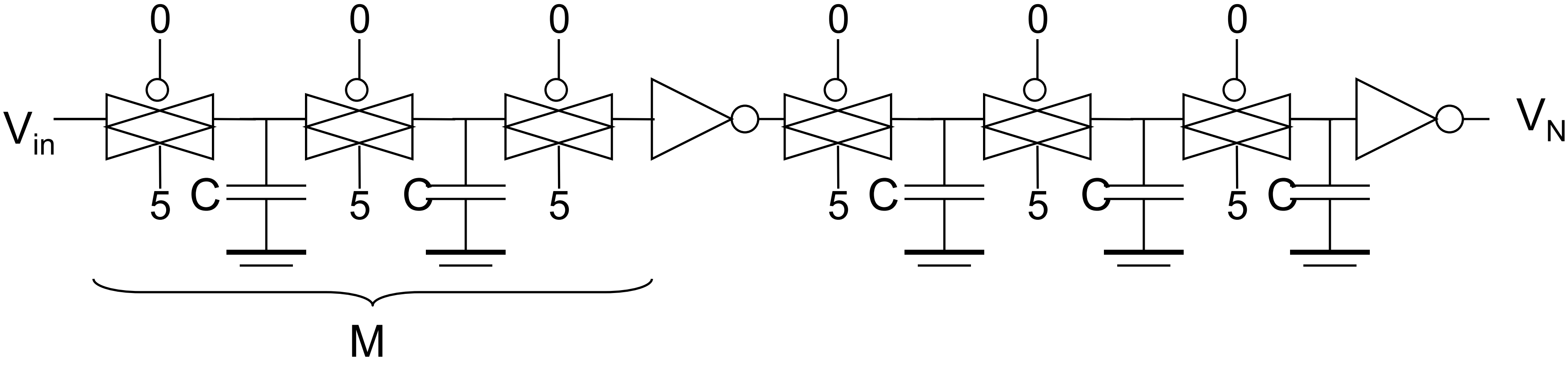


- Delay of the RC chain (N TG's in series) is

$$t_p(V_n) = 0.69 \sum_{k=1}^{N} kCR_{eq} = 0.69\ CR_{eq}\ (N(N+1))/2 \approx 0.35\ CR_{eq}N^2$$

TG Delay Optimization

- Can speed it up by inserting buffers every M switches



- Delay of buffered chain (M TG's between buffer)

$$t_p = 0.69 \lfloor N/M\ CR_{eq}\ (M(M+1))/2 \rfloor + (N/M - 1)\ t_{pbuf}$$

$$M_{opt} = 1.7 \sqrt{}\ (t_{pbuf}/CR_{eq}) \approx 3 \text{ or } 4$$

Next Lecture and Reminders

- Next lecture
  - Designing energy efficient logic
    - Reading assignment – Rabaey, et al, 5.5 & 6.2.1
- Reminders
  - HW3 due Oct 10th (hand in to TA)
  - Class cancelled on Oct 10th as make up for evening midterm
  - I will be out of town Oct 10th through Oct 15th and Oct 18th through Oct 23rd, so office hours during those periods are cancelled
  - We will have a guest lecturer on Oct 22nd
  - Evening midterm exam scheduled
    - Wednesday, October 16th from 8:15 to 10:15pm in 260 Willard
    - Only one midterm conflict filed for so far