Logical Effort

(Sizing, Staging, Fan out)

Design for Performance

- Reduce C_L
  - internal diffusion capacitance of the gate itself
    - keep the drain diffusion as small as possible
  - 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! – when the intrinsic capacitance dominates the extrinsic load

- Increase V_{DD}
  - can trade-off energy for performance
  - increasing V_{DD} above a certain level yields only very minimal improvements
  - reliability concerns enforce a firm upper bound on V_{DD}

NMOS/PMOS Ratio

- So far have sized the PMOS and NMOS so that the R_{eq}'s match (ratio of 3 to 3.5)
  - symmetrical VTC
  - equal high-to-low and low-to-high propagation delays

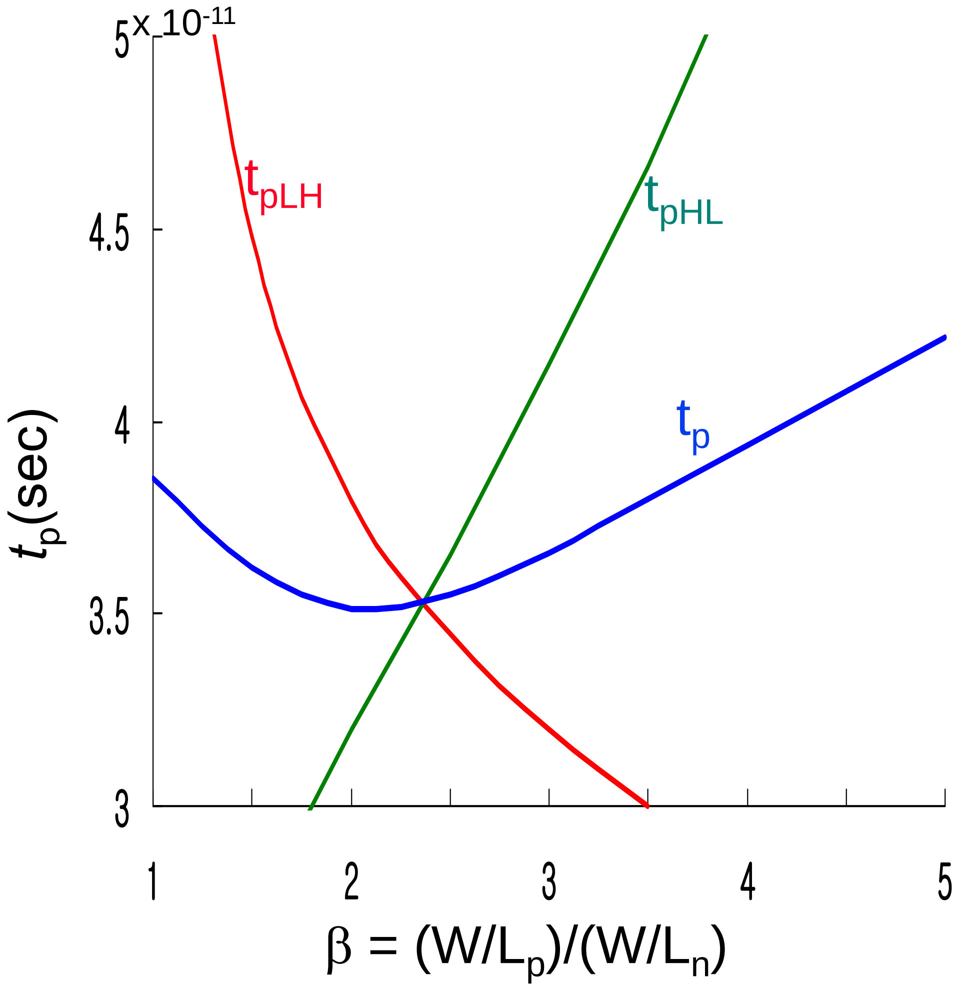
- If speed is the only concern, reduce the width of the PMOS device!
  - widening the PMOS degrades the t_{pHL} due to larger parasitic capacitance

$$\beta = (W/L_p)/(W/L_n)$$

$r = R_{eqp}/R_{eqn}$ (resistance ratio of identically-sized PMOS and NMOS)

$\beta_{opt} = \sqrt{r}$ when wiring capacitance is negligible

PMOS/NMOS Ratio Effects



β of 2.4 (= 31 kΩ/13 kΩ) gives symmetrical response

β of 1.6 to 1.9 gives optimal performance

Device Sizing for Performance

- Divide capacitive load, C_L, into
  - C_{int} : intrinsic - diffusion and Miller effect
  - C_{ext} : extrinsic - wiring and fanout

$$t_p = 0.69\, R_{eq}\, C_{int}\, (1 + C_{ext}/C_{int}) = t_{p0}\, (1 + C_{ext}/C_{int})$$

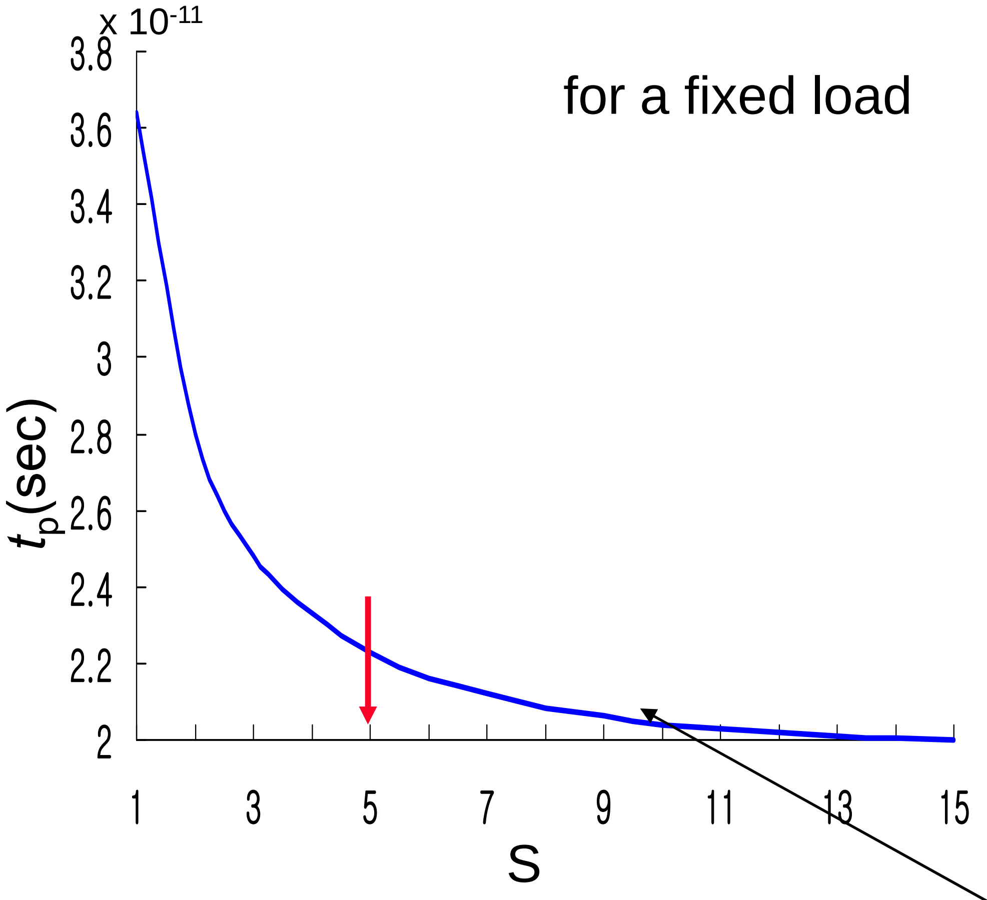
  - where $t_{p0} = 0.69\, R_{eq}\, C_{int}$ is the intrinsic (unloaded) delay of the gate

- Widening both PMOS and NMOS by a factor S reduces R_{eq} by an identical factor ($R_{eq} = R_{ref}/S$), but raises the intrinsic capacitance by the same factor ($C_{int} = SC_{iref}$)

$$t_p = 0.69\, R_{ref}\, C_{iref}\, (1 + C_{ext}/(SC_{iref})) = t_{p0}(1 + C_{ext}/(SC_{iref}))$$

  - t_{p0} is independent of the sizing of the gate; *with no load the drive of the gate is totally offset by the increased capacitance*
  - any S sufficiently larger than (C_{ext}/C_{int}) yields the best performance gains with least area impact

Sizing Impacts on Delay



for a fixed load

The majority of the improvement is already obtained for S = 5. Sizing factors larger than 10 barely yield any extra gain (and cost significantly more area).

self-loading effect (intrinsic capacitance dominates)

Impact of Fanout on Delay

- Extrinsic capacitance, C_{ext}, is a function of the fanout of the gate - the larger the fanout, the larger the external load.

- First determine the input loading effect of the inverter. Both C_g and C_{int} are proportional to the gate sizing, so $C_{int} = \gamma C_g$ is independent of gate sizing and

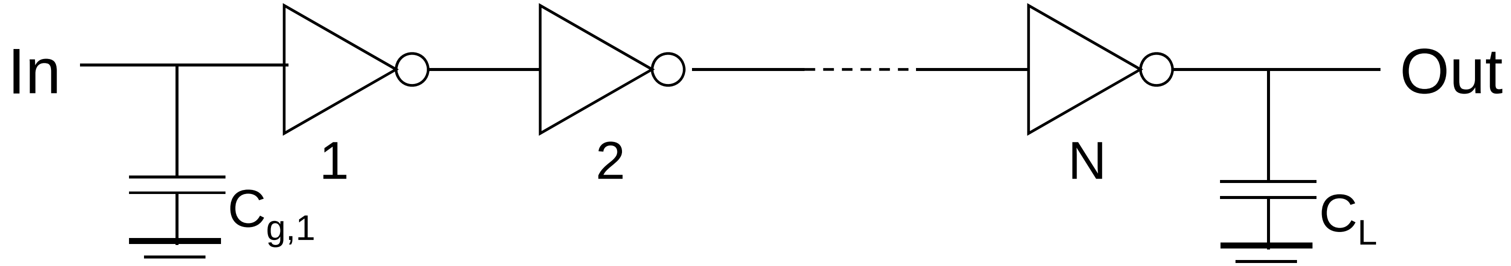
$$t_p = t_{p0}\,(1 + C_{ext}/\ \gamma C_g) = t_{p0}\,(1 + f/\gamma)$$

i.e., the delay of an inverter is a function of the ratio between its external load capacitance and its input gate capacitance: the effective fan-out f

$$f = C_{ext}/C_g$$

Inverter Chain

- Real goal is to minimize the delay through an inverter chain



the delay of the j-th inverter stage is

$$t_{p,j} = t_{p0}\,(1 + C_{g,j+1}/(\gamma C_{g,j})) = t_{p0}(1 + f_j/\gamma)$$

and $\quad t_p = t_{p1} + t_{p2} + \ldots + t_{pN}$

so $\quad t_p = \sum t_{p,j} = t_{p0} \sum (1 + C_{g,j+1}/(\gamma C_{g,j}))$

- If C_L is given
  - How should the inverters be sized?
  - How many stages are needed to minimize the delay?

Sizing the Inverters in the Chain

- The optimum size of each inverter is the geometric mean of its neighbors – meaning that if each inverter is sized up by the same factor f wrt the preceding gate, it will have the same effective fan-out and the same delay

$$f = \sqrt[N]{C_L/C_{g,1}} = \sqrt[N]{F}$$

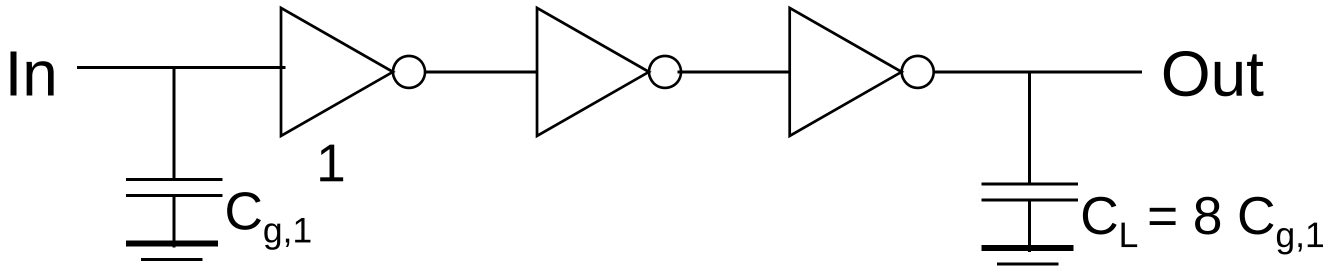
where F represents the overall effective fan-out of the circuit ($F = C_L/C_{g,1}$)

and the minimum delay through the inverter chain is

$$t_p = N\ t_{p0}\ (1 + (\sqrt[N]{F}\,)\ /\ \gamma)$$

- The relationship between t_p and F is linear for one inverter, square root for two, etc.

Example of Inverter Chain Sizing

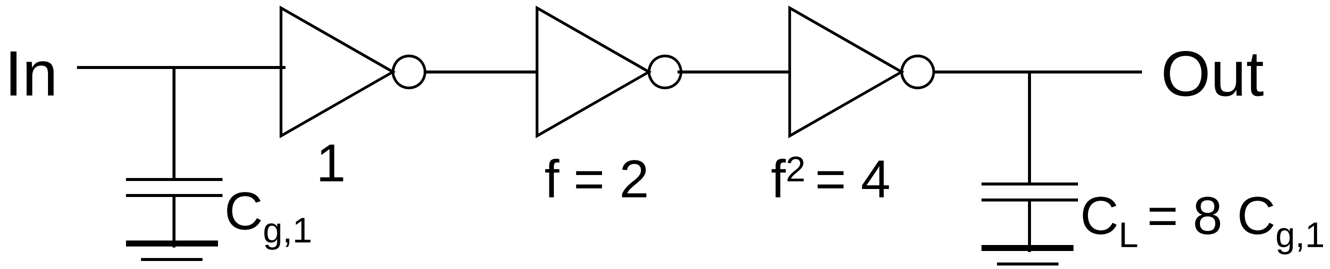


- $C_L/C_{g,1}$ has to be evenly distributed over N = 3 inverters

$$C_L/C_{g,1} = 8/1$$

$$f\ =$$

Example of Inverter Chain Sizing



- $C_L/C_{g,1}$ has to be evenly distributed over N = 3 inverters

$$C_L/C_{g,1} = 8/1$$

$$f = \sqrt[3]{8} = 2$$

Determining N: Optimal Number of Inverters

- What is the optimal value for N given F (=f^N) ?
  - if the number of stages is too large, the intrinsic delay of the stages becomes dominate
  - if the number of stages is too small, the effective fan-out of each stage becomes dominate

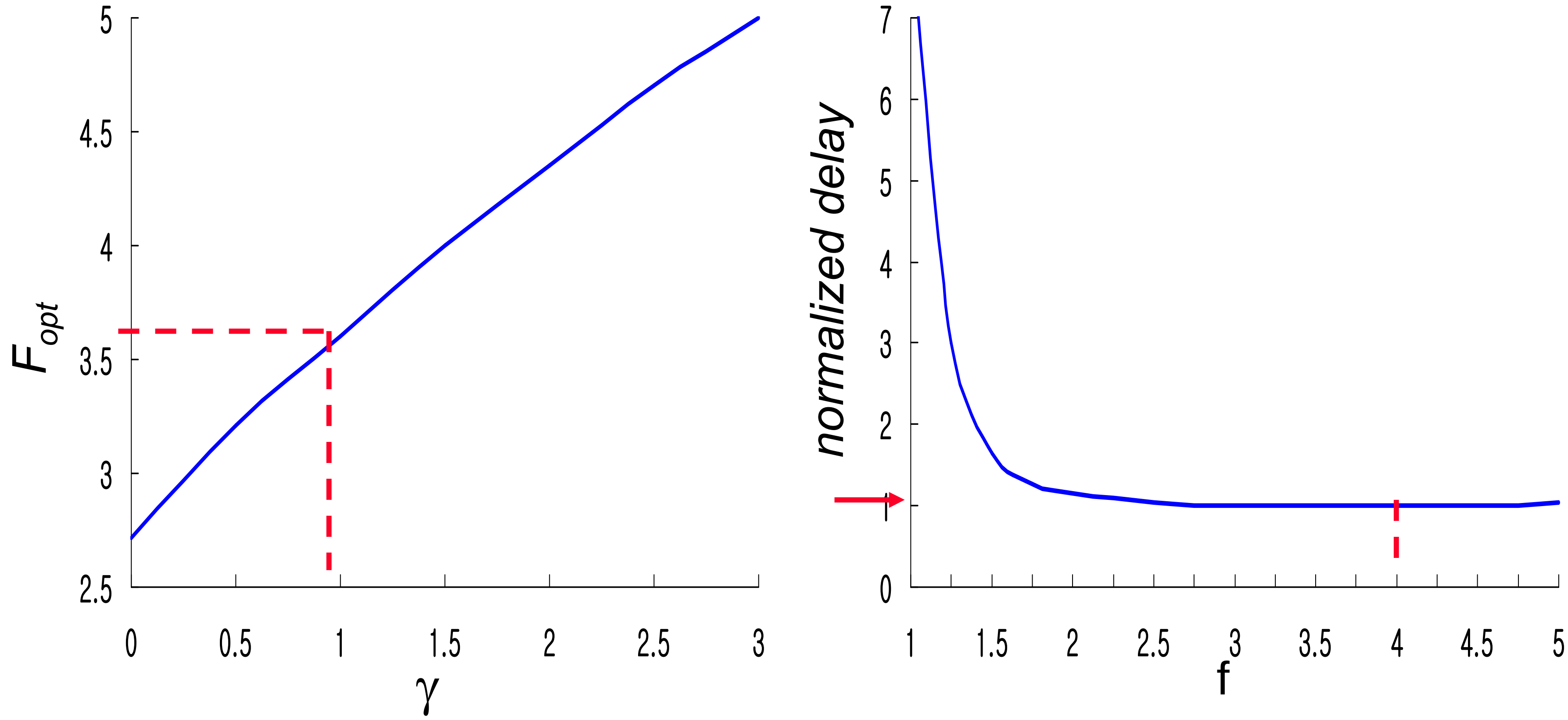
- The optimum N is found by differentiating the minimum delay expression divided by the number of stages and setting the result to 0, giving

$$\gamma + \sqrt[N]{F} - (\sqrt[N]{F} \; \mathrm{ln}F)/N = 0$$

- For γ = 0 (ignoring self-loading) N = ln (F) and the effective-fan out becomes f = e = 2.71828

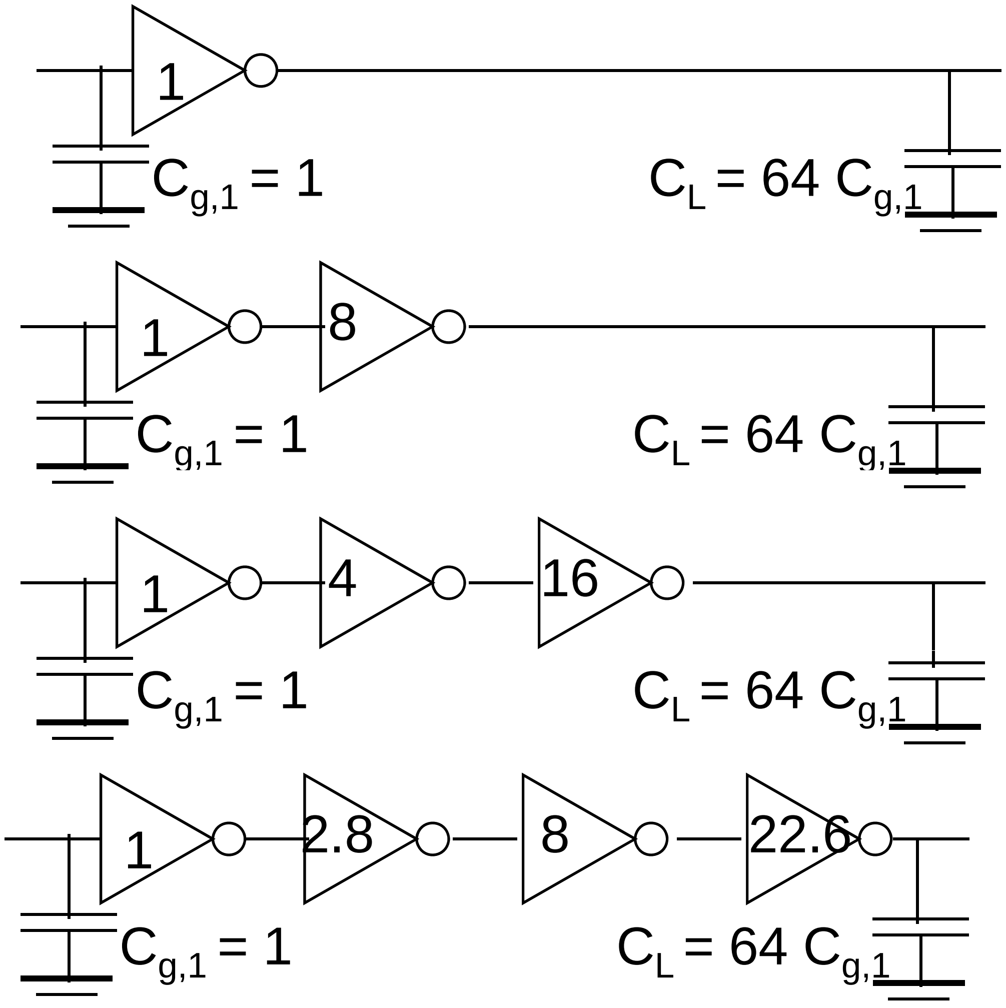
- For γ = 1 (the typical case) the optimum effective fan-out (tapering factor) turns out to be close to 3.6

Optimum Effective Fan-Out



- ❑ Choosing f larger than optimum has little effect on delay and reduces the number of stages (and area).
  - Common practice to use f = 4 (for γ = 1)
  - But too many stages has a substantial negative impact on delay

Example of Inverter (Buffer) Staging



N	f	t_p
1	64	65
2	8	18
3	4	15
4	2.8	15.3

Impact of Buffer Staging for Large C_L

F (γ = 1)	Unbuffered	Two Stage Chain	Opt. Inverter Chain
10	11	8.3	8.3
100	101	22	16.5
1,000	1001	65	24.8
10,000	10,001	202	33.1

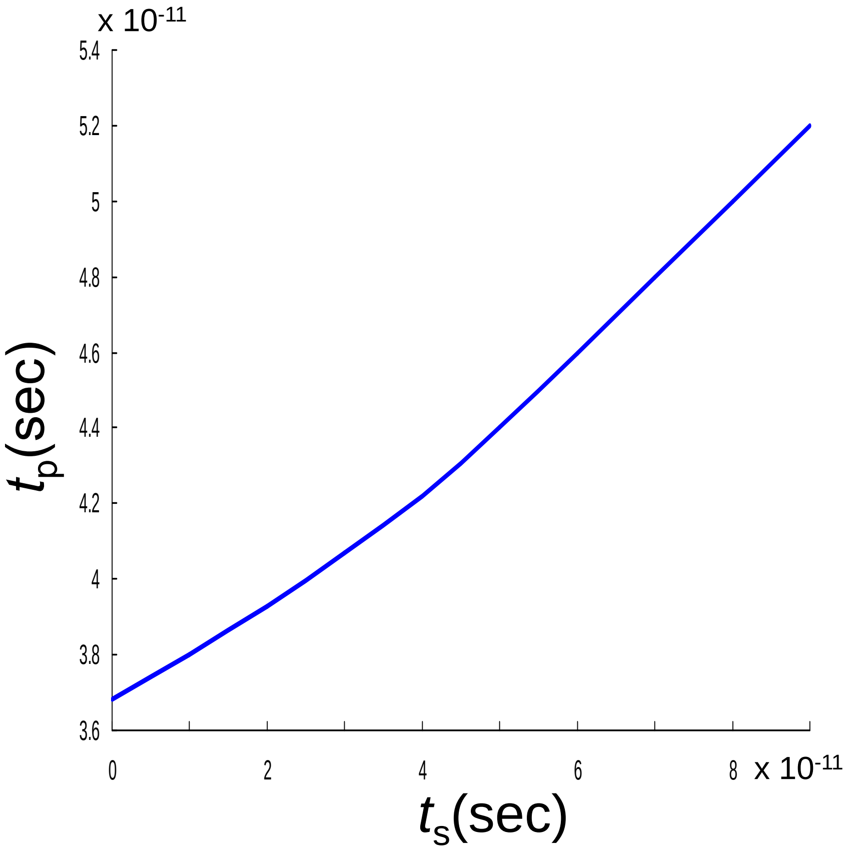
- Impressive speed-ups with optimized cascaded inverter chain for very large capacitive loads.

Input Signal Rise/Fall Time

- In reality, the input signal changes gradually (and both PMOS and NMOS conduct for a brief time). This affects the current available for charging/discharging C_L and impacts propagation delay.

- t_p increases linearly with increasing input slope, t_s, once $t_s > t_p$

- t_s is due to the limited driving capability of the preceding gate



for a minimum-size inverter
with a fan-out of a single gate

Design Challenge

- A gate is never designed in isolation: its performance is affected by both the fan-out and the driving strength of the gate(s) feeding its inputs.

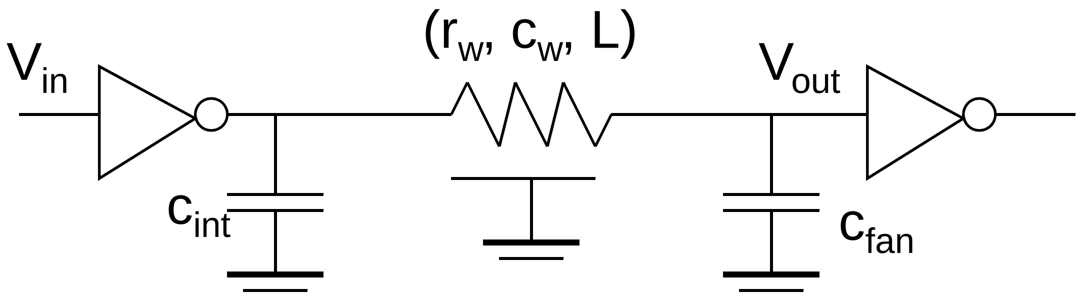
$$t^i_p = t^i_{step} + \eta \, t^{i-1}_{step} \qquad (\eta \approx 0.25)$$

- Keep signal rise times smaller than or equal to the gate propagation delays.
  - good for performance
  - good for power consumption

- Keeping rise and fall times of the signals small and of approximately equal values is one of the major challenges in high-performance designs - slope engineering.

Delay with Long Interconnects

- When gates are farther apart, wire capacitance and resistance can no longer be ignored.



$$t_p = 0.69R_{dr}C_{int} + (0.69R_{dr}+0.38R_w)C_w + 0.69(R_{dr}+R_w)C_{fan}$$

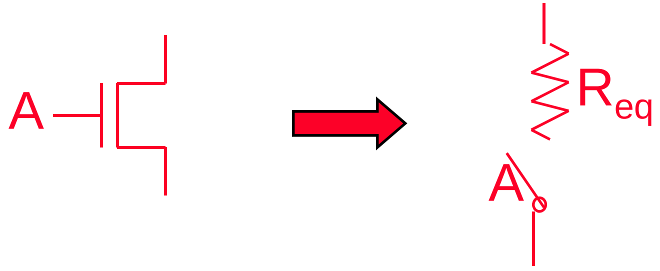
where $R_{dr} = (R_{eqn} + R_{eqp})/2$

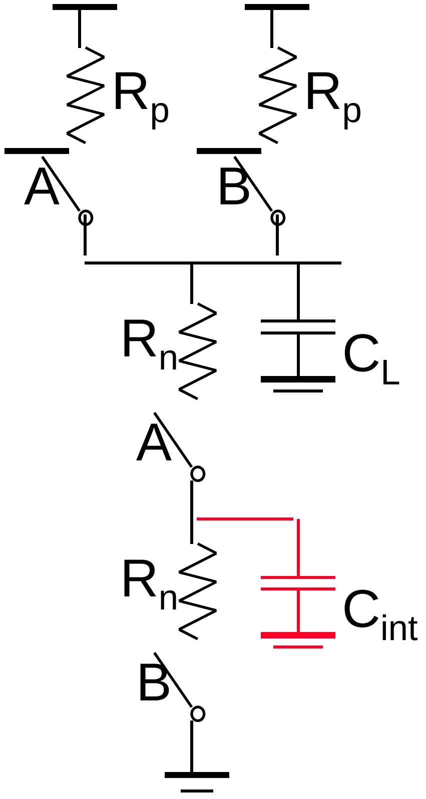
$$= 0.69R_{dr}(C_{int}+C_{fan}) + 0.69(R_{dr}c_w+r_wC_{fan})L + 0.38r_wc_wL^2$$

- Wire delay rapidly becomes the dominate factor (due to the quadratic term) in the delay budget for longer wires.

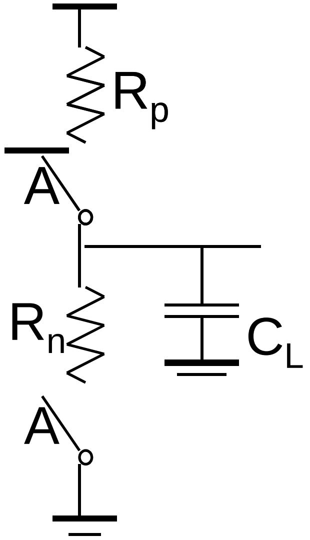
Rabaey 5.4.2

Switch Delay Model

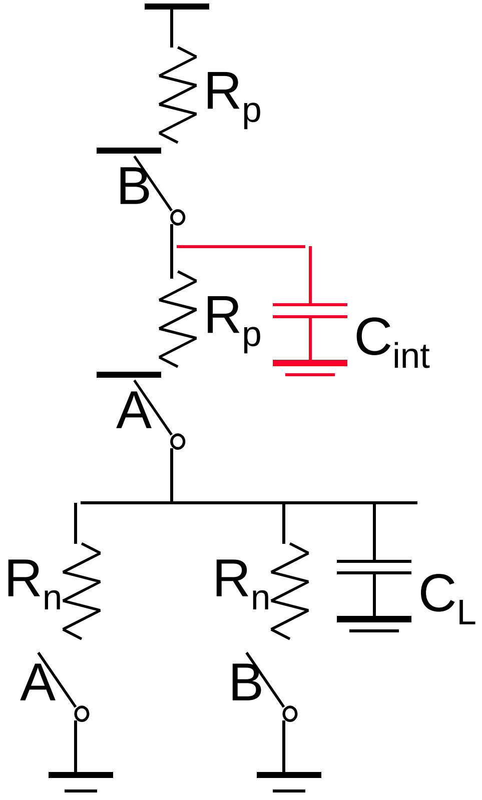




NAND

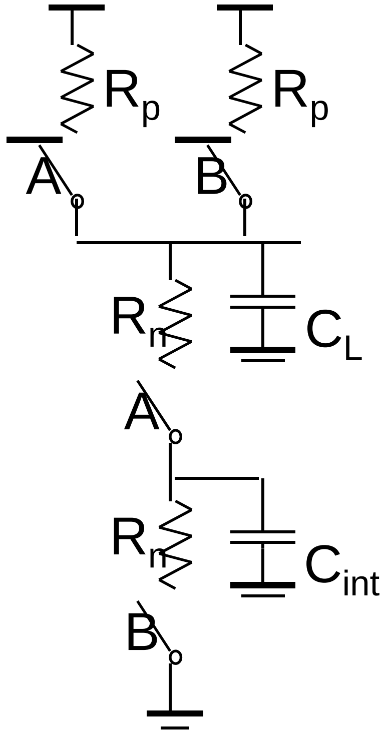


INVERTER



NOR

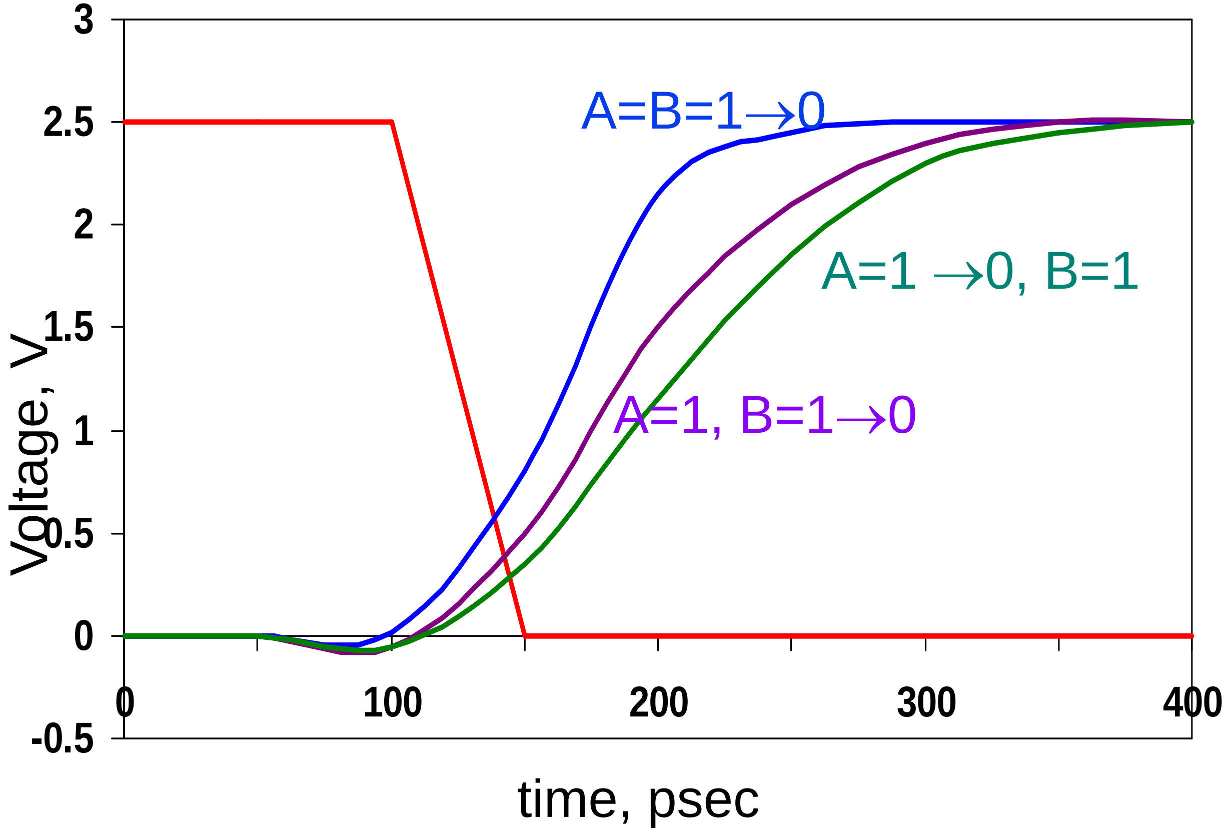
Input Pattern Effects on Delay



- ❑ 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

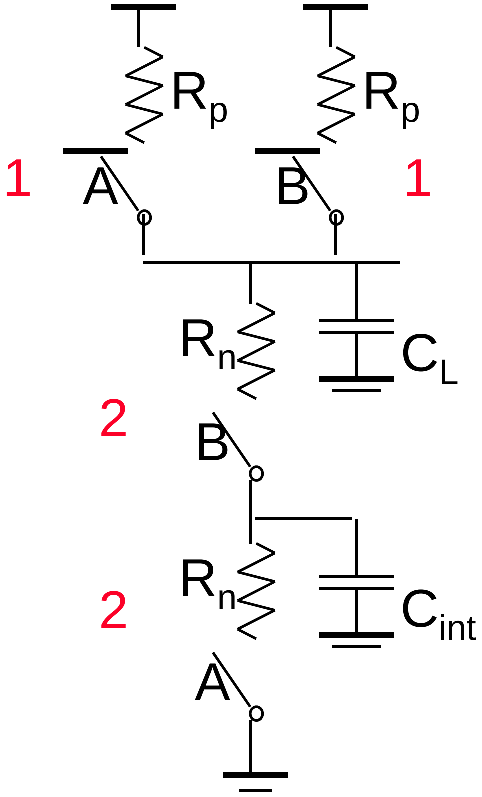
Delay Dependence on Input Patterns

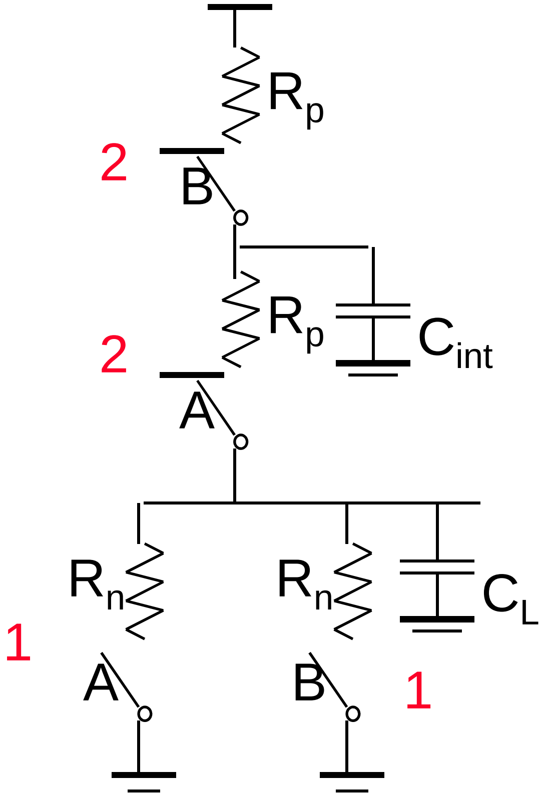
2-input NAND with
NMOS = 0.5μm/0.25 μm
PMOS = 0.75μm/0.25 μm
C_L = 10 fF



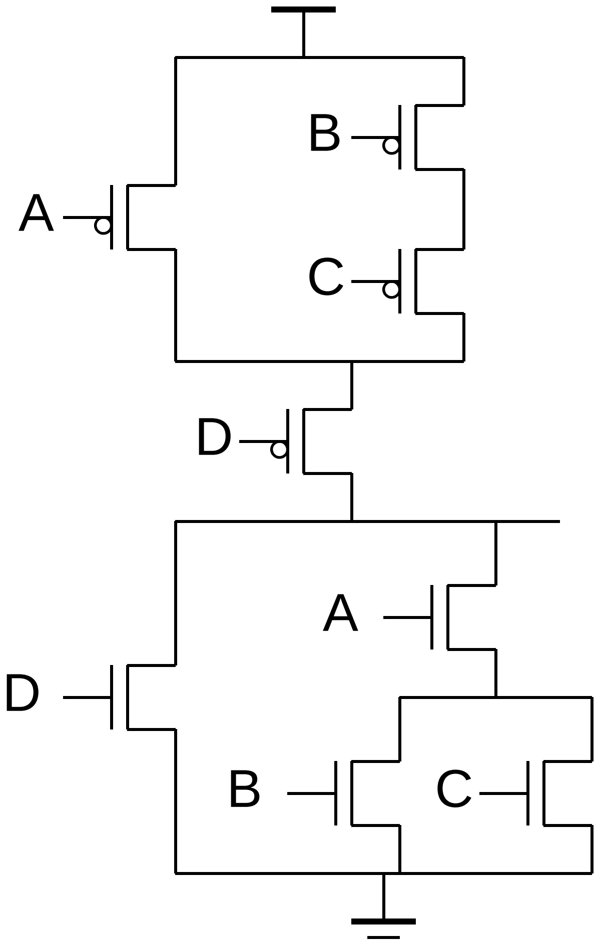
Input Data Pattern	Delay (psec)
A=B=0→1	69
A=1, B=0→1	62
A= 0→1, B=1	50
A=B=1→0	35
A=1, B=1→0	76
A= 1→0, B=1	57

Transistor Sizing



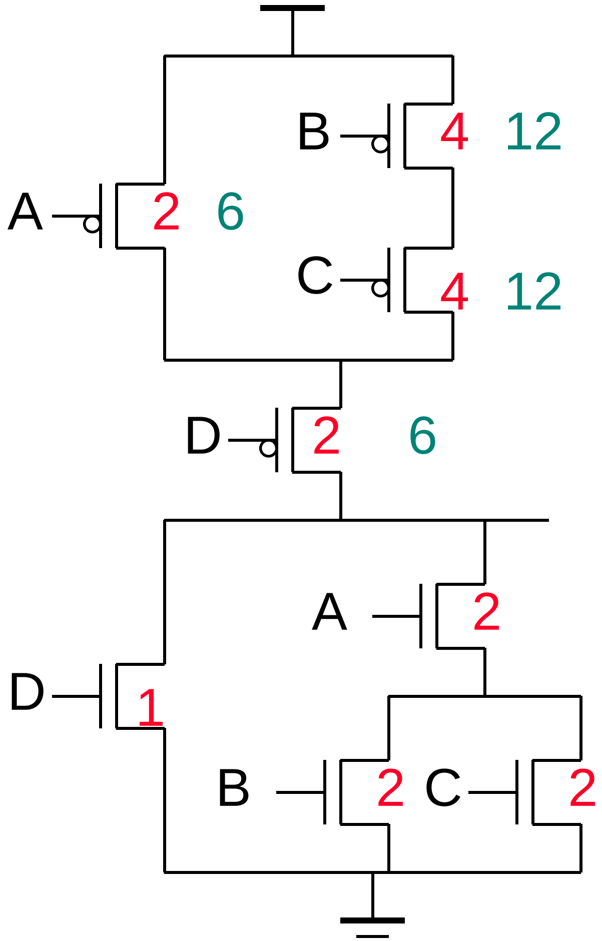


Transistor Sizing a Complex CMOS Gate



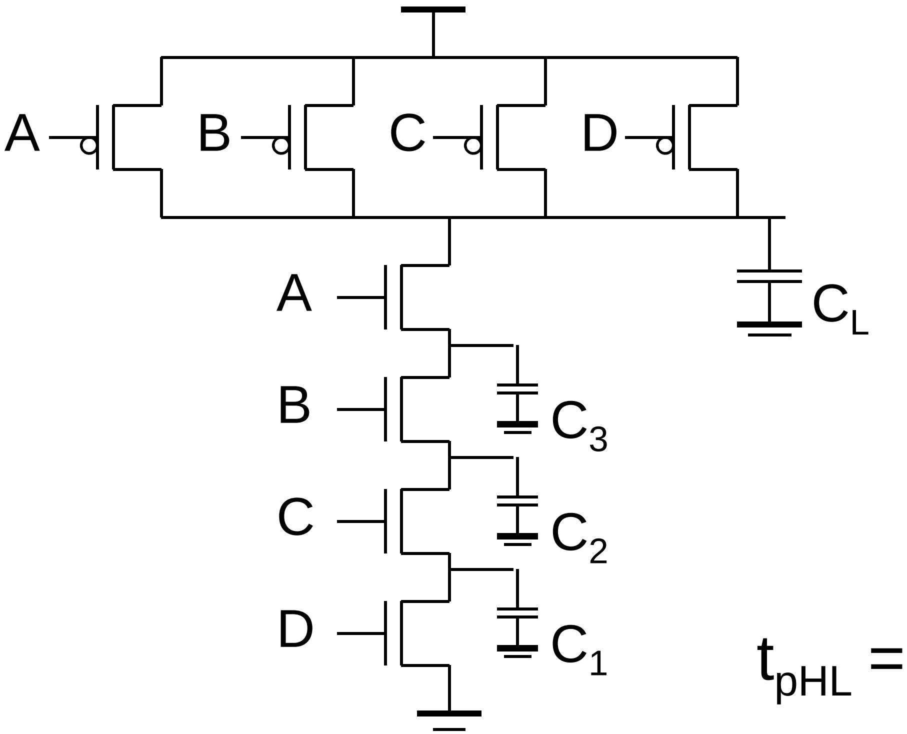
OUT = !(D + A • (B + C))

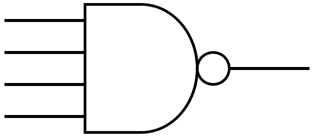
Transistor Sizing a Complex CMOS Gate



OUT = !(D + A • (B + C))

Fan-In Considerations



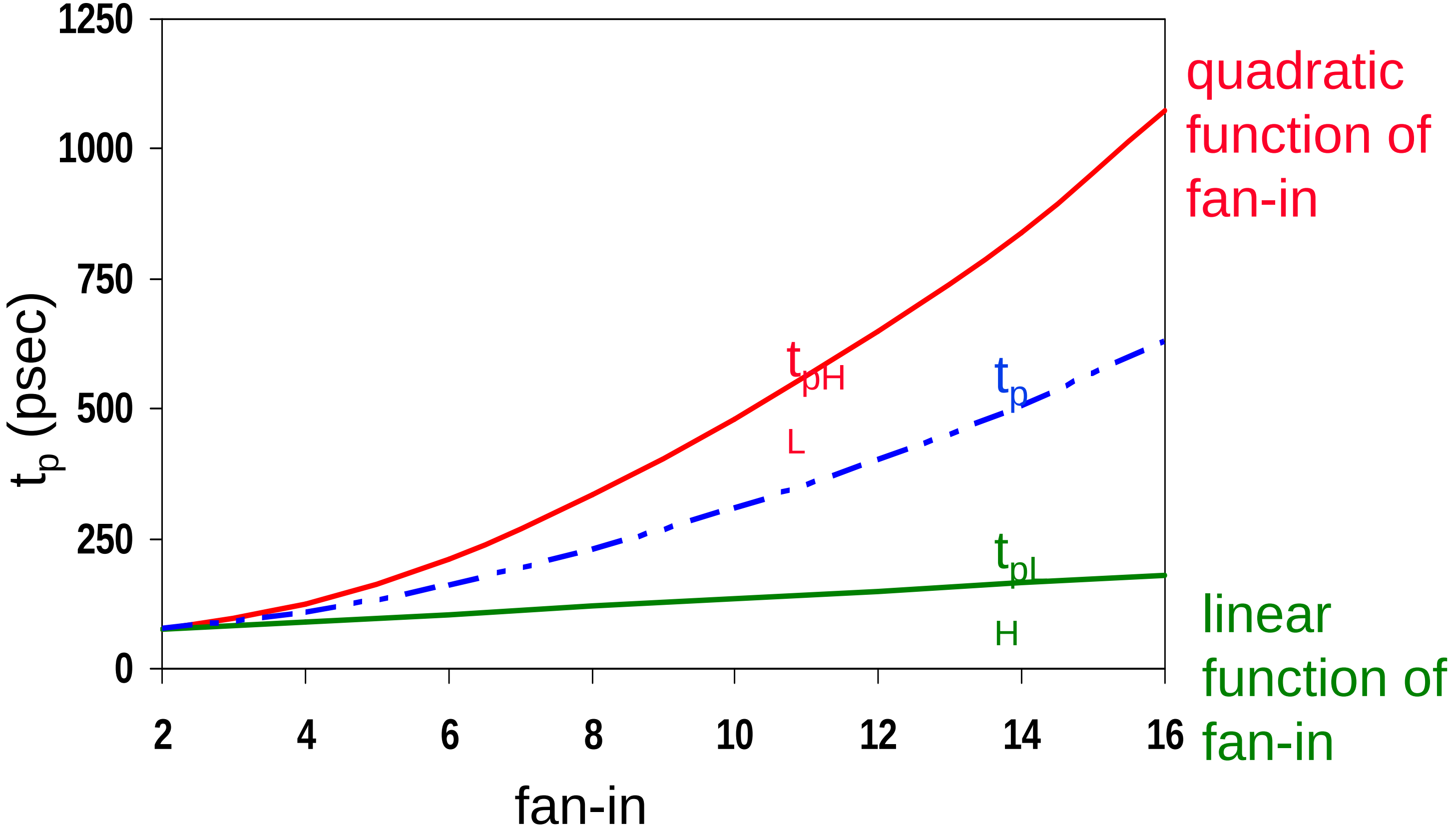


Distributed RC model
(Elmore delay)

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

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

t_p as a Function of Fan-In

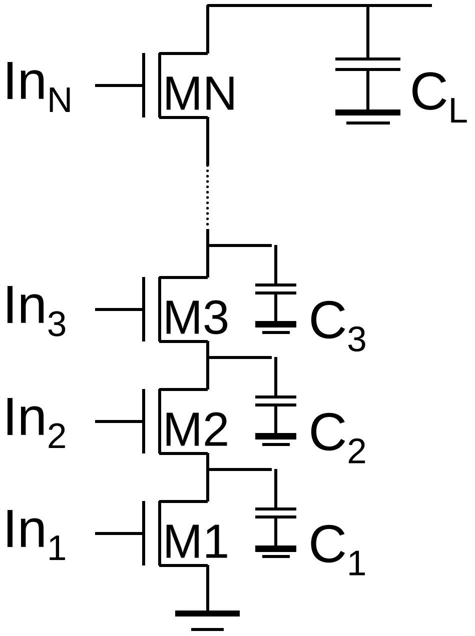


- Gates with a fan-in greater than 4 should be avoided.

Fast Complex Gates: Design Technique 1

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

- Progressive sizing



Distributed RC line

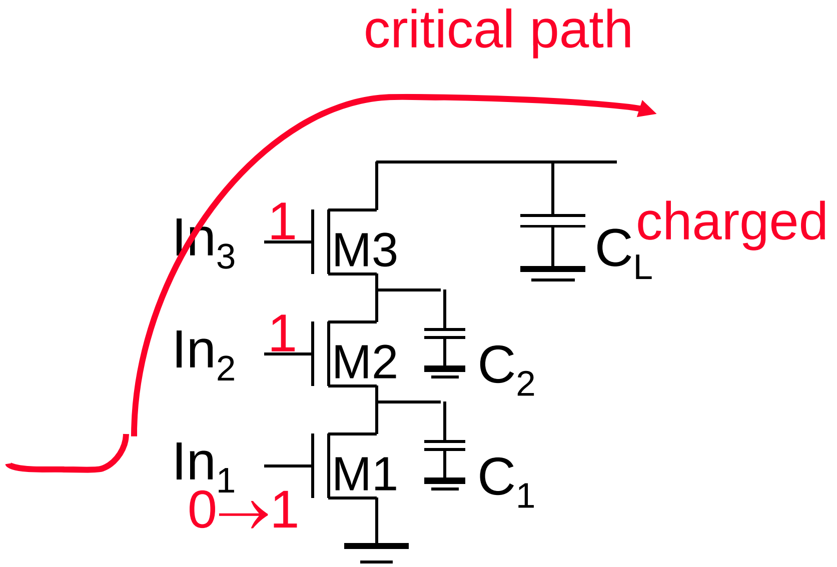
M1 > M2 > M3 > … > MN

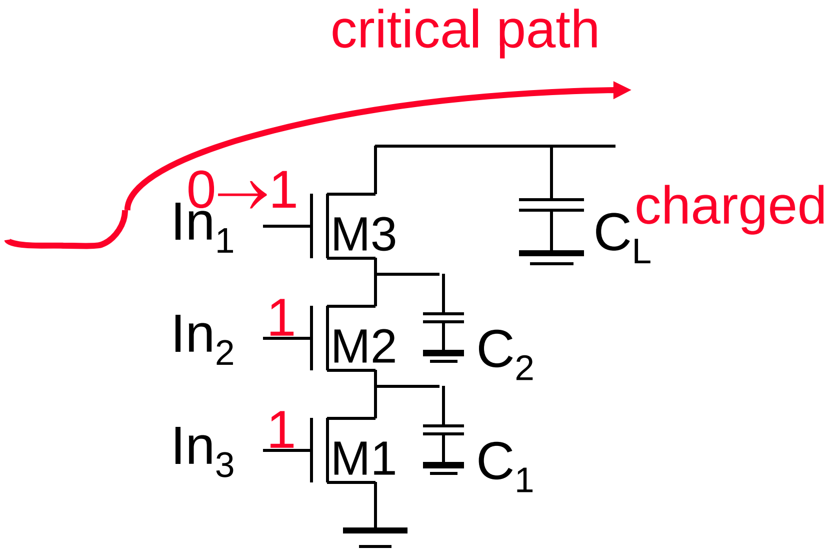
(the fet closest to the output should be the smallest)

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

Fast Complex Gates: Design Technique 2

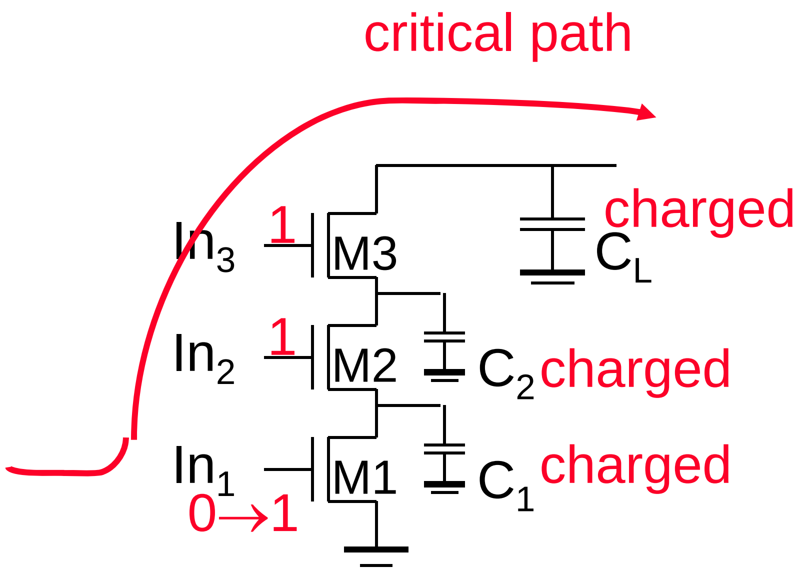
- ❑ Input re-ordering
  - when not all inputs arrive at the same time



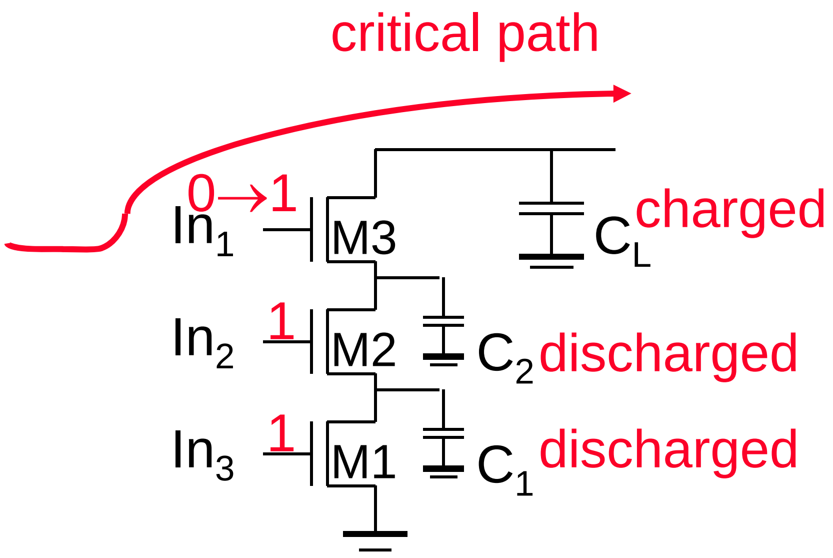


Fast Complex Gates: Design Technique 2

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

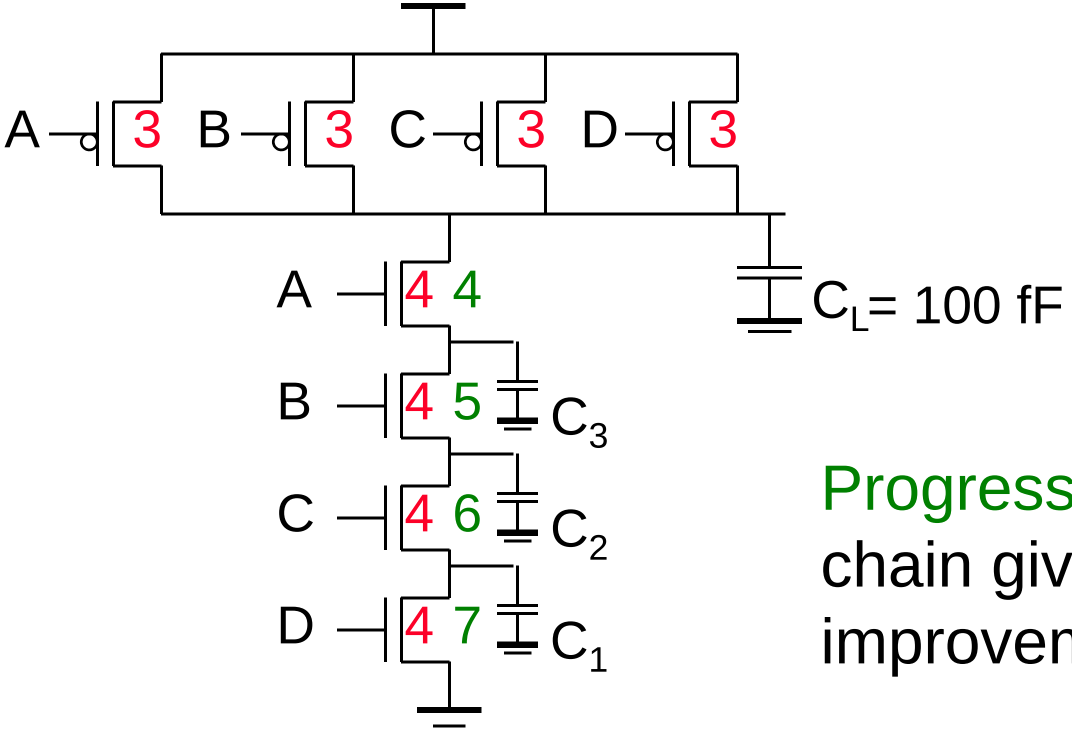


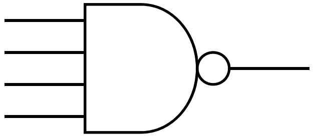
delay determined by time to discharge C_L, C_1 and C_2



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%

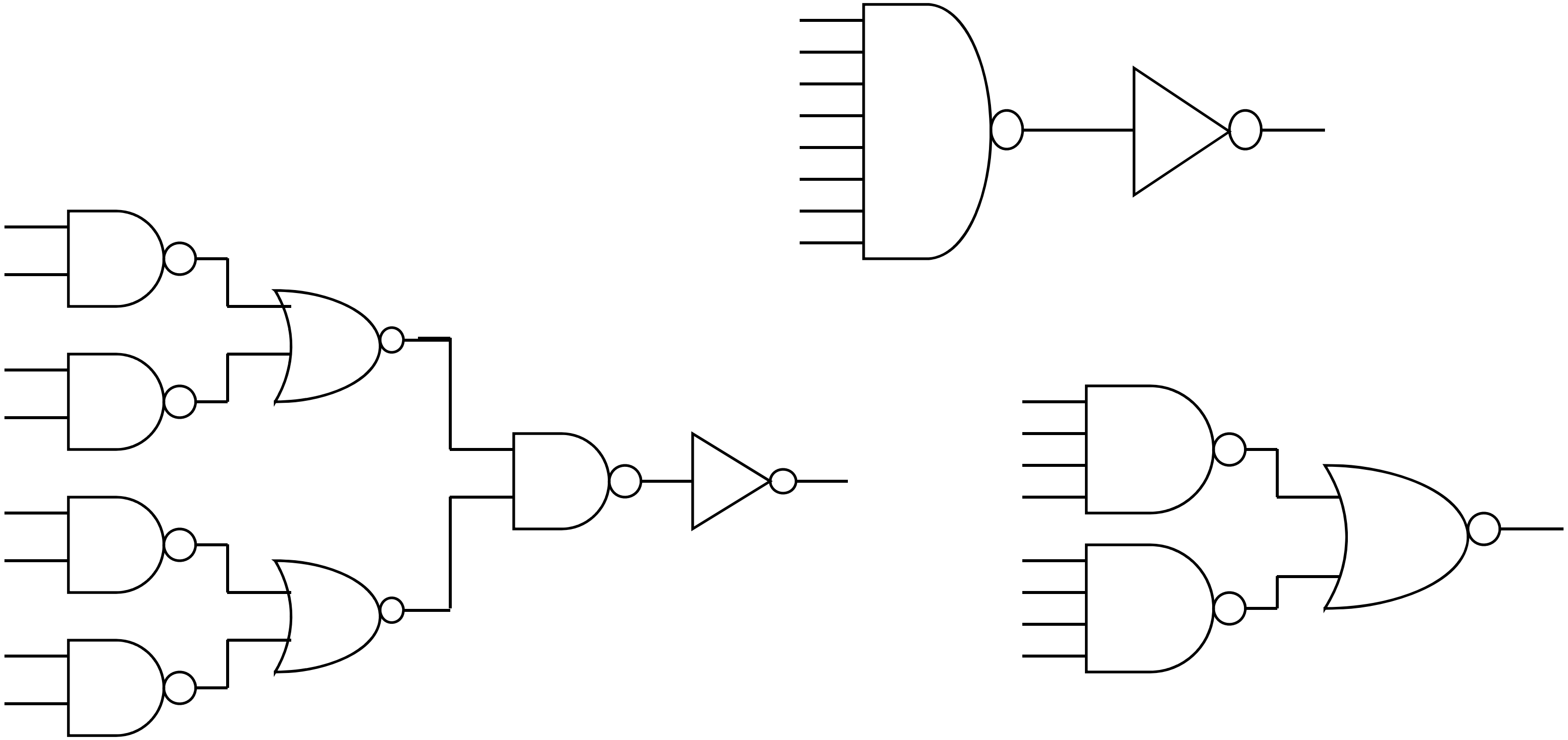
critical path A – 23%

critical path D – 17%

Fast Complex Gates: Design Technique 3

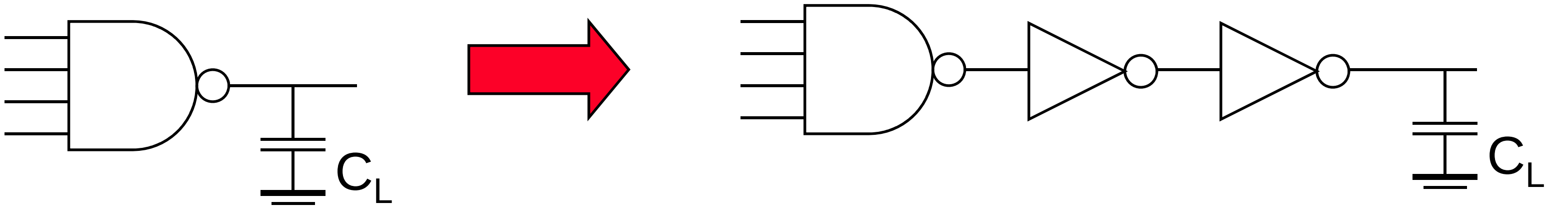
- Alternative logic structures

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