Digital Integrated Circuit Design
CMPE 530/630

Delay Models

Dr. Cory Merkel

Learning Objectives

- Understand the tradeoffs between different methods of estimating delay
- Be able to estimate the best and worst-case delays of CMOS circuits using RC and linear delay models
- Understand the impact of physical layout on delay of CMOS gates

Some Definitions

t_{pdf} - Propagation (worst-case) delay when output is falling

t_{pdr} - Propagation (worst-case) delay when output is rising

t_{pcf} - Contamination (best-case) delay when output is falling

t_{pcr} - Contamination (best-case) delay when output is rising

$t_{pd} = (t_{pdf} + t_{pdr})/2$ – Avg. prop. delay

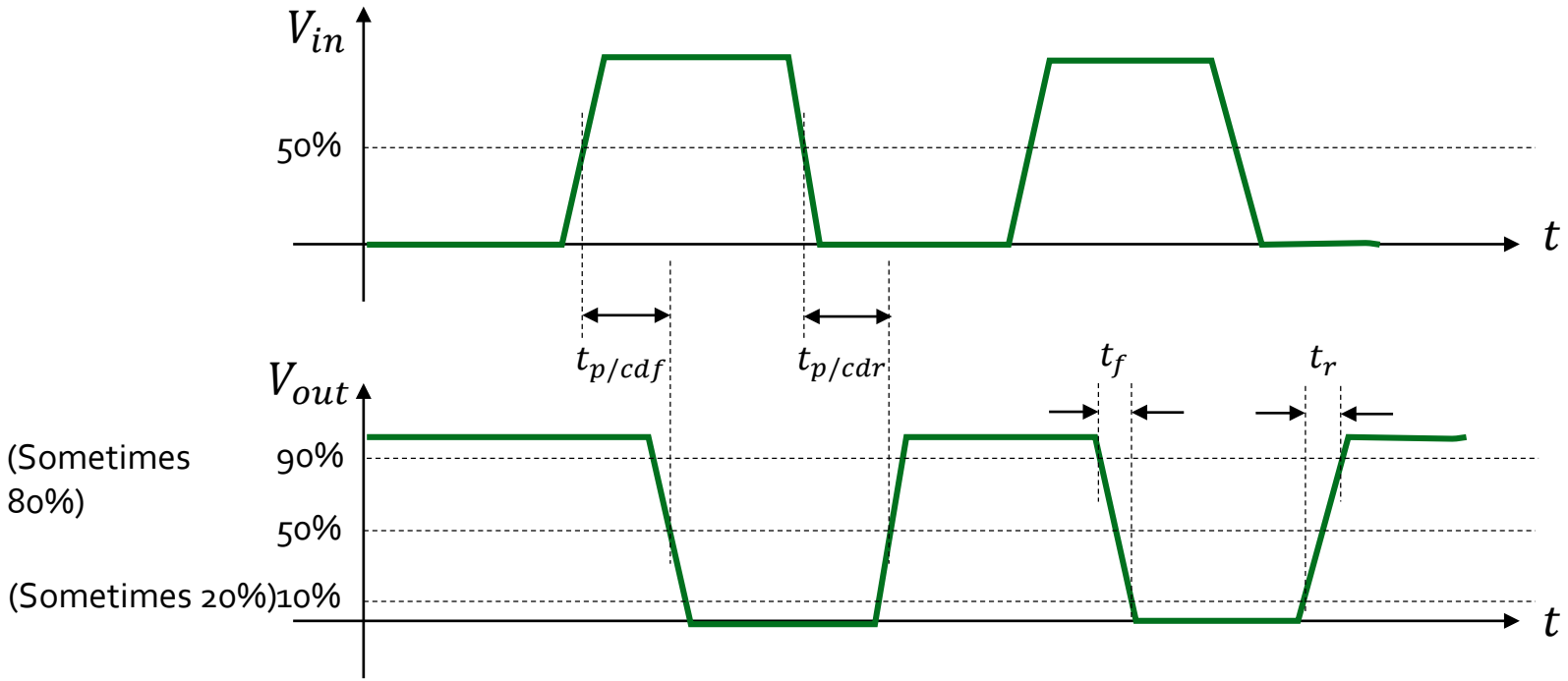
t_r - Rise time of output

t_f - Fall time of output

$t_{rf} = \frac{t_r + t_f}{2}$ - Edge rate

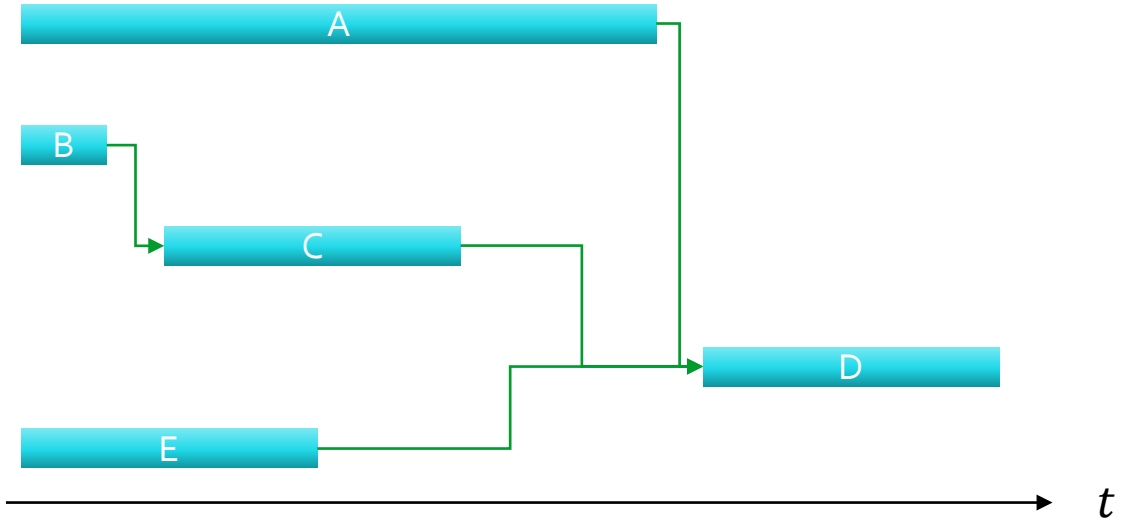
$f_{input,max} = \frac{1}{t_r + t_f}$ - Maximum input frequency

$f_{throughput,max} = \frac{1}{t_{pdf} + t_{pdr}}$ - Maximum output frequency



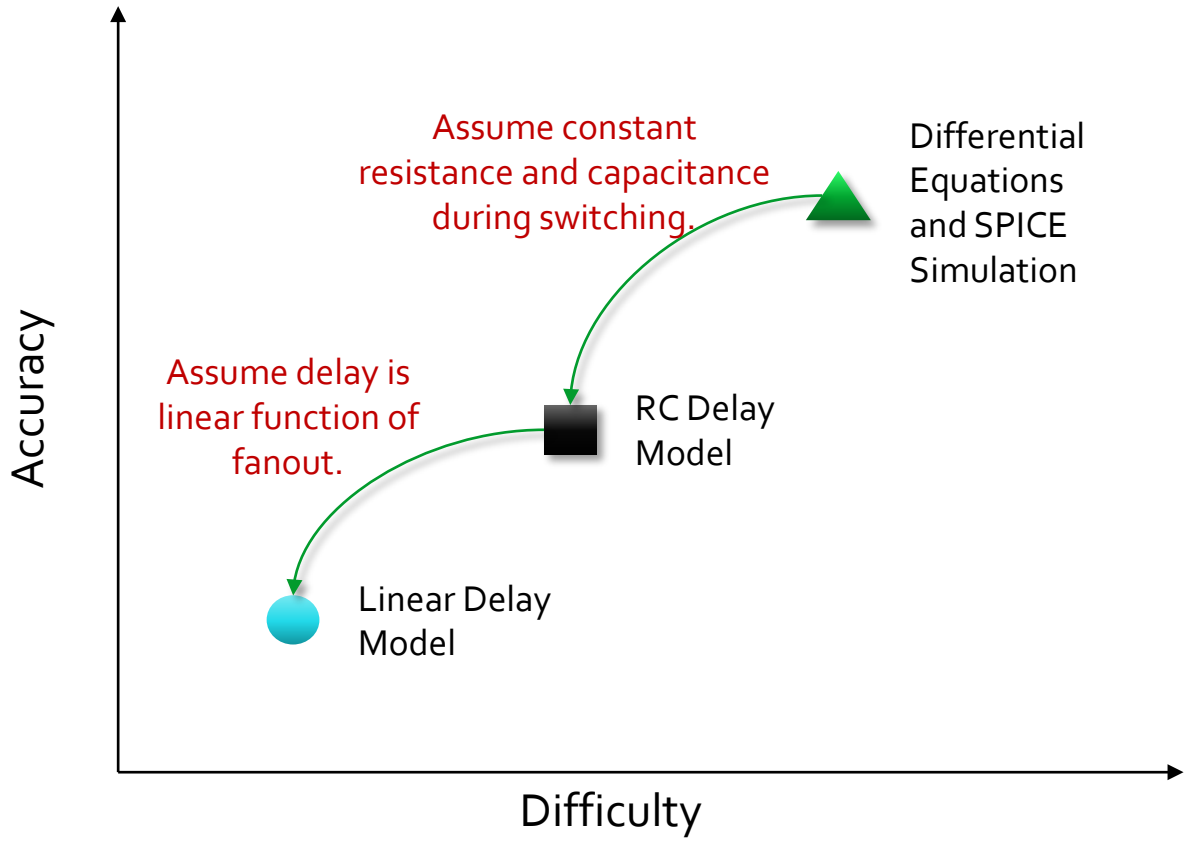
Timing Optimization

- Focus on reducing delay of *critical path* (path with max. delay)
- Optimizing critical path can change critical path



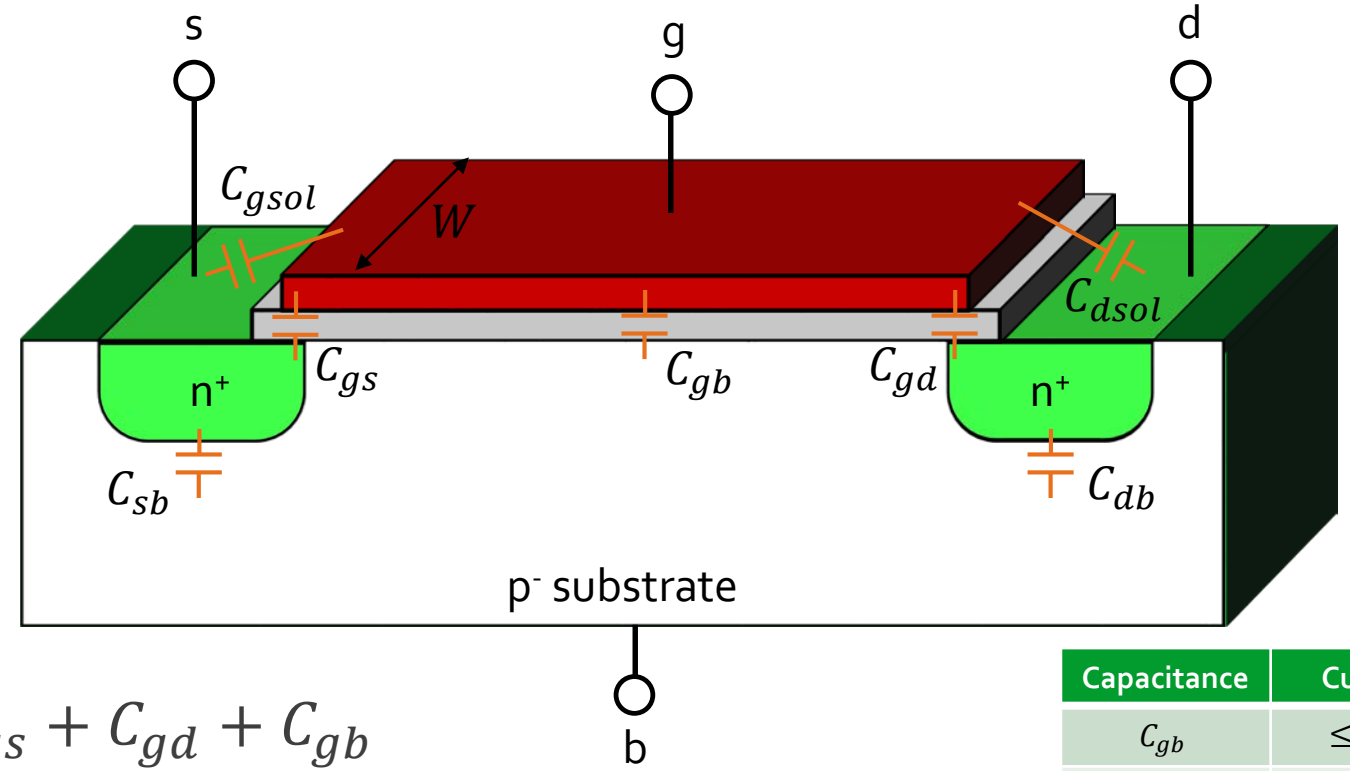
- Critical path is A-D
  - Optimizing B-C or E will not affect total delay
  - If delay of A is reduced below B-C, then new critical path is B-C

Ways of Measuring Delay



MOSFET Capacitances

» Definition: $C \equiv \mathrm{d}Q/\mathrm{d}V$



» $C_g = C_{gs} + C_{gd} + C_{gb}$

» $C_0 \equiv C_{ox}WL$

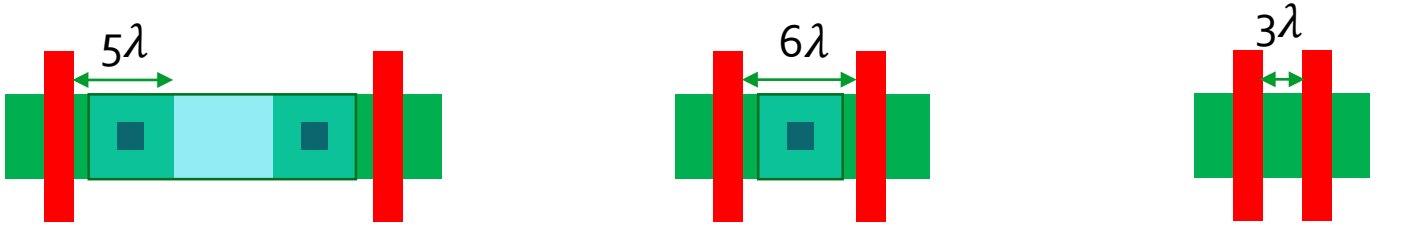
» C_{sb} and C_{db} are called <u>diffusion capacitances</u>

Gate apacitances depend on operating region

Capacitance	Cutoff	Linear	Saturation
C_{gb}	$\leq C_0$	0	0
C_{gs}	0	$C_0/2$	$2/3C_0$
C_{gd}	0	$C_0/2$	0
C_g	C_0	C_0	$2/3C_0$

MOSFET Capacitances

- Diffusion capacitances (C_{sb} and C_{db}) are called *parasitic*, since they are not "intentional," like the gate capacitance.
- Diffusion capacitance is formed at the p-n junctions of source/drain diffusion regions and body
  - Depends on **area** and depth of diffusion
  - Wider transistors have bigger diffusion area
  - Layout also affects diffusion area:



Isolated contacted diffusion

Shared contacted diffusion

Shared uncontacted diffusion

MOSFET Capacitances

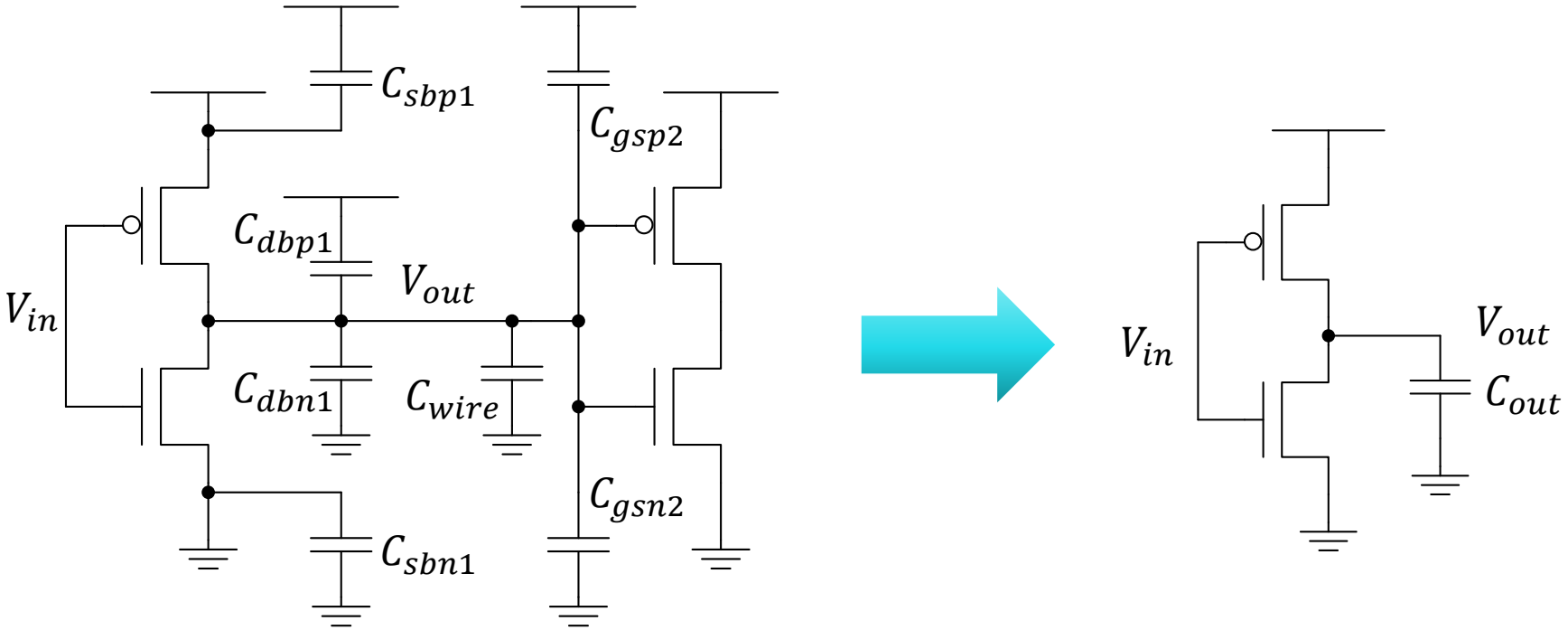
- C_{sb} and C_{sd} are complicated functions of the process parameters and geometries, but we can estimate that the per-micron capacitances are approximately:

$$C_g \approx C_{sb} \approx C_{db}$$

Vendor		Orbit	HP	AMI	AMI	TSMC	TSMC	TSMC	IBM	IBM	IBM
Model		MOSIS	MOSIS	MOSIS	MOSIS	MOSIS	MOSIS	TSMC	IBM	IBM	IBM
Feature Size f	nm	2000	800	600	600	350	250	180	130	90	65
V_{DD}	V	5	5	5	3.3	3.3	2.5	1.8	1.2	1.0	1.0
Gates											
C_g (delay)	fF/μm	1.77	1.67	1.55	1.48	1.90	2.30	1.67	1.04	0.97	0.80
C_g (power)	fF/μm	2.24	1.70	1.83	1.76	2.20	2.92	2.06	1.34	1.23	1.07
FO4 Inv. Delay	ps	856	297	230	312	210	153	75.6	45.9	37.2	17.2
nMOS											
C_d (isolated)	fF/μm	1.19	1.11	1.14	1.21	1.63	1.88	1.12	0.94	0.89	0.76
C_d (shared)	fF/μm	1.62	1.43	1.41	1.50	2.04	2.60	1.62	1.56	1.60	1.28
C_d (merged)	fF/μm	1.48	1.36	1.19	1.24	1.60	2.16	1.41	1.40	1.51	1.20
R_n (single)	kΩ · μm	30.3	10.1	9.19	11.9	5.73	4.02	2.69	2.34	2.35	1.34
R_n (series)	kΩ · μm	22.1	6.95	6.28	8.59	4.01	3.10	2.00	1.93	1.81	1.13
V_{tn} (const. I)	V	0.65	0.65	0.70	0.70	0.59	0.48	0.41	0.32	0.32	0.31
V_{tn} (linear ext.)	V	0.65	0.75	0.76	0.76	0.67	0.57	0.53	0.43	0.43	0.43
I_{dsat}	μA/μm	152	380	387	216	450	551	566	478	497	755
I_{off}	pA/μm	2.26	9.36	2.21	1.45	6.57	56.3	93.9	1720	4000	33400
I_{gate}	pA/μm	n/a	n/a	n/a	n/a	n/a	n/a	n/a	1.22	3620	8520
pMOS											
C_d (isolated)	fF/μm	1.42	1.17	1.31	1.42	1.89	2.07	1.24	0.94	0.74	0.73
C_d (shared)	fF/μm	1.92	1.62	1.73	1.86	2.37	2.89	1.79	1.56	1.25	1.25
C_d (merged)	fF/μm	1.52	1.23	1.35	1.43	1.83	2.40	1.56	1.41	1.16	1.18
R_p (single)	kΩ · μm	67.1	26.7	19.9	29.6	16.1	8.93	6.51	6.39	5.47	2.87
R_p (series)	kΩ · μm	53.9	21.4	15.4	23.6	13.3	6.91	5.41	5.48	4.92	2.42
$\lvert V_{tp}\rvert$ (const. I)	V	0.72	0.91	0.90	0.90	0.83	0.46	0.43	0.33	0.35	0.33
$\lvert V_{tp}\rvert$ (linear ext.)	V	0.71	0.94	0.93	0.93	0.88	0.52	0.51	0.42	0.43	0.42
I_{dsat}	μA/μm	70.5	154	215.3	99.0	181	245	228	177	187	360
I_{off}	pA/μm	2.18	1.57	2.08	1.38	2.06	30.1	25.2	1330	2780	19500
I_{gate}	pA/μm	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.06	1210	2770

Delay of an Inverter

» Consider a CMOS inverter driving another CMOS inverter:



» $C_{out} = C_{dbp1} + C_{dbn1} + C_{wire} + C_{gsp2} + C_{gsn2}$

Delay of an Inverter

- Let's consider t_{pdf} of the inverter
  - Time $t = 0$, $V_{out} = 0$ and V_{in} changes from 1 to 0
  - PMOS is cutoff, NMOS on, initially in saturation

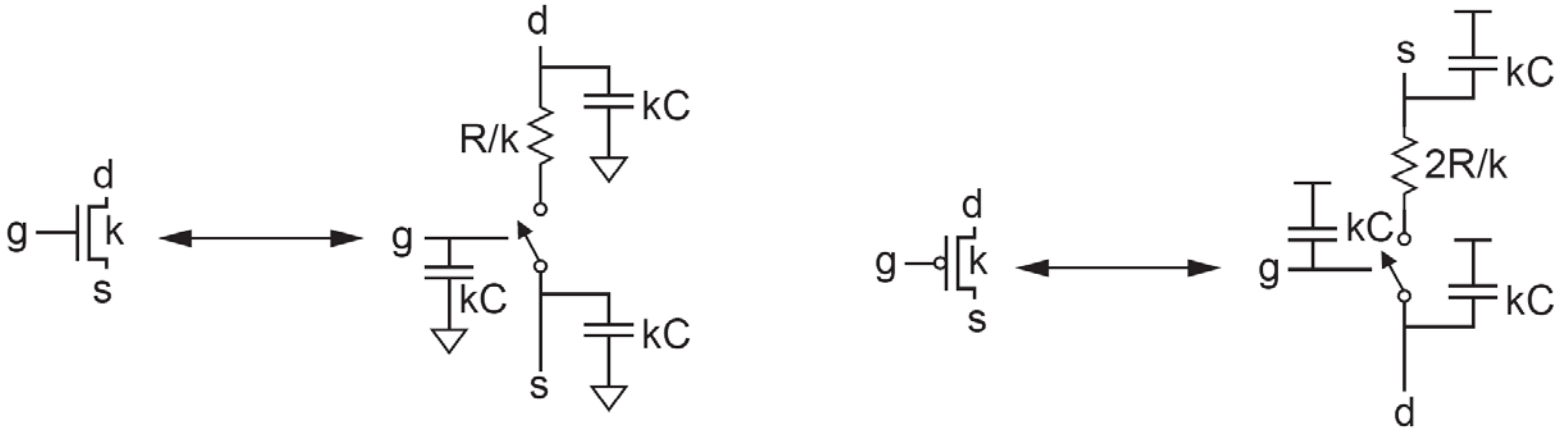
$$Q = VC \rightarrow i = \frac{\mathrm{d}v}{\mathrm{d}t}C \rightarrow \mathrm{d}t = \frac{\mathrm{d}v}{i}C \rightarrow t_{pd}$$

$$= \int_0^{V_{dd}/2} \frac{C_{out}}{i_{dsn1}} \mathrm{d}V_{out}$$

- But, i_{dsn1} depends on V_{out}
  - Initially in saturation, then linear mode, then cutoff
  - Messy non-linear differential equation, has to be solved numerically

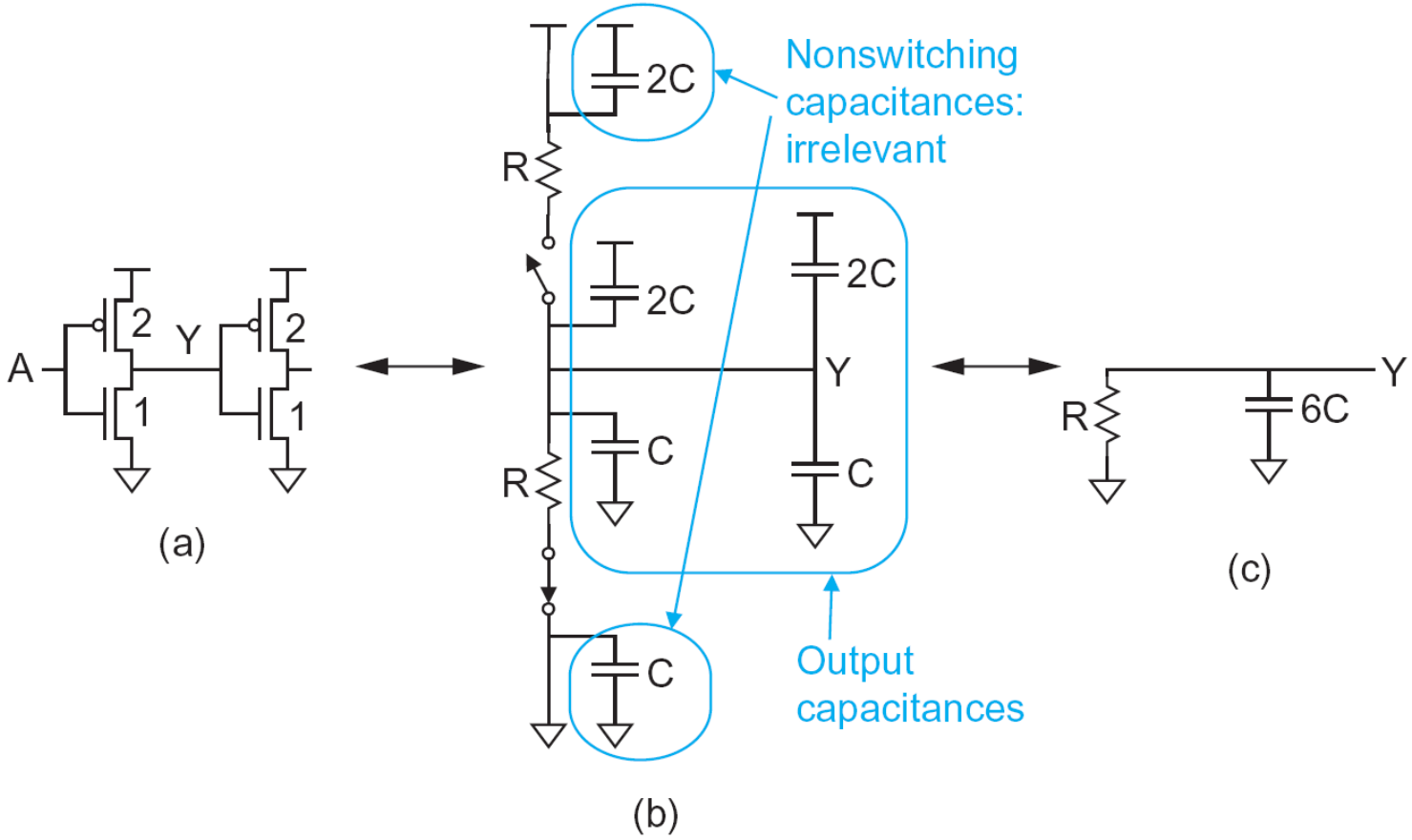
RC Delay Model

- Idea: Model transistor as resistor that captures its average V_{ds} vs. I_{ds} slope
- Recall $R \propto \frac{1}{\beta} = \frac{L}{W\mu C_{ox}}$, $C_g \propto WL$, $C_g \approx C_{sb} \approx C_{db}$
- For NMOS and PMOS with k times the unit transistor widths and $\mu_n = 2\mu_p$:



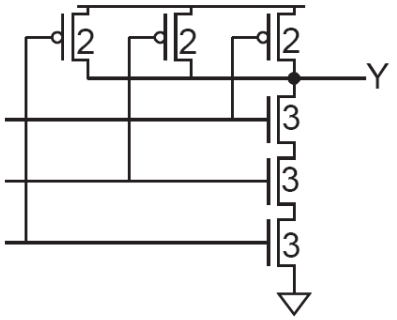
RC Delay Model

» Equivalent circuit for a fanout-of-1 inverter:

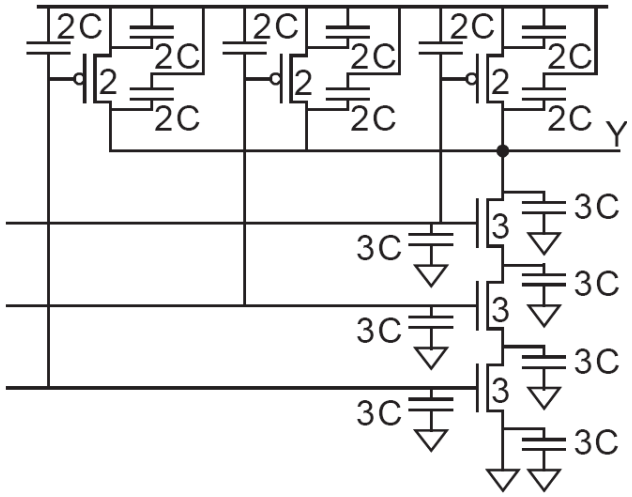


RC Delay Model

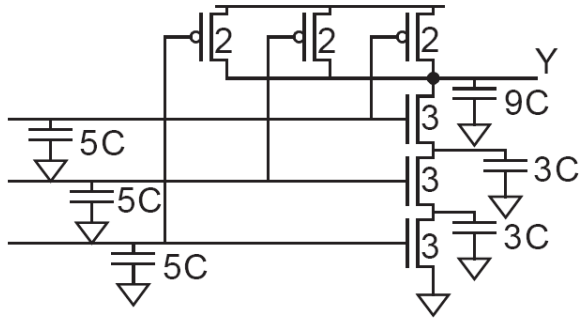
» 3-input NAND gate equivalent RC circuits:



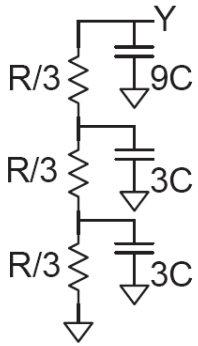
(a)



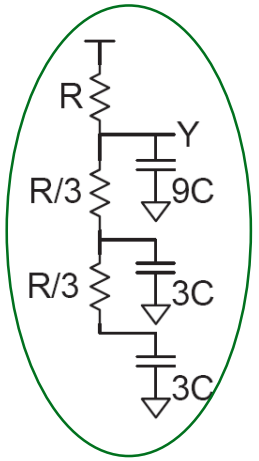
(b)



(c)



(d)

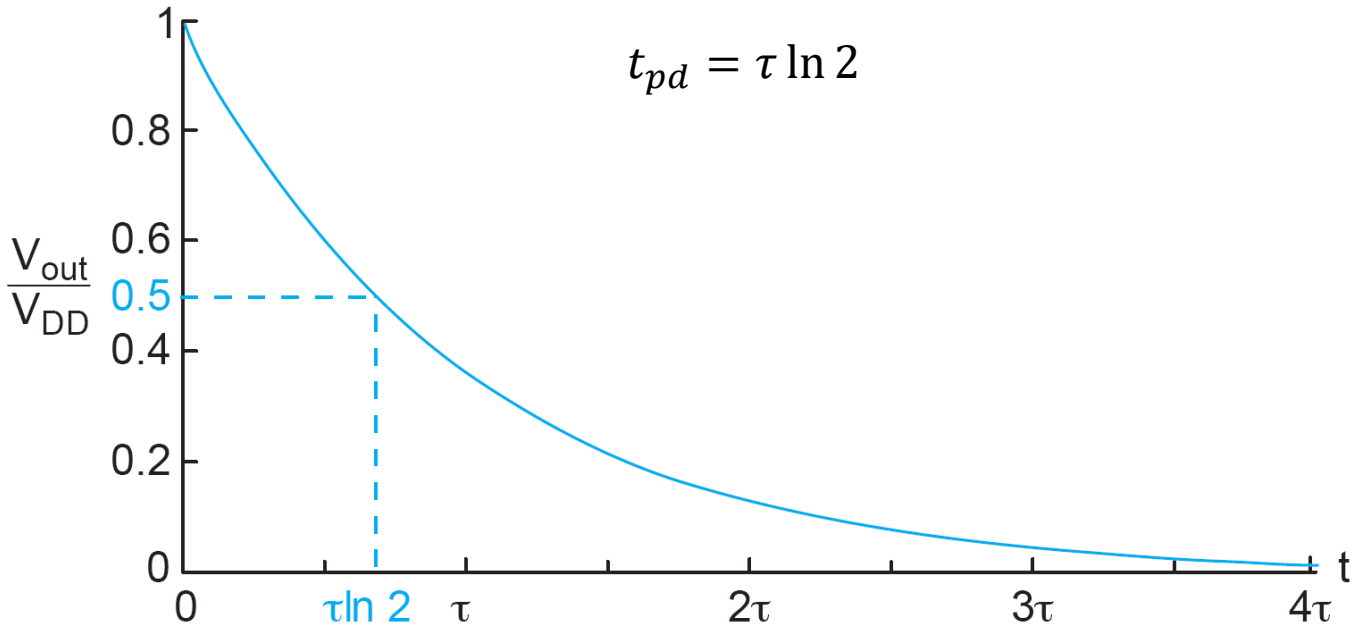


(e)

Worst-case rising

RC Delay Model

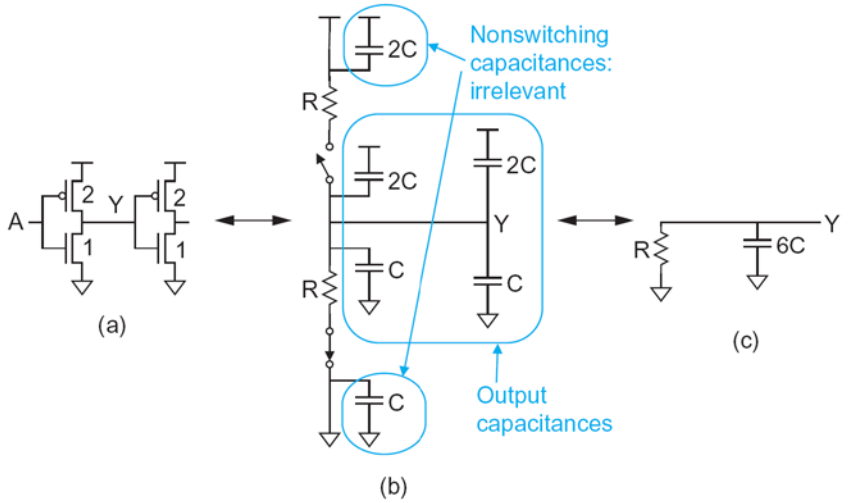
» We can get the time constant from the equivalent τ RC circuit and use it to estimate the delay:



» The ln 2 factor is often ignored

RC Delay Model

» Inverter with fanout-of-1:



» Falling propagation delay: $V_{out} = V_{DD}\mathrm{e}^{-t/6RC} \rightarrow \tau = 6RC \rightarrow t_{pd} = 6RC \ln 2$

» Absorb the $\ln 2$ factor into R: $t_{pd} = 6RC$

RC Delay Model

- Fanout-of-4 delay (FO4) is a common figure of merit
- For the inverter in the previous slide, the FO4 delay is

$$t_{pd} = 15RC$$

- It's often convenient to define τ' as the delay of a fanout-of-1 inverter with *no parasitic capacitances*
- Usually, we use $\tau' = 3RC$
- Now we can define delays independent of RC:

$$d \equiv \frac{t_{pd}}{\tau'} = 5$$

$$d = h + 1$$ (FO4 inverter)

(fanout-of-h inverter)

Elmore Delay Model

- Calculating the time constant for a first-order RC circuit is easy
- Most circuits are more complicated
- Elmore delay approximates the time constant of an RC tree as

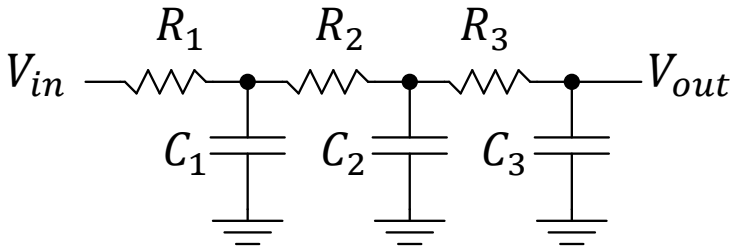
$$\tau \approx \sum_i R_{is} C_i$$

where R_{is} is the shared resistance from switching source to node i

- Related to "zero-value time constant"

Elmore Delay Model

» Example of simple RC circuit:



» When V_{in} switches, we first have to charge C_1
- $\tau_1 = R_1 C_1$

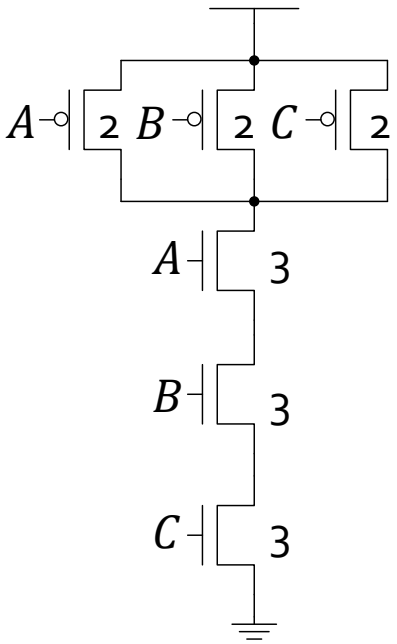
» Then, C_2
- $\tau_2 = (R_1 + R_2) C_2$

» Then, C_3
- $\tau_3 = (R_1 + R_2 + R_3) C_3$

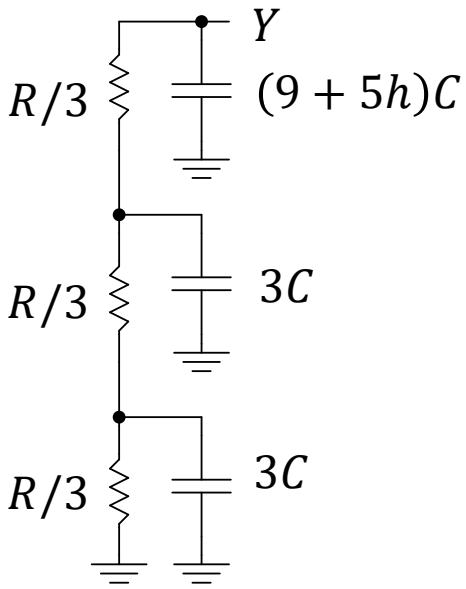
$$\tau \approx R_1 C_1 + (R_1 + R_2) C_2 + (R_1 + R_2 + R_3) C_3$$

Elmore Delay Model

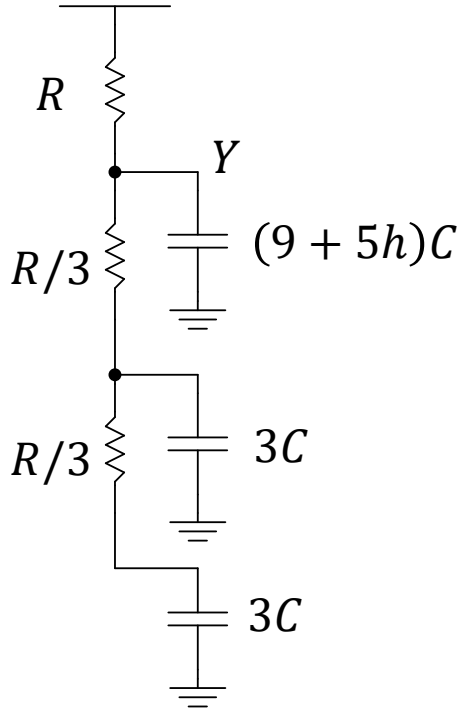
» Example: Estimate the rising and falling propagation (worst-case) delays for a 3-input NAND gate with fanout of h



Falling Delay Equivalent Circuit



Rising Delay Equivalent Circuit

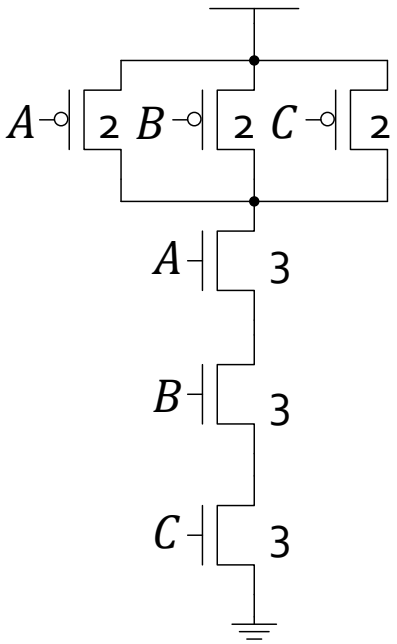


$$t_{pdf} = \frac{R}{3}3C + \frac{2R}{3}3C + R(9+5h)C = (12+5h)RC$$

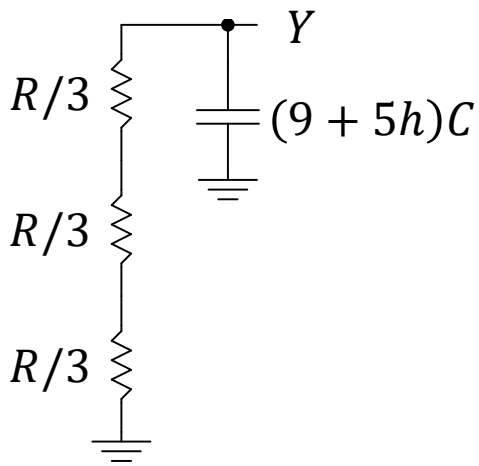
$$t_{pdr} = R(9+5h)C + R3C + R3C = (15+5h)RC$$

Elmore Delay Model

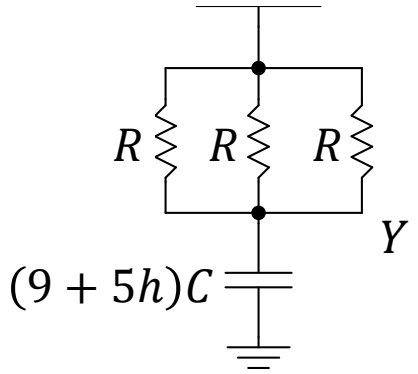
» Estimate the rising and falling *contamination* (best-case) delay for a 3-input NAND gate with fanout of h



Falling Delay Equivalent Circuit



Rising Delay Equivalent Circuit



$$t_{cdf} = \left(\frac{R}{3} + \frac{R}{3} + \frac{R}{3}\right)(9 + 5h)C = (9 + 5h)RC \qquad t_{cdr} = \left(\frac{R}{3}\right)(9 + 5h)C = \left(3 + \frac{5}{3}h\right)RC$$

Elmore Delay Model

» Note that we can write the delays as

$$d = f + p = gh + p$$

- f is the "effort" delay
- p is the "parasitic" delay associated with charging a gate's own internal parasitic capacitances
- g is called the "logical effort" and represents the complexity of the gate being driven
- h is the fanout or "electrical effort"

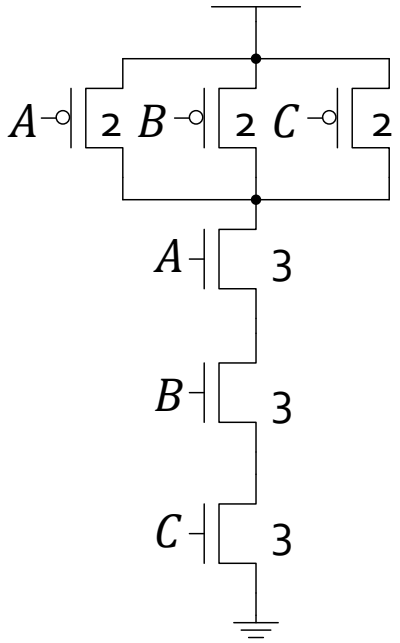
» For the 3-input NAND gate, the worst-case delay was $t_{pdr} = (15 + 5h)RC$, so

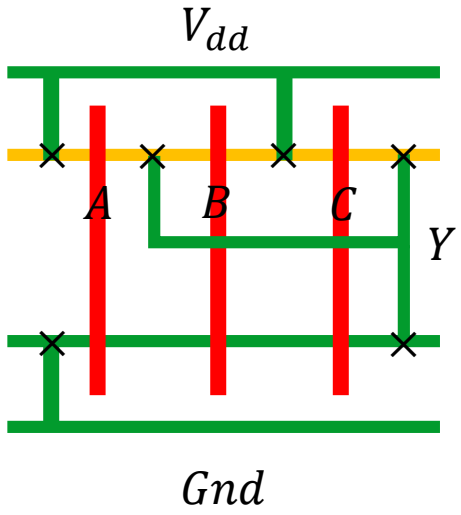
$$d \equiv \frac{t_{pdr}}{\tau'} = \frac{5}{3}h + 5$$

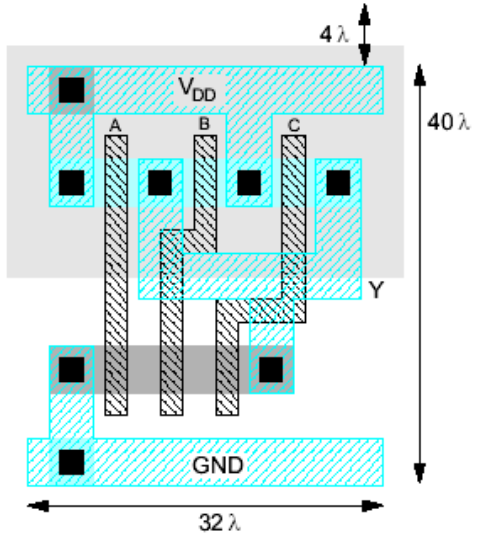
- The logical effort is 5/3, and the parasitic delay is 5

Elmore Delay Model

» Diffusion (parasitic) capacitance depends on the layout and how diffusions are shared:







» A and B PMOS drains are shared, so only count diffusion capacitance once, so capacitance at Y is $(7 + 5h)C$ instead of $(9 + 5h)C$

» Sometimes uncontacted diffusion is counted as only (½)C