ENEE 359a
Digital VLSI Design

Transistor Sizing
& Logical Effort

Prof. Bruce Jacob

blj@ece.umd.edu

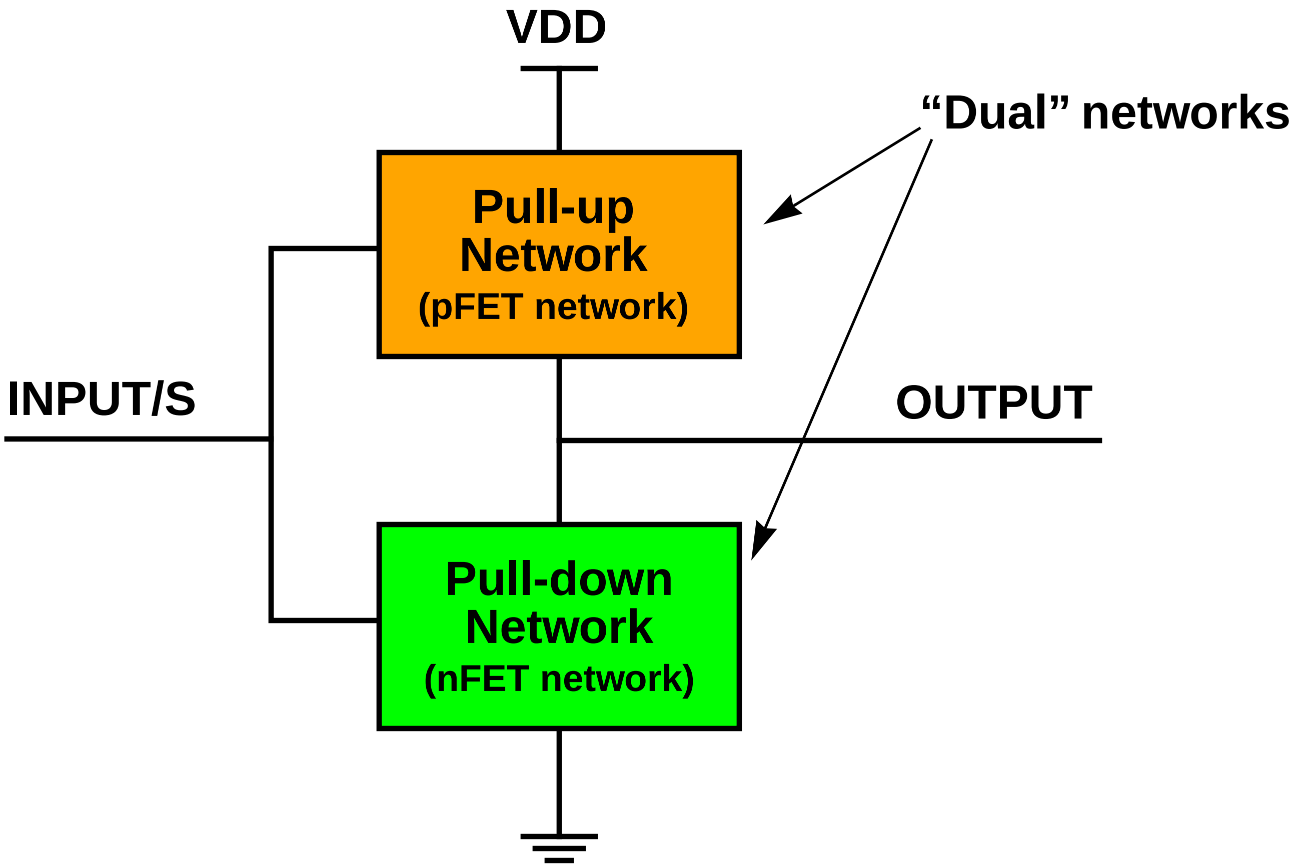
Credit where credit is due:
Slides contain original artwork (© Jacob 2004) as well as material taken liberally from Irwin & Vijay’s CSE477 slides (PSU), Schmit & Strojwas’s 18-322 slides (CMU), Dally’s EE273 slides (Stanford), Wolf’s slides for *Modern VLSI Design*, and/or Rabaey’s slides (UCB).

Overview

- Sizing of transistors to balance performance of single inverter
- More on RC time constant, first-order approximation of time delays
- Sizing in complex gates, examples
- Sizing of inverter chains for driving high capacitance loads (off-chip wires)

Resistance

WOULD LIKE BALANCED NETWORKS:



Resistance

Resistance of MOSFET:

$$R_{\mathrm{n}} = \frac{1}{\mu_{\mathrm{n}} C_{\mathrm{ox}}(V_{\mathrm{GS}} - V_{\mathrm{Tn}})}\left(\frac{L}{W}\right)$$

- Increasing W decreases the resistance; allows more current to flow

Oxide capacitance $C_{\mathrm{ox}} = \varepsilon_{\mathrm{ox}} / t_{\mathrm{ox}}$ **[F/cm^2]**

Gate capacitance $C_{\mathrm{G}} = C_{\mathrm{ox}} W L$ **[F]**

Transconductance $\beta_{\mathrm{n}} = \mu_{\mathrm{n}} C_{\mathrm{ox}}\left(\frac{W}{L}\right) = k'_{\mathrm{n}}\left(\frac{W}{L}\right)$

(units [A/V^2])

Resistance

nFET vs. pFET

$$\boldsymbol{R}_{\mathrm{n}} = \frac{1}{\beta_{\mathrm{n}}(\boldsymbol{V}_{\mathrm{DD}} - \boldsymbol{V}_{\mathrm{Tn}})} \qquad \beta_{\mathrm{n}} = \mu_{\mathrm{n}} \boldsymbol{C}_{\mathrm{ox}} \left(\frac{\boldsymbol{W}}{\boldsymbol{L}}\right)_{\mathrm{n}}$$

$$\boldsymbol{R}_{\mathrm{p}} = \frac{1}{\beta_{\mathrm{p}}(\boldsymbol{V}_{\mathrm{DD}} - |\boldsymbol{V}_{\mathrm{Tp}}|)} \qquad \beta_{\mathrm{p}} = \mu_{\mathrm{p}} \boldsymbol{C}_{\mathrm{ox}} \left(\frac{\boldsymbol{W}}{\boldsymbol{L}}\right)_{\mathrm{p}}$$

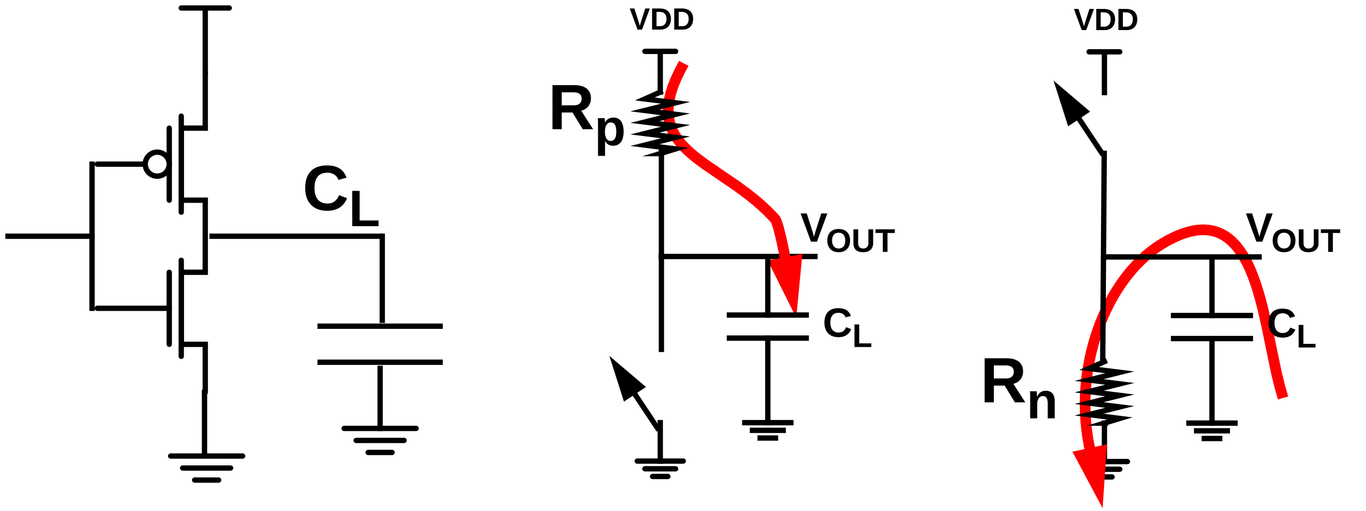
$$\frac{\mu_{\mathrm{n}}}{\mu_{\mathrm{p}}} = \boldsymbol{r}$$

Typically (2 .. 3)

(μ is the carrier mobility through device)

Transistor Sizing

SIMPLE CASE: Inverter



Charging: Vout rising

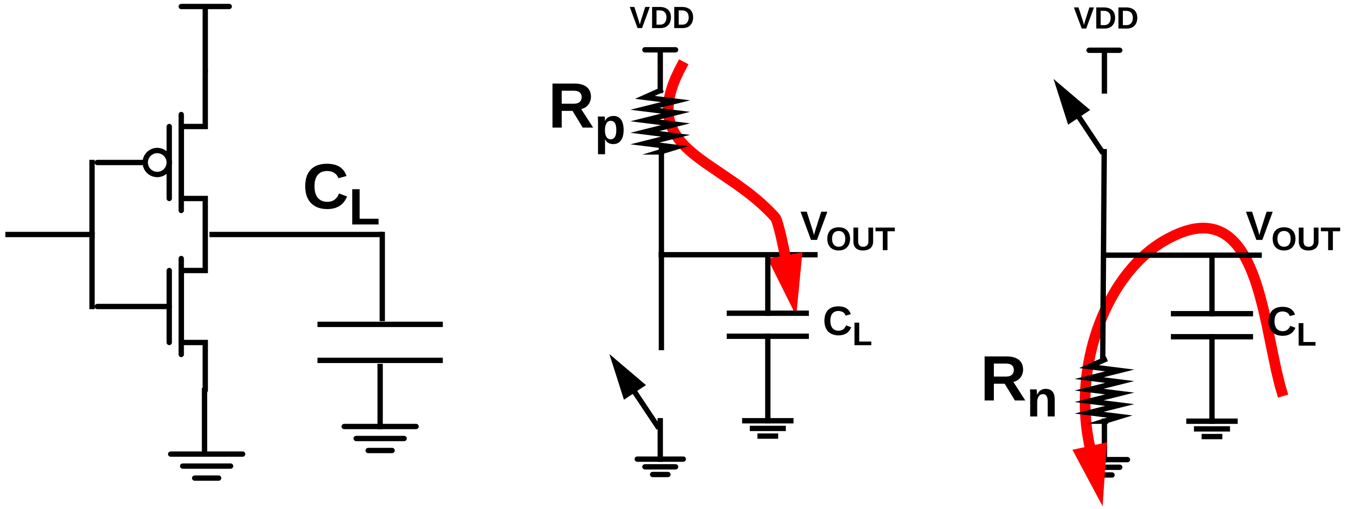
Discharging: Vout falling

If $(W/L)_p = r(W/L)_n$ then $\beta_n = \beta_p$
(and $R_n = R_p$)
… symmetric inverter

Make pFET bigger (wider) by factor of *r*

Transistor Sizing

SIMPLE CASE: Inverter



Charging: Vout rising

Discharging: Vout falling

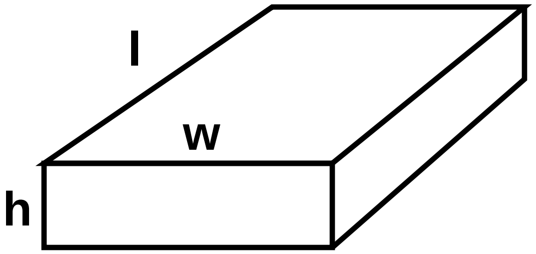
$$t_{pLH} = \ln(2)\, R_p\, C_L = 0.69\, R_p\, C_L$$

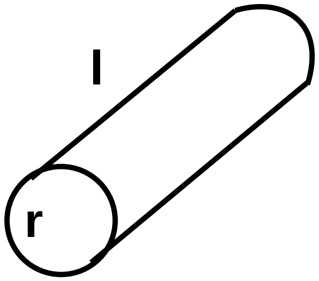
$$t_{pHL} = \ln(2)\, R_n\, C_L = 0.69\, R_n\, C_L$$

$$t_p = (t_{pHL} + t_{pLH})/2 = 0.69\, C_L(R_n + R_p)/2$$

(note: the ln(2)RC term comes from first-order analysis of simple RC circuit’s respose to step input ... time for output to reach 50% value … more detail on this in a moment, after we discuss *capacitance* …)

Wire Resistance





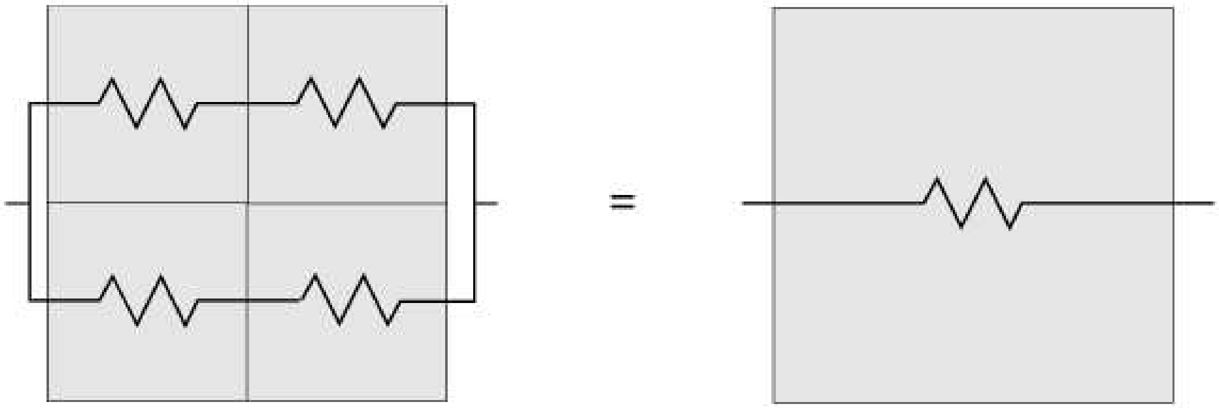
- $R = \rho l/A = \rho l/(wh)$ **for rectangular wires (on-chip wires & vias, PCB traces)**
- $R = \rho l/A = \rho l/(\pi r^2)$ **for circular wires (off-chip, off-PCB)**

Material	**Resistivity ρ (Ω-m)**
Silver (Ag)	1.6 x 10-8
Copper (Cu)	1.7 x 10-8
Gold (Au)	2.2 x 10-8
Aluminum (Al)	2.7 x 10-8
Tungsten (W)	5.5 x 10-8

Bruce Jacob

University of
Maryland
ECE Dept.

Sheet Resistance



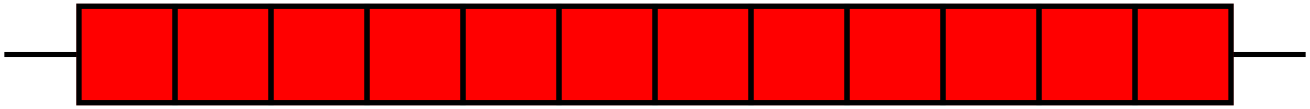
R = ρl/(wh) = l/w•ρ/h for rectangular wires

Sheet resistance R_{sq} = ρ/h ***(h=thickness)***

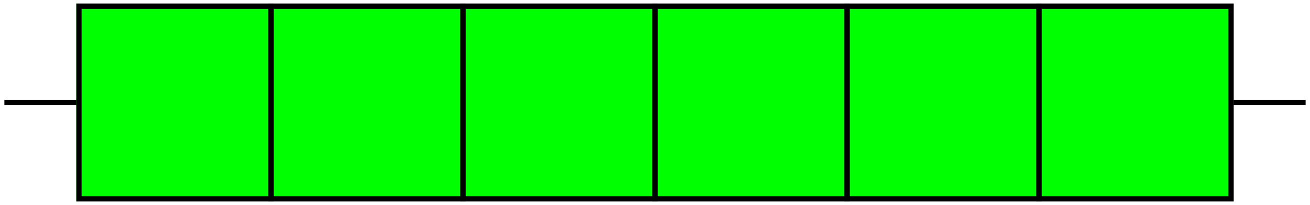
Material	**Sheet resistance R_{sq} (Ω/sq)**
n, p well diffusion	1000 to 1500
n+, p+ diffusion	50 to 150
polysilicon	150 to 200
polysilicon with silicide	4 to 5
Aluminum	0.05 to 0.1

More on Resistance

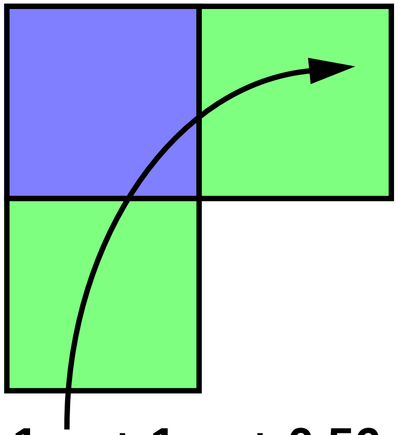
Sheet resistance R_{sq}



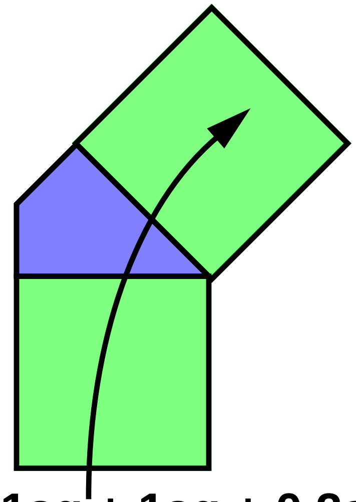
12 squares



6 squares

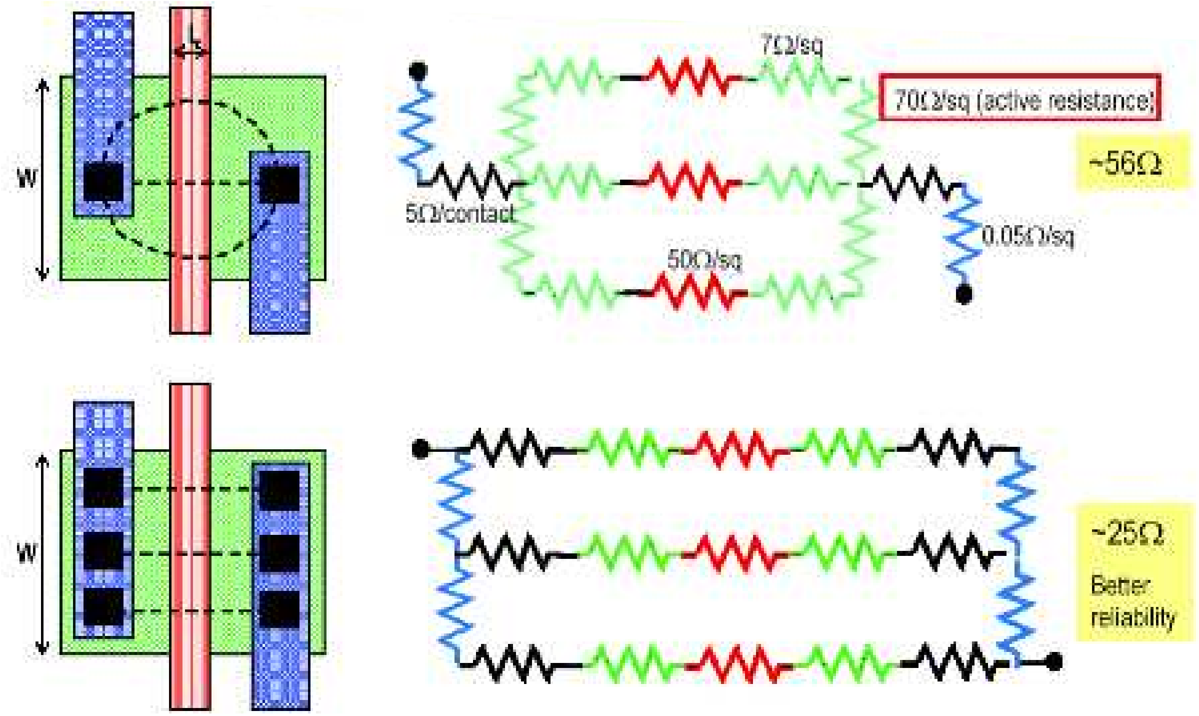


$= 1sq + 1sq + 0.56sq$



$= 1sq + 1sq + 0.2sq$

More on Resistance

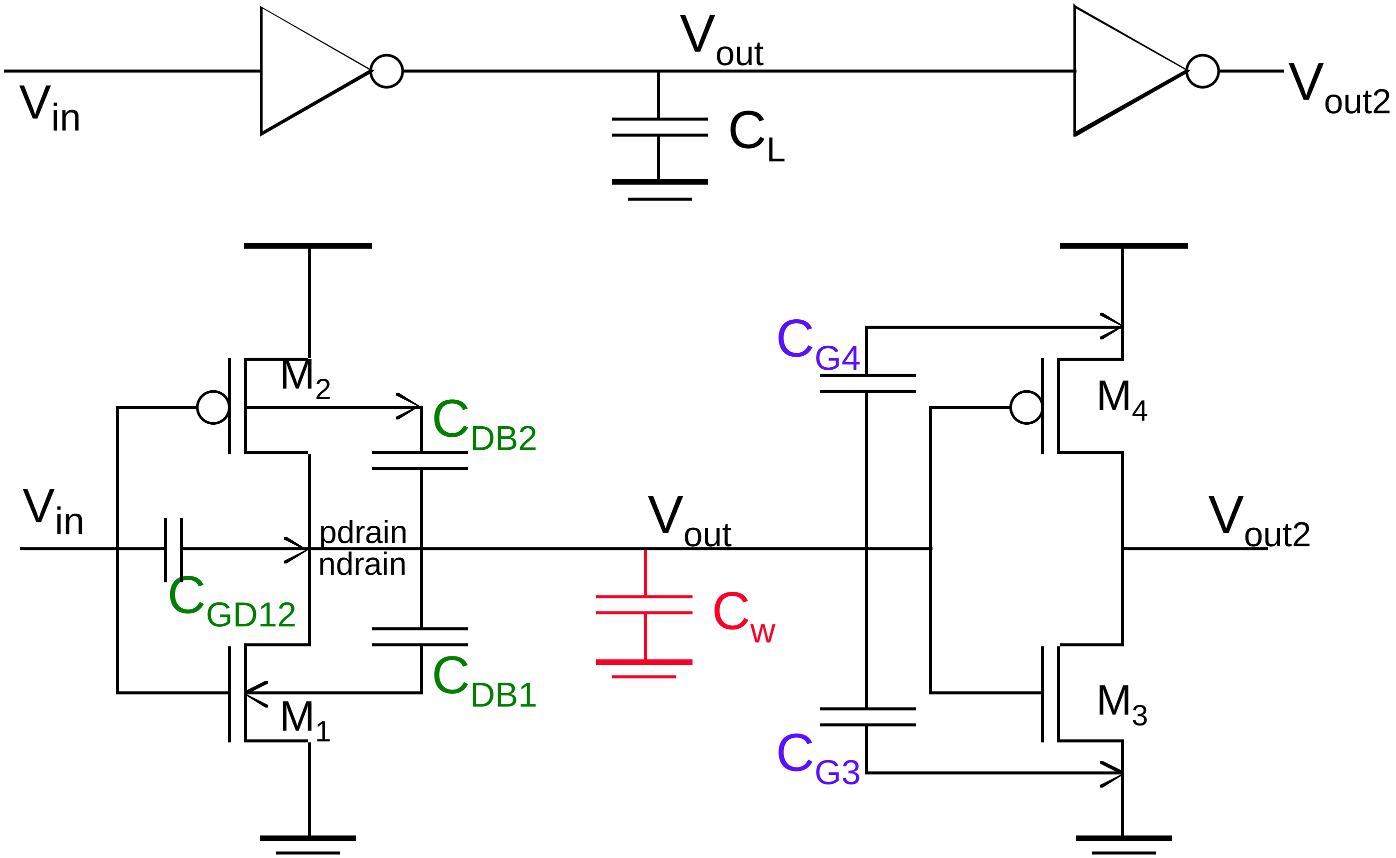


(it’s not just the channel that counts)

Bruce Jacob

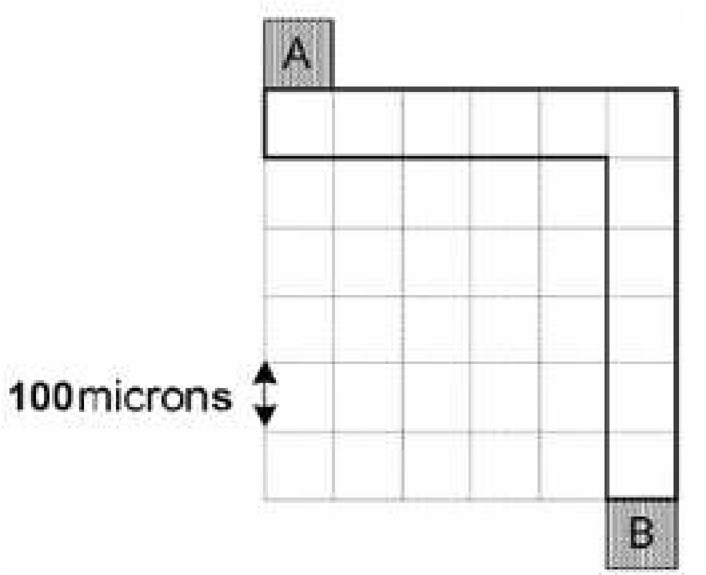
University of
Maryland
ECE Dept.

And Now … Capacitance (C_L)



- intrinsic MOS transistor capacitances
- extrinsic MOS transistor (fanout) capacitances
- wiring (interconnect) capacitance

C_W, a Large Example



Given:

R = 40 mΩ/□

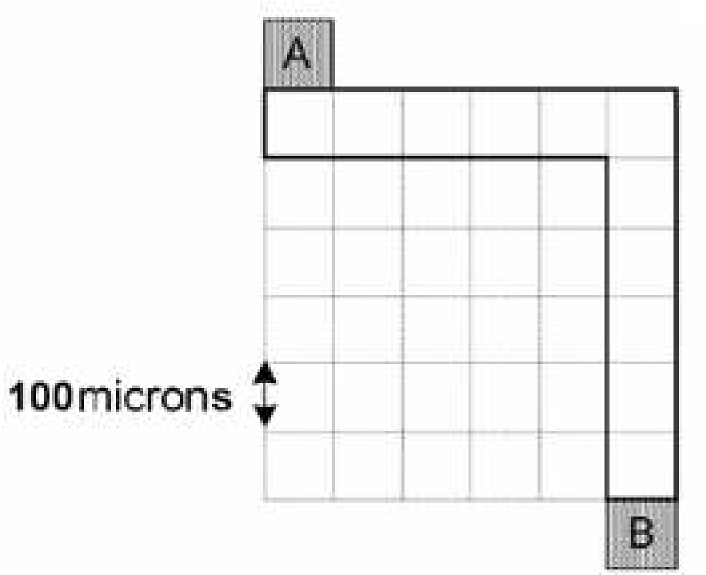
C_{fringe} = 0.044 fF/µm

C_{plate} = 0.031 fF/µm^2

Determine:

the resistance between A and B, the plate and fringe capacitances to ground.

C_W, a Large Example



Given:

R = 40 mΩ/□

C_{fringe} = 0.044 fF/μm

C_{plate} = 0.031 fF/μm^2

Determine:

the resistance between A and B, the plate and fringe capacitances to ground.

R = (9 + 2∗ 0.56 squares) * 40 mΩ/□ = 404.8 mΩ

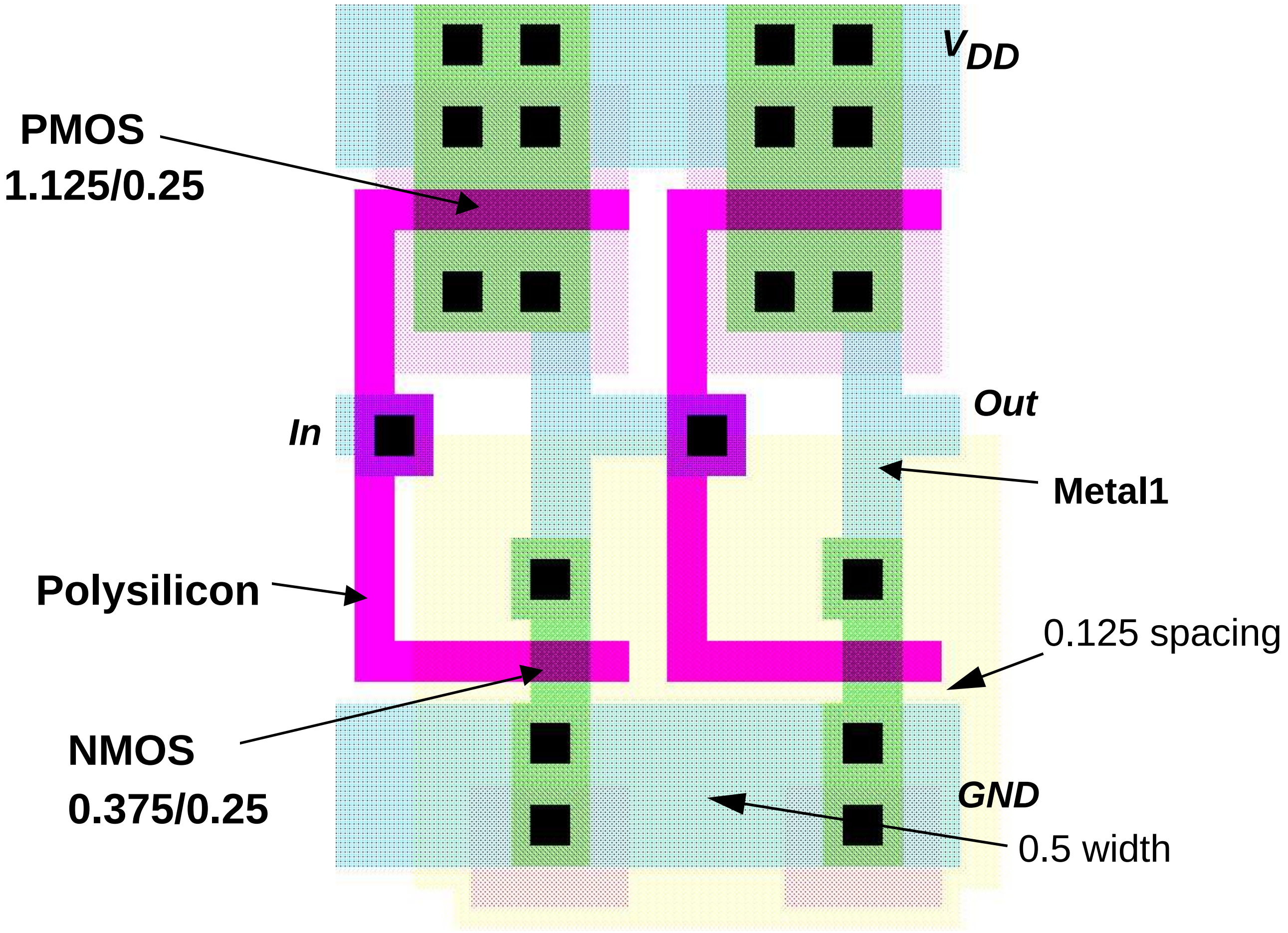
$C_{fringe,g}$ = Perimeter C_{fringe} = 96.8fF

(Perimeter = 2200 μm)

$C_{plate,g}$ = Area C_{plate} = 3.41pF

(Area = 110,000 μm^2)

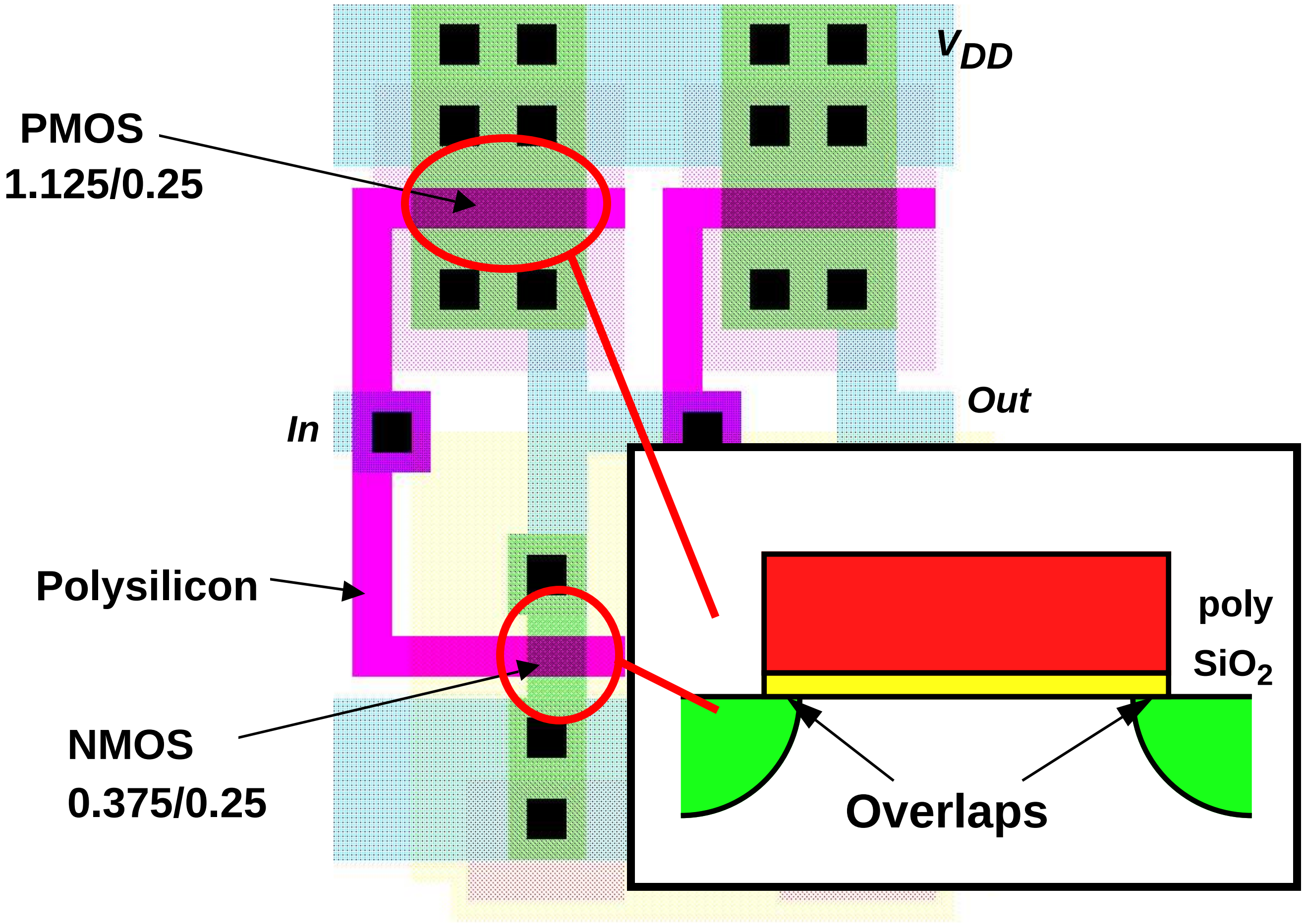
Two Chained Inverters



Bruce Jacob

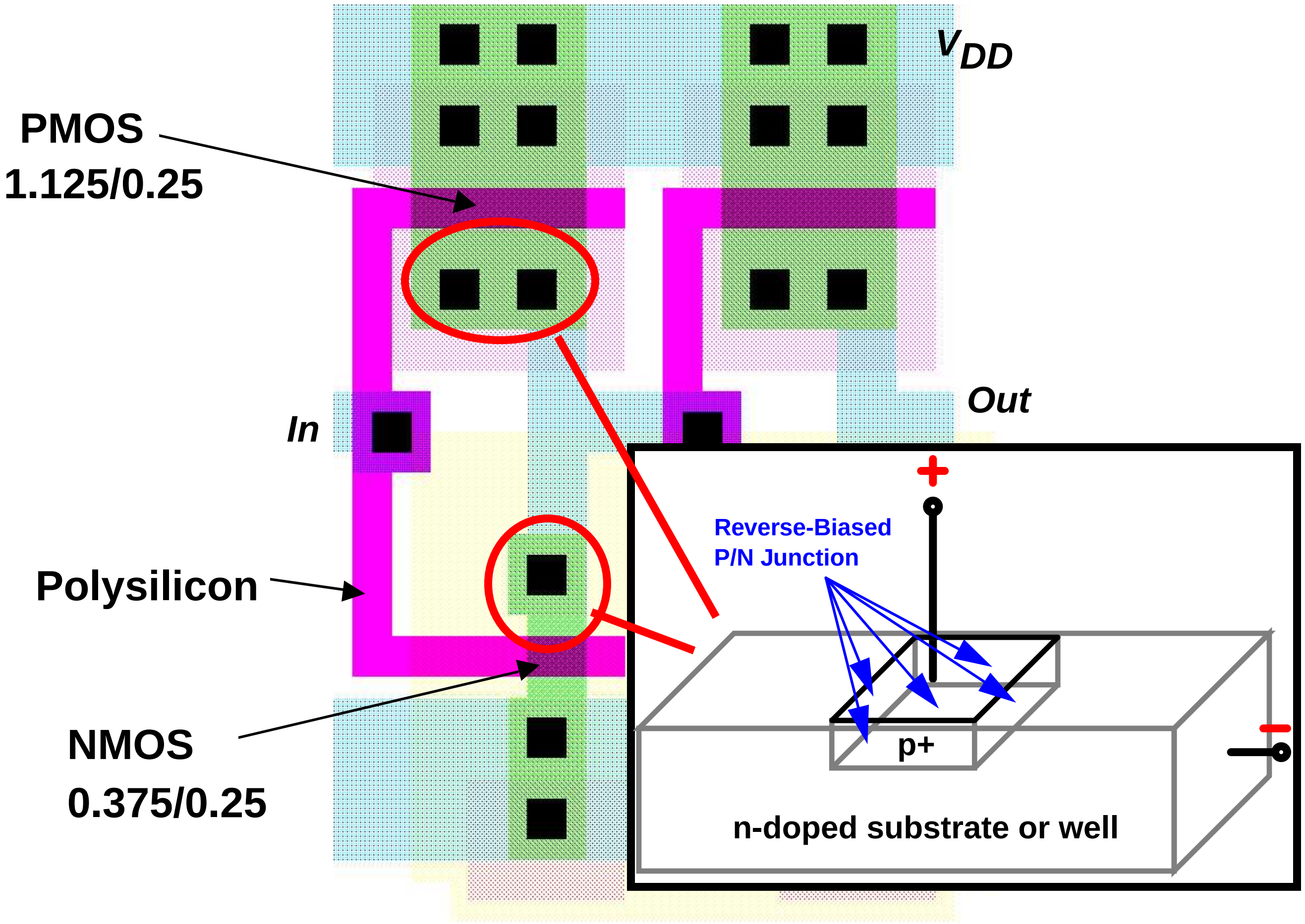
University of
Maryland
ECE Dept.

Gate-Drain Capacitance C_{GD}



Scales with transistor width W

Diffusion Capacitance C_{DB}



PMOS
1.125/0.25

V_{DD}

In

Out

Polysilicon

NMOS
0.375/0.25

Reverse-Biased
P/N Junction

+

−

p+

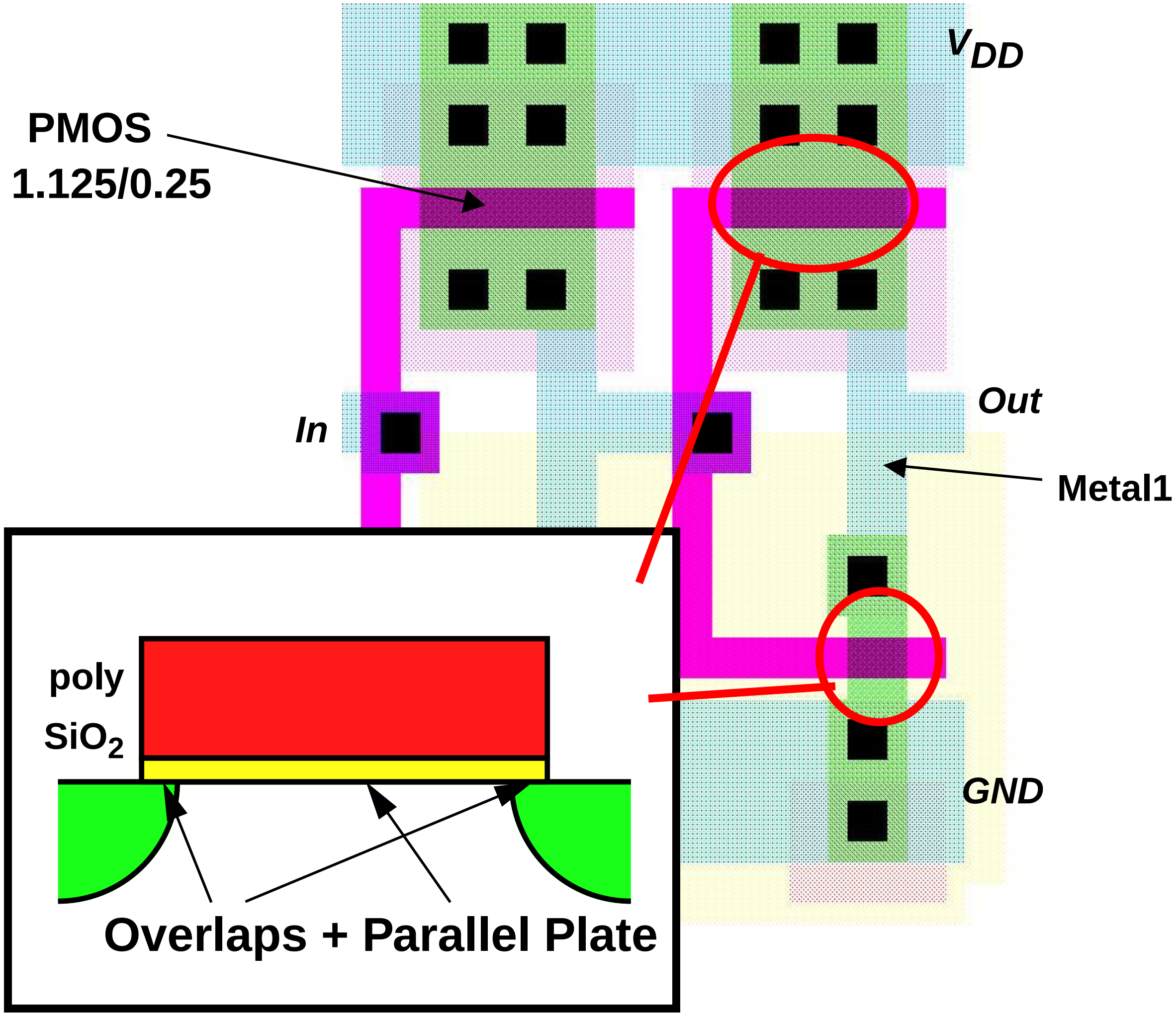
n-doped substrate or well

- **Drain is reverse-biased diode, non-linear C dependent on drain voltage (approx. nonlinearity with linear eqn, using K terms for bottom plate and sidewalls)**

Bruce Jacob

University of Maryland
ECE Dept.

Gate/Fan-out Capacitance C_G



Scales with both W *and* L

ENEE 359a
Lecture/s 9
Transistor Sizing

Bruce Jacob

University of
Maryland
ECE Dept.

SLIDE 19

Two Chained Inverters: C_L

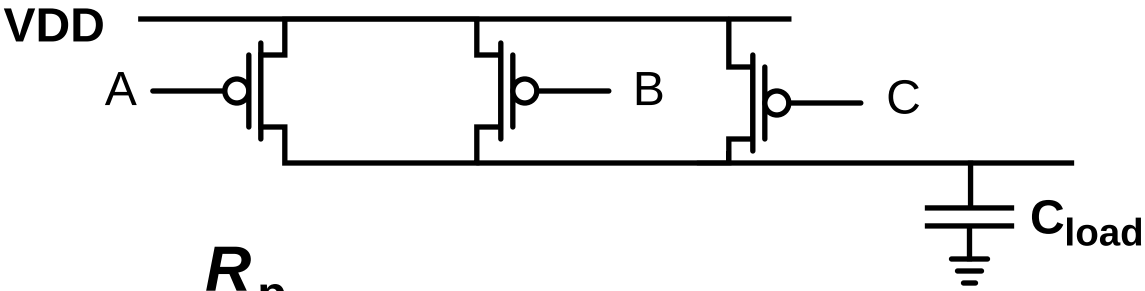
C Term	Expression	Value (fF) H → L	Value (fF) L → H
C_{GD1}	$2\ C_{on}\ W_n$	0.23	0.23
C_{GD2}	$2\ C_{op}\ W_p$	0.61	0.61
C_{DB1}	$K_{eqbpn} AD_n C_j + K_{eqswn} PD_n C_{jsw}$	0.66	0.90
C_{DB2}	$K_{eqbpp} AD_p C_j + K_{eqswp} PD_p C_{jsw}$	1.50	1.15
C_{G3}	$2\ C_{on}\ W_n + C_{ox}\ W_n\ L_n$	0.76	0.76
C_{G4}	$2\ C_{op}\ W_p + C_{ox}\ W_p\ L_p$	2.28	2.28
C_W	From extraction	0.12	0.12
C_L	Sum	6.1	6.0

- **Terms in red: under control of designer**
- **C_L split between intrinsic and extrinsic/wire sources**

UNIVERSITY OF MARYLAND

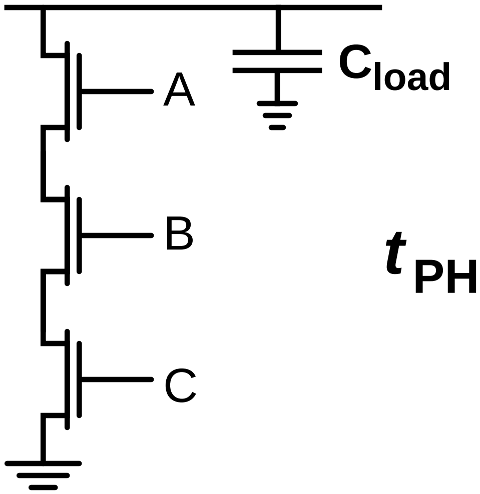
MOSFET Switching

Parallel switching (all switch at same time):



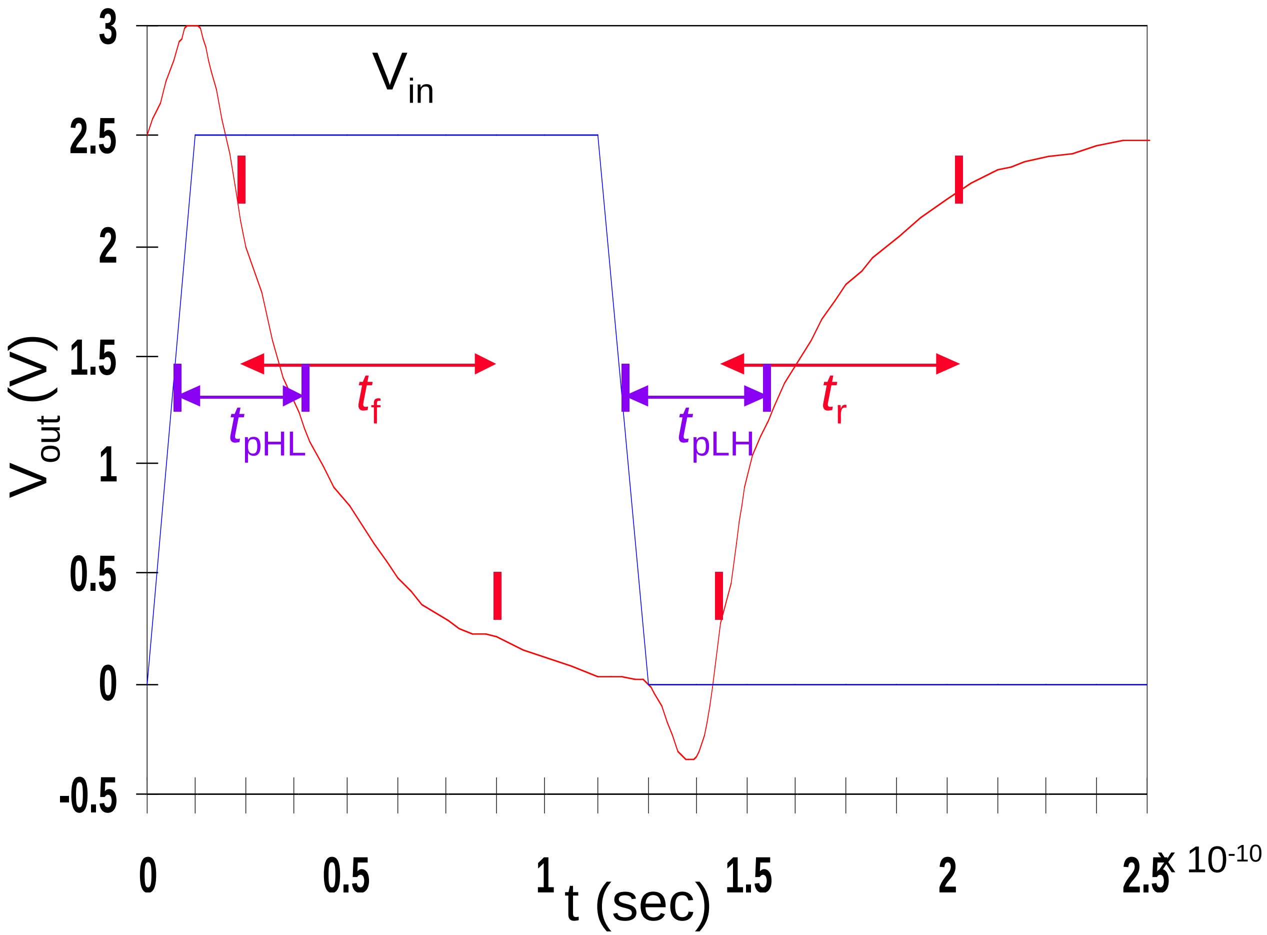
$$t_{PLH} = 0.7 \bullet \frac{R_p}{N} \bullet (N \bullet C_{oxp} + C_{load})$$

Series switching (all switch at same time):



$$t_{PHL} = 0.35 \bullet R_n C_{oxn} \bullet N^2 + 0.7 \bullet N \bullet R_n \bullet C_{load}$$

RC Delay, Two Inverters



- VDD=2.5V, 0.25mm
- $W/L_n = 1.5, W/L_p = 4.5$
- R_{eqn}= 13 kΩ (÷ 1.5)
- R_{eqp}= 31 kΩ (÷ 4.5)

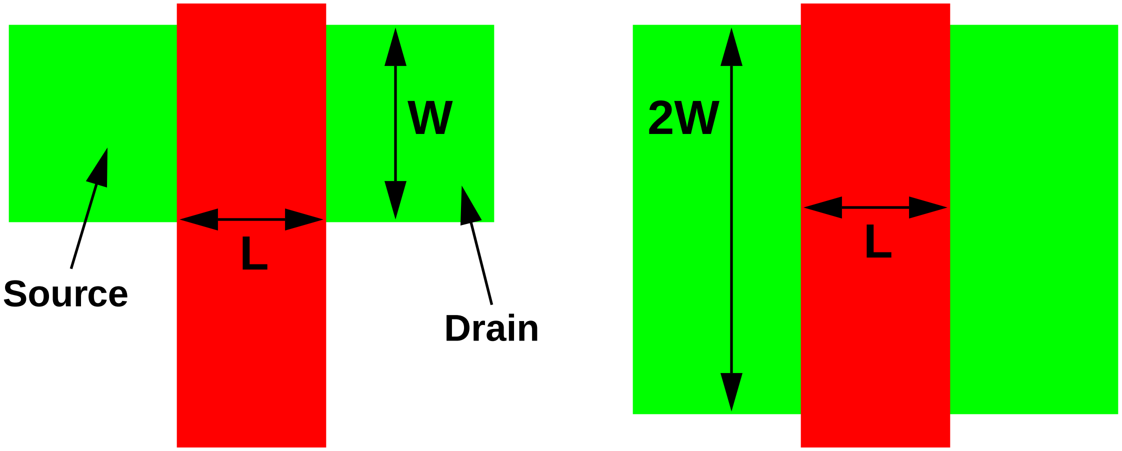
$t_{pHL} = 0.69\ R_nC = 36$ ps
$t_{pLH} = 0.69\ R_pC = 29$ ps

From SPICE simulation:
t_{pHL} = 39.9 ps, t_{pLH} = 31.7 ps

Bruce Jacob

University of
Maryland
ECE Dept.

Transistor Sizing I



The electrical characteristics of transistors determine the switching speed of a circuit

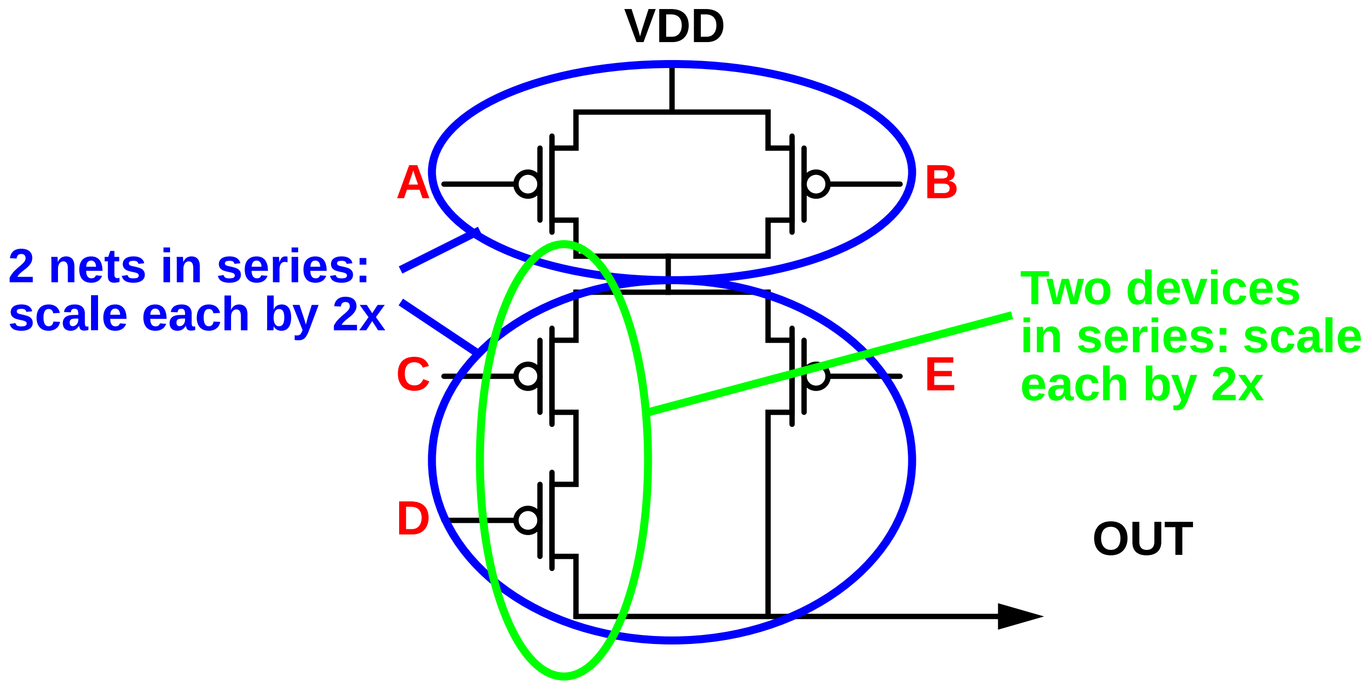
- Need to select the aspect ratios $(W/L)_n$ and $(W/L)_p$ of *every* FET in the circuit

Define *Unit Transistor* (R_1, C_1)

- L/W_{min}-> highest resistance (needs scaling)
- $R_2 = R_1 \div 2$ and $C_2 = 2 \bullet C_1$
- Separate nFET and pFET unit transistors
- Unit devices are *not* restricted to individual transistors

Sizing I: Complex Gates

Critical transistors: those in series



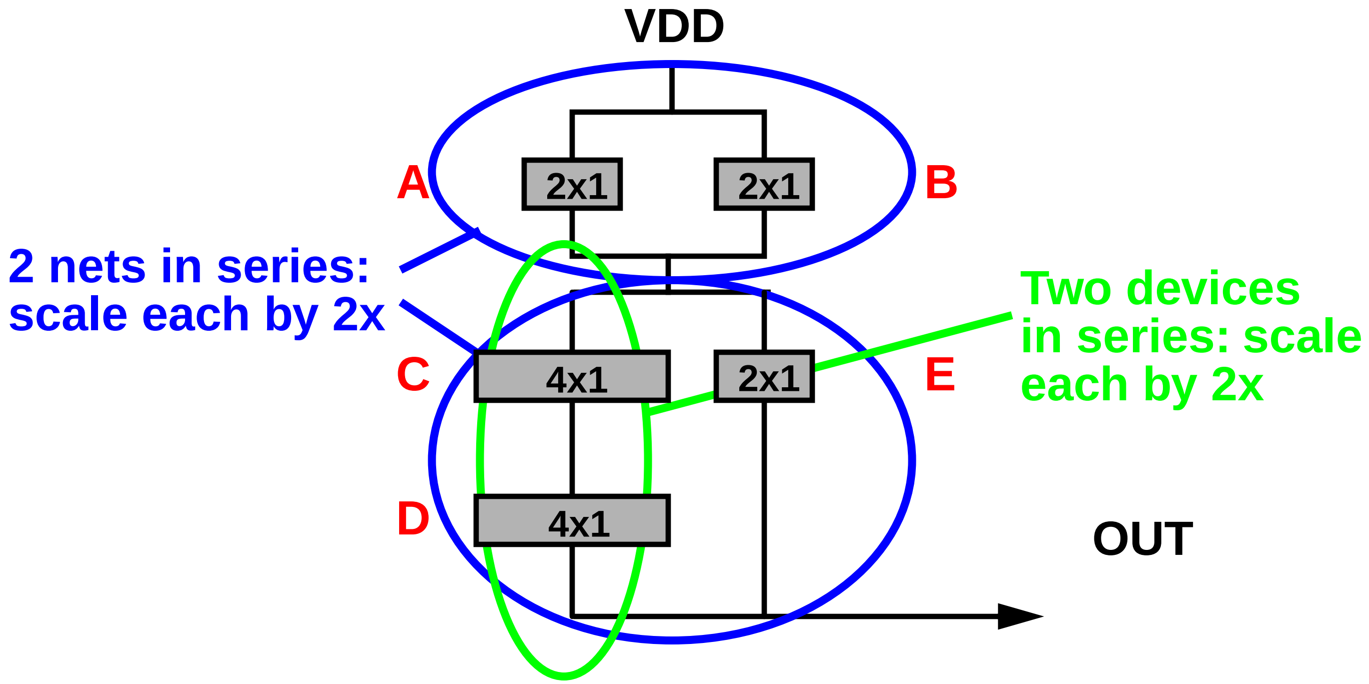
2 nets in series: scale each by 2x

Two devices in series: scale each by 2x

- N FETs in series => scale each by factor of N
- Ignore FETs in parallel (assume worst case: only 1 on)
- Ultimate goal: total resistance of net = 1 square

Sizing I: Complex Gates

Critical transistors: those in series



VDD

A

B

C

D

E

2x1

2x1

4x1

2x1

4x1

OUT

2 nets in series:
scale each by 2x

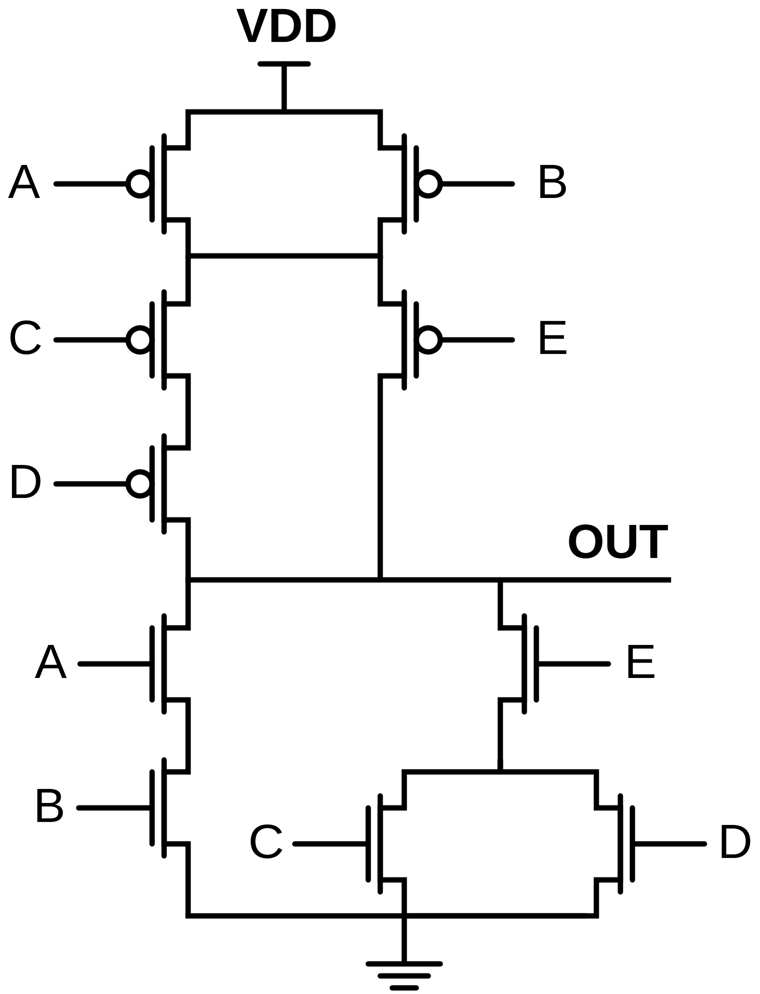
Two devices
in series: scale
each by 2x

- N FETs in series => scale each by factor of N
- Ignore FETs in parallel (assume worst case: only 1 on)
- Ultimate goal: total resistance of net = 1 square

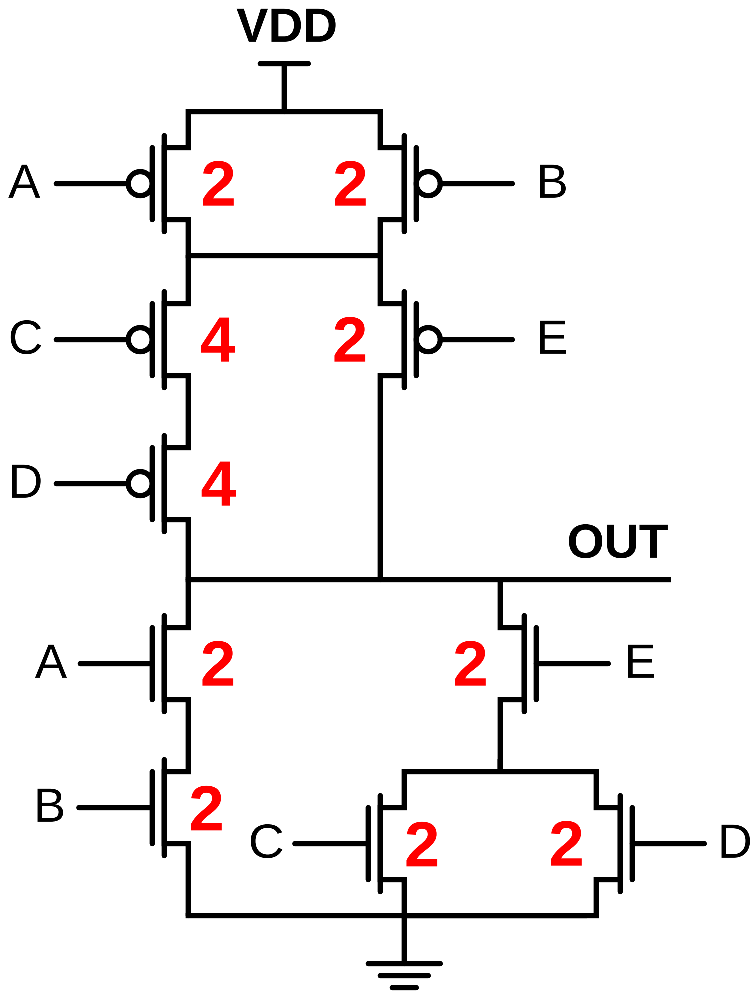
Bruce Jacob

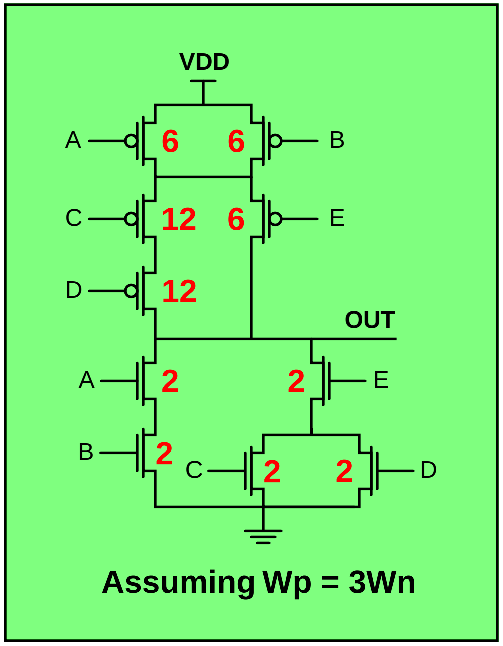
University of
Maryland
ECE Dept.

Examples



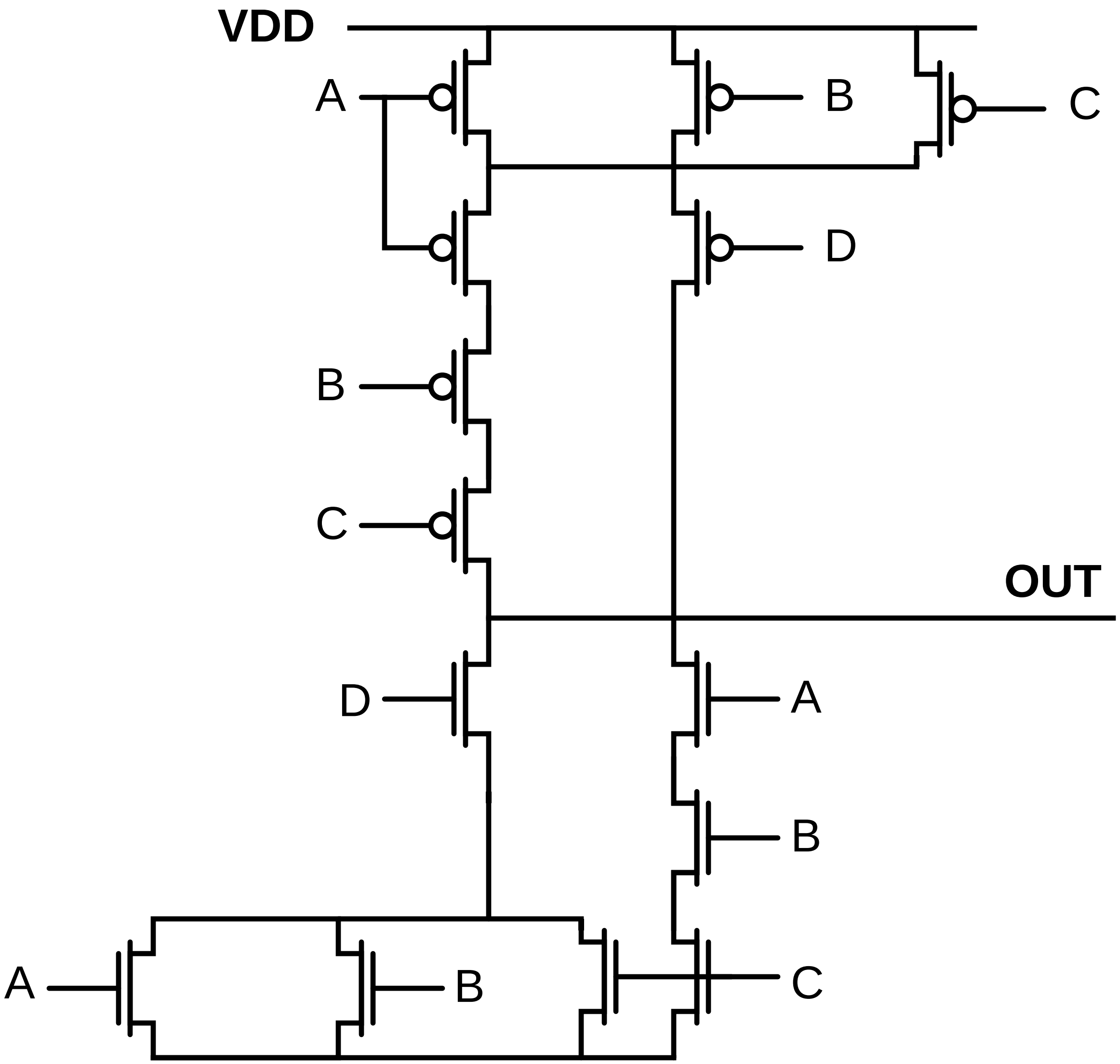
Examples



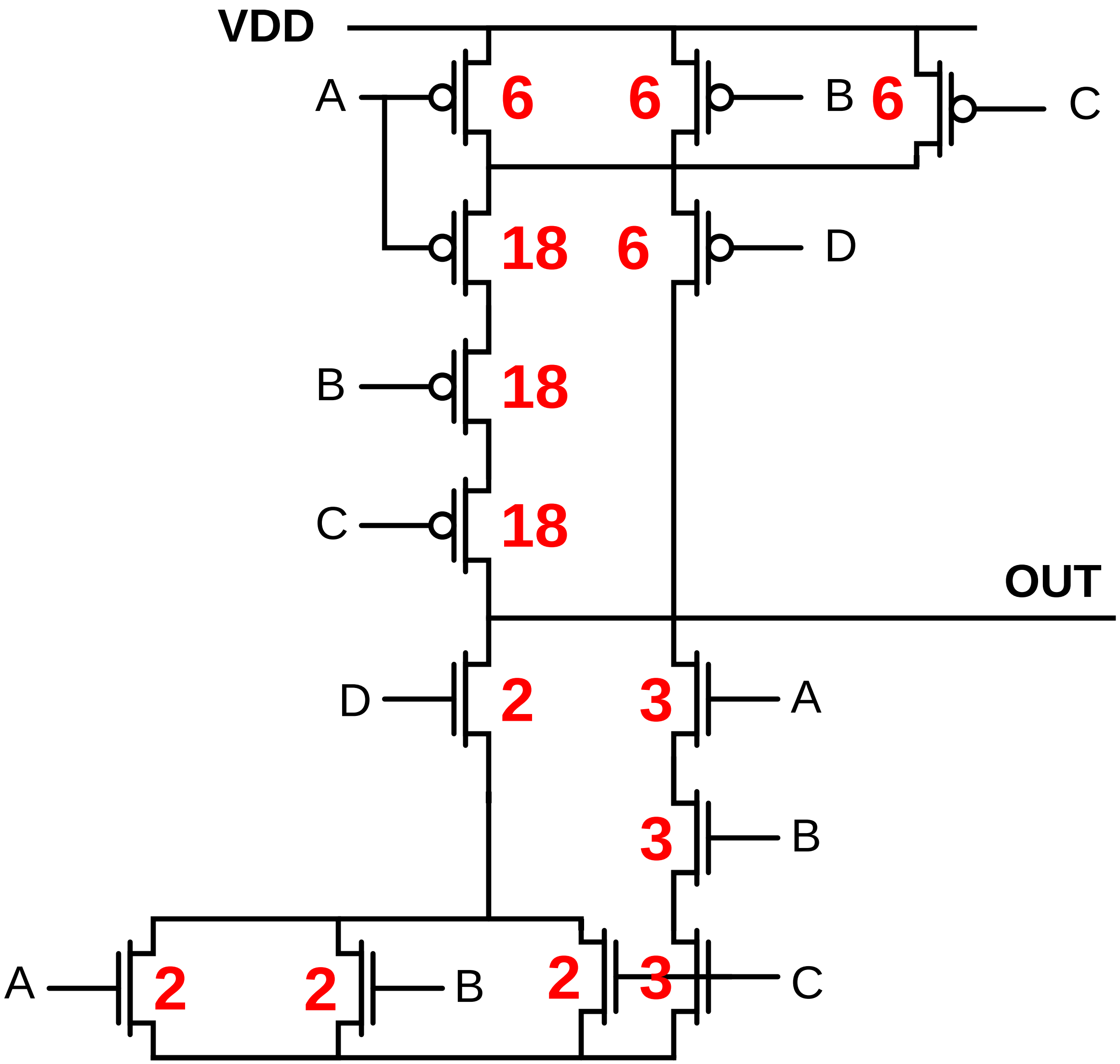


Assuming Wp = 3Wn

Examples



Examples



Ways to Improve Gate Delay

$$t_p \approx (t_{pHL} + t_{pLH}) \approx [C_L \div (k' W/L\ V_{DD})]$$

Reduce C_L

- internal diffusion capacitance of the gate itself (keep the drain diffusion as small as possible)
- other terms: interconnect capacitance & fanout

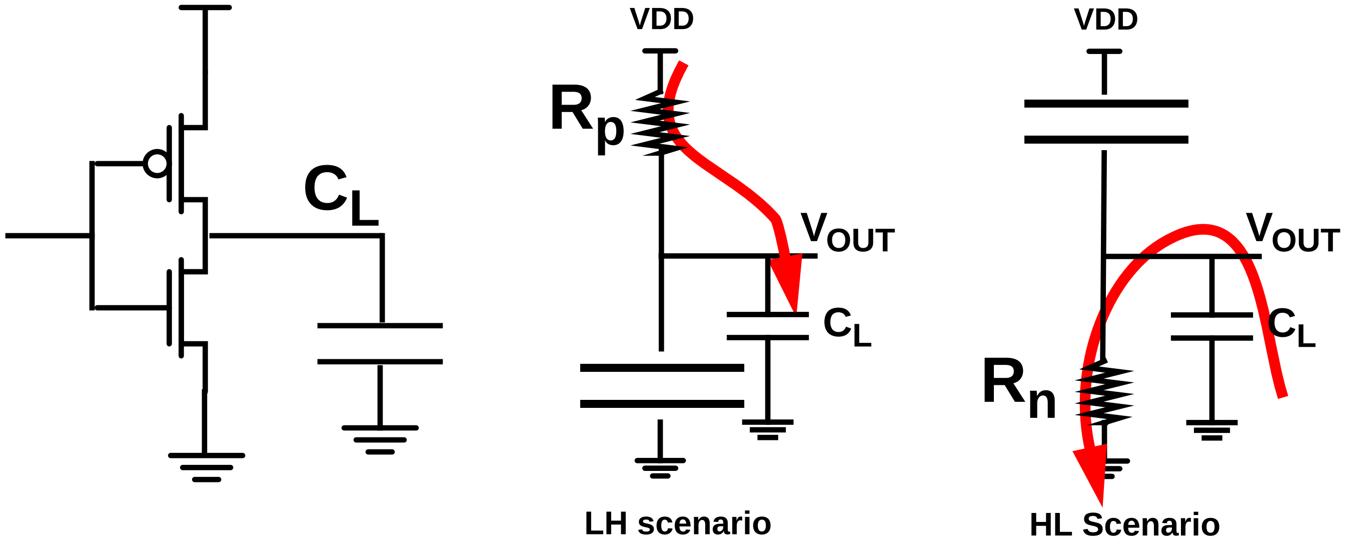
Increase W/L ratio of the transistor

- 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}

Gate Delay, Revisited



$$t_p \approx (t_{pHL} + t_{pLH}) \approx 0.7R_{ref}C_{ref}\ (1 + C_{ext}/SC_{iref})$$

- widening the PMOS improves t_{pLH} (R_p is lower) but degrades t_{pHL} (increases intrinsic capacitance G_{GD} and G_{DB})
- widening the NMOS improves t_{pHL} (R_n is lower) but degrades t_{pLH} (increases intrinsic capacitance G_{GD} and G_{DB})

Gate Delay, Revisited

So far have sized the PMOS and NMOS so that the R_{eq}'s match (ratio between 2 & 3.5)

- symmetrical VTC
- equal high-to-low and low-to-high propagation delays

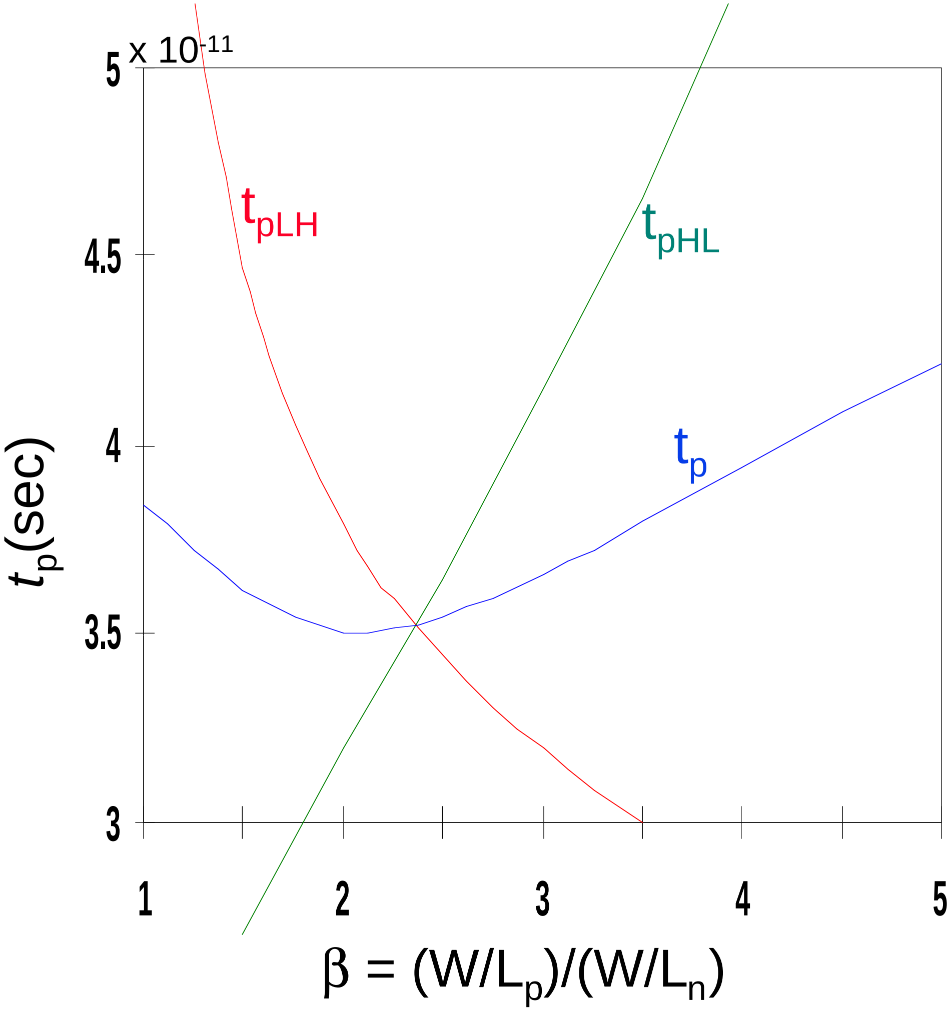
If speed is the only concern, reduce the width of the PMOS device!

- widening the PMOS degrades t_{pHL} due to larger parasitic capacitance (intrinsic capacitance)

$B = (W/L_p)/(W/L_n)$

- $r = R_{eqp}/R_{eqn}$ (resistance ratio of identically-sized PMOS and NMOS)
- $B_{opt} \approx \sqrt{r}$ if wiring capacitance negligible

Gate Delay, Revisited



- ß of 2.4 (R_p/R_n = 31 kΩ/13 kΩ) [what we’ve looked at] gives symmetric response
- ß of 1.6 to 1.9 gives optimal performance

Inverter Delay

$$t_p = 0.7R_{ref}C_{ref}\,(1 + C_{ext}/SC_{iref})$$

$$C_{int} = \gamma C_g$$

$$t_p = t_{p0}\left(1 + \frac{C_{ext}}{\gamma C_g}\right)$$

$$t_p = t_{p0}\left(1 + \frac{f}{\gamma}\right)$$

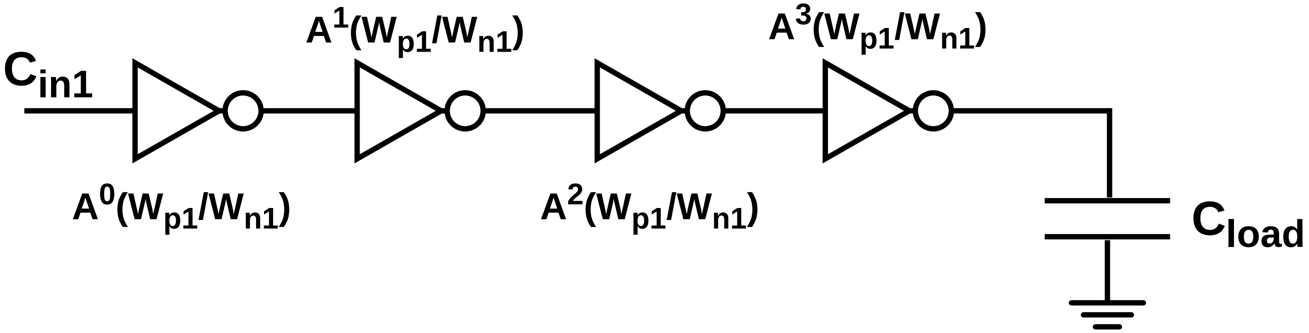
Propagation time is function of ratio of external to internal capacitance

This ratio is called fan-out, f

Gamma term is function of technology, $\gamma \approx 1$

Sizing & Big Gates

Sizing for Large Capacitive Loads



Supose C_{load} large (e.g. off-chip wires)

- Scale each *inverter* (both FETs in the circuit) by a factor A (input capacitances scale by A)
- if input C to last inverter * A = C_{load}
(i.e., C_{load} looks like N+1th inverter) then we have:

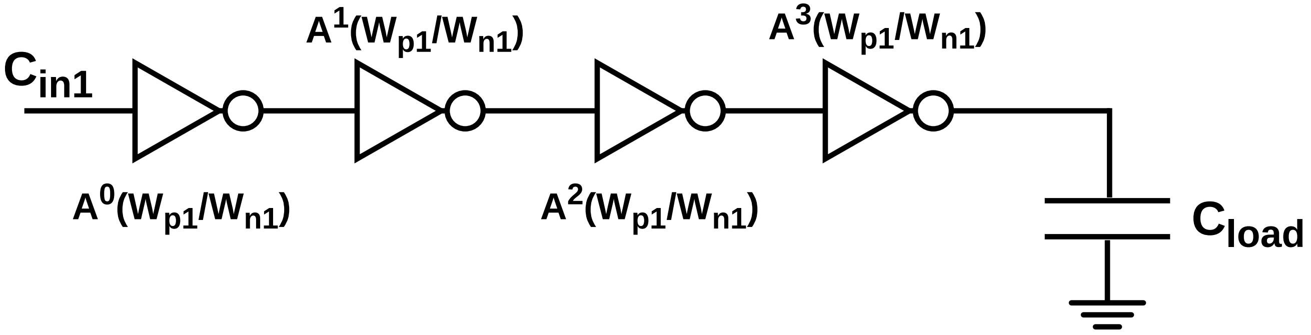
$$\text{Input C of last inverter} = C_{in1}\, A^N = C_{load}$$

- Rearranging:

$$A = [C_{load} \div C_{in1}]^{1/N}$$

Sizing & Big Gates

Sizing for Large Capacitive Loads



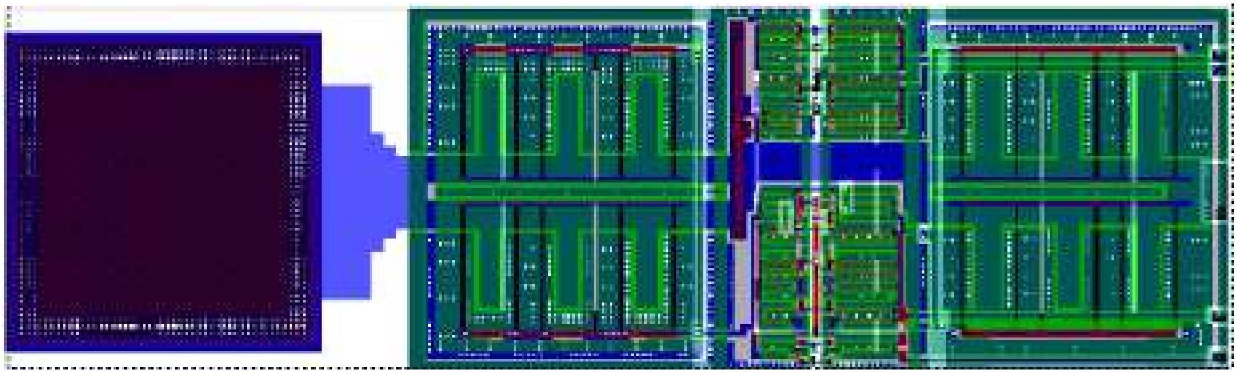
- Capacitances increase by factor of A left to right
- Resistances decrease by factor of A left to right
- Total delay ($t_{pHL} + t_{pLH}$):

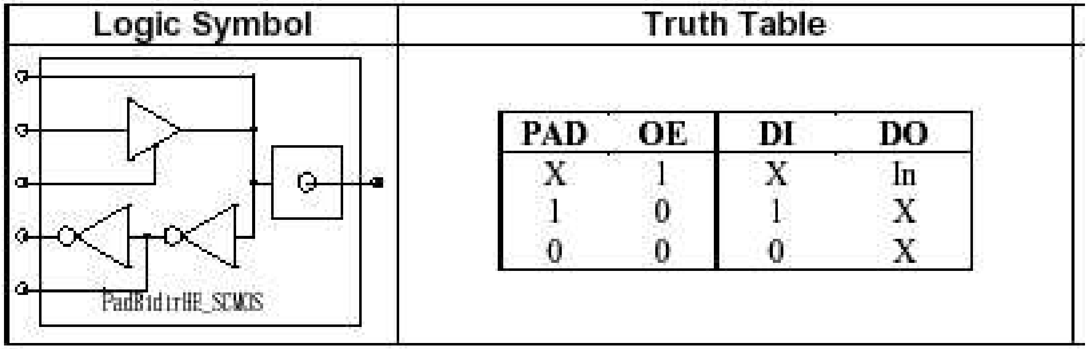
$$(R_{n1}+R_{p1}) \bullet (C_{out1}+AC_{in1}) +$$
$$(R_{n1}+R_{p1})/A \bullet (AC_{out1}+A^2C_{in1}) + \ldots$$
$$= N\ (R_{n1}+R_{p1}) \bullet (C_{out1}+AC_{in1})$$

- Find optimal chain length:

$$N_{opt} = \ln(C_{load} \div C_{in1})$$

Sizing & Big Gates





Logic Symbol

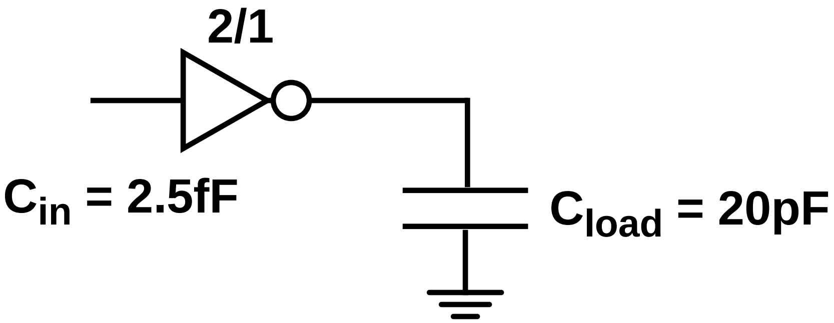
PadBidirHE_SCMOS

Truth Table

PAD	OE	DI	DO
X	1	X	In
1	0	1	X
0	0	0	X

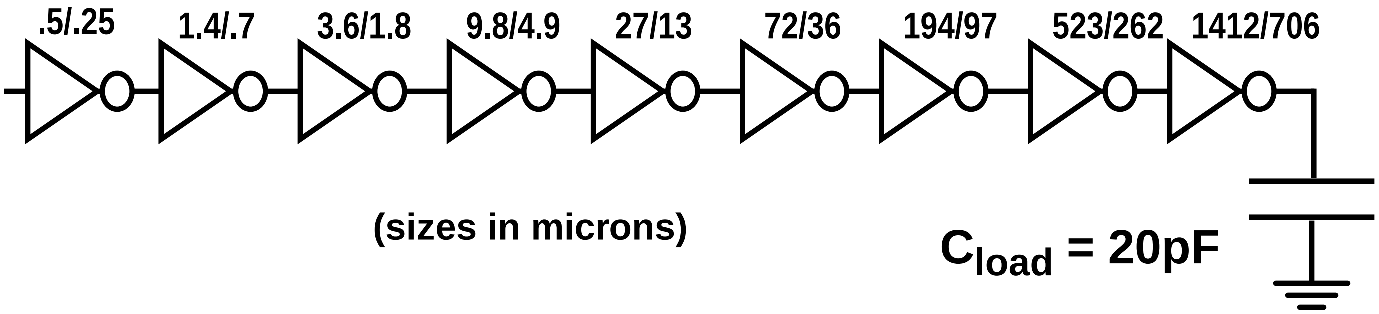
I/O Pad: large structures are ESD diodes and inverter chains (scale: pad is ~65 µm)

Example



Load is ~8000x that of single inverter’s input capacitance: find optimal solution.

Example



N_{opt} = ln(20pF/2.5fF) = 8.98 => 9 stages

Scaling factor A = $(20pF/2.5fF)^{1/9}$ = 2.7

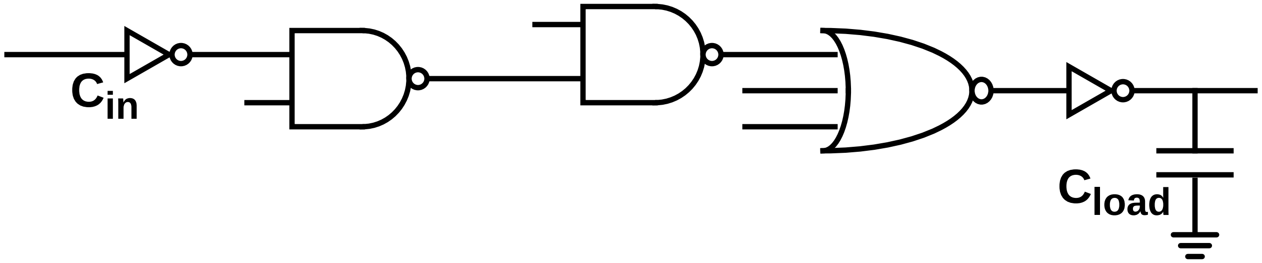
Total delay = $(t_{pHL} + t_{pLH})$
= $N (R_{n1}+R_{p1}) \bullet (C_{out1}+AC_{in1})$
= $N (R_{n1}+R_{p1}) \bullet (C_{out1} + [C_{load} \div C_{in1}]^{1/N} C_{in1})$

(assume $C_{in1} = 1.5C_{out1}$ = 2.5 fF)

= 9 • (31/9 + 13/3) • (1.85fF + 2.7 • 2.5fF)
= 602 ps (0.6 ns)

Generalize: *Logical Effort*

Want to find minimum delay for chains:



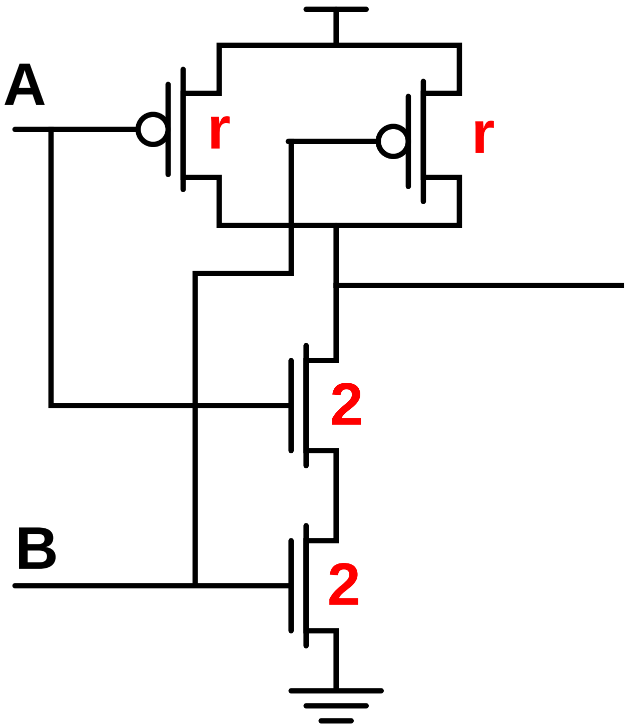
Main Points:

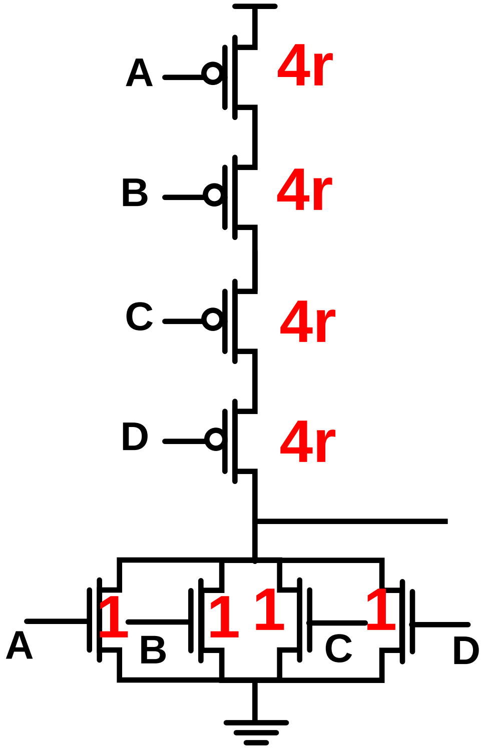
- **Path length is (maybe) fixed; find scaling**
- **Want constant scaling factor along path**
  [this gives same ***gate effort*** at each stage]
- **RC delay of a gate uses sum of internal C (its own C_{out}) and input of next gate (C_{in})**

Definitions

g = Gate-level logical effort

**= ratio of its input capacitance
to that of INVERTER**





$$g_{\text{nand}} = \frac{n + r}{1 + r}$$

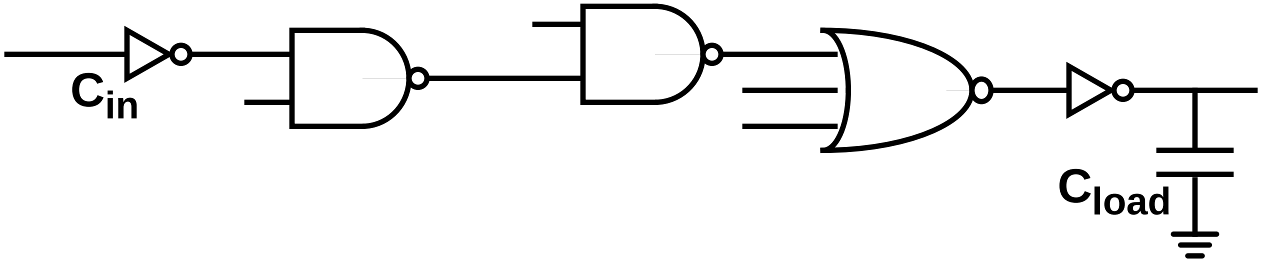
$$g_{\text{nor}} = \frac{1 + nr}{1 + r}$$

Definitions

Total Path Effort $H = GFB$

Optimal gate effort $h = \sqrt[N]{H}$

G = Path Logical Effort



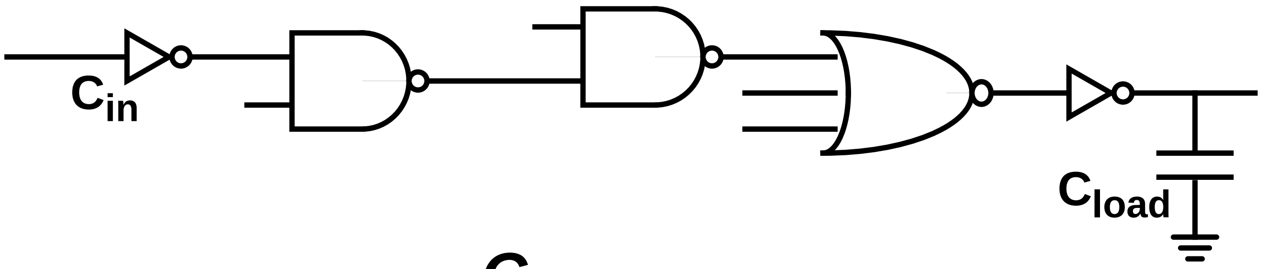
$$G_{path} = g_{inv} \cdot g_{nand} \cdot g_{nand} \cdot g_{nor} \cdot g_{inv}$$

Definitions

Total Path Effort H = GFB

Optimal gate effort $h = \sqrt[N]{H}$

F = Effective Fan-Out of Chain



$$F = \frac{C_{load}}{C_{in}}$$

Also called *Electrical Effort*

Bruce Jacob

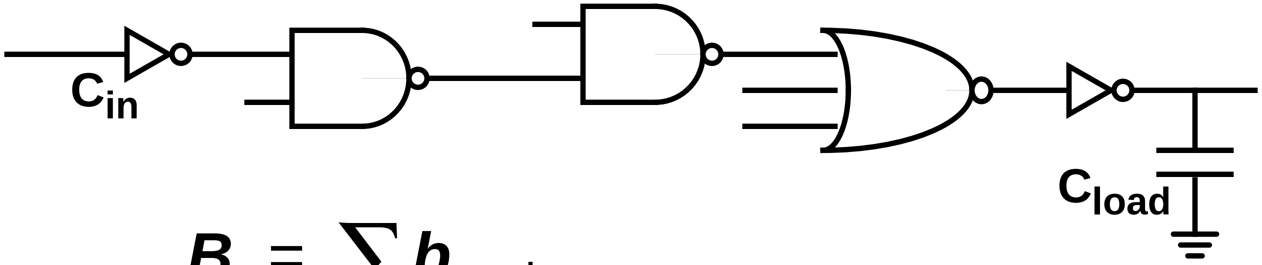
University of
Maryland
ECE Dept.

Definitions

Total Path Effort H = GFB

Optimal gate effort h = $\sqrt[N]{H}$

B = Path Branching Effort



$$B = \sum b_{\text{node}}$$

$$b_{\text{node}} = \frac{C_{\text{on-path}} + C_{\text{off-path}}}{C_{\text{on-path}}}$$

Definitions

Total Path Effort $H = GFB$

Optimal gate effort $h = \sqrt[N]{H}$

Redefine inverter delay:

$$t_p = t_{p0}\left(1 + \frac{f}{\gamma}\right) \quad \Rightarrow \quad t_p = t_{p0}\left(p + \frac{fg}{\gamma}\right)$$

Total delay through path:

$$D = t_{p0}\sum\left(p_i + \frac{f_i g_i}{\gamma}\right)$$

Minimum delay through path:

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

Definitions

Total Path Effort $H = GFB$

Optimal gate effort $h = \sqrt[N]{H}$

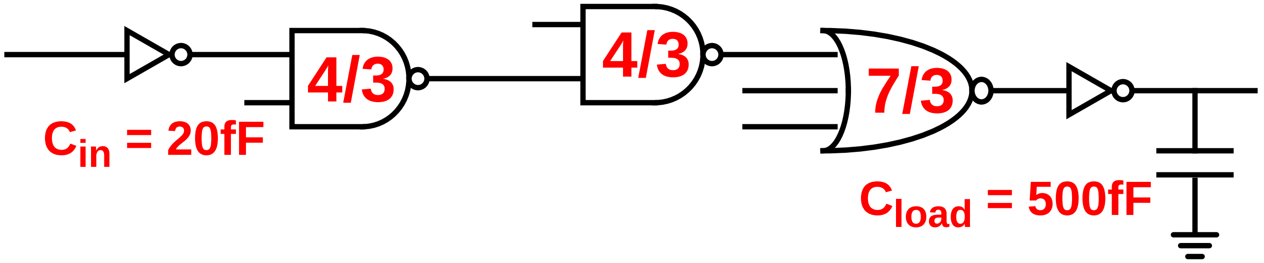
Gate effort $h_i = g_i f_i$

Sizing s_i for gate i in chain:

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

Analysis

Find minimum delay for chain (assume r=2):



G = (1)(4/3)(4/3)(7/3)(1) = 4.15

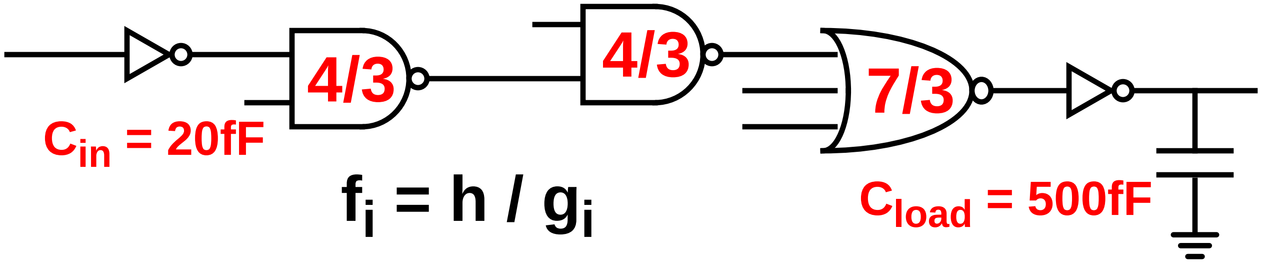
F = 500/20 = 25

B = 1 (no branching)

$$h = \sqrt[N]{H} = \sqrt[5]{103.75} = 2.53$$

Analysis

Find minimum delay for chain (assume r=2):



$f_i = h / g_i$

$f_1 = 2.53$
$f_2 = 2.53 \bullet 3/4 = 1.9$
$f_3 = 2.53 \bullet 3/4 = 1.9$
$f_4 = 2.53 \bullet 3/7 = 1.1$
$f_5 = 2.53$

$s_1 = 1$
$s_2 = f_1 \bullet g_1/g_2 = 1.9$
$s_3 = f_1 f_2 \bullet g_1/g_3 = 3.6$
$s_4 = f_1 f_2 f_3 \bullet g_1/g_4 = 3.9$
$s_5 = f_1 f_2 f_3 f_4 \bullet g_1/g_5 = 10.0$