EN1600
Design and Implementation of VLSI Systems
Fall 2016

Lectures 6 & 7: MOSFET Capacitance, Resistance, and CMOS Performance

Reading: Chapter 2, sections 3
Chapter 4, sections 4.3, 4.4, 4.5
Weste & Harris

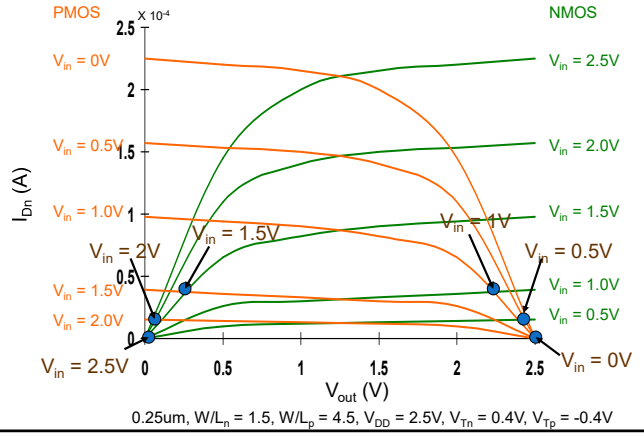
September 26 & 28, 2016
Prof. R. Iris Bahar

BROWN

Homework #1

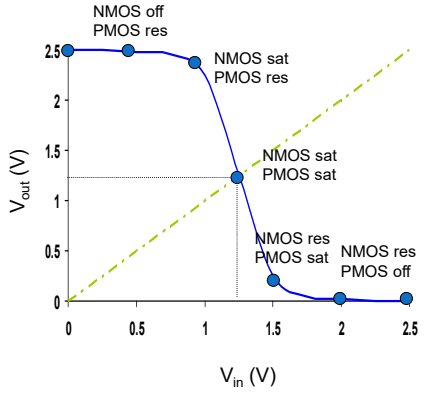
- Assignment is posted on the course webpage
  - Due Wednesday, Sept. 28
- Please keep in mind the following for your layouts:
  - Try to create a minimum physical implementation (without violating any design rules)
  - Avoid using multiple metals if possible (HW#1 only requires M1)
  - Keep metal and poly lines as short as possible.
  - Transistors of same type can share source/drain diffusion regions if they are physically connected.
  - Be consistent with your layout (e.g., orient transistors the same)
- Bad layout designs will lower your homework grade

CMOS inverter I-V curves



0.25um, W/L_n = 1.5, W/L_p = 4.5, V_{DD} = 2.5V, V_{Tn} = 0.4V, V_{Tp} = -0.4V

CMOS inverter VTC



Taking those intersection points from the load curves, we obtain the voltage-transfer characteristic

Switching threshold

- Define V_M to be the point where $V_{in} = V_{out}$ (both PMOS and NMOS in saturation since $V_{DS} = V_{GS}$)
- If $V_M = V_{DD}/2$, then this implies *symmetric rise/fall* behavior for the CMOS gate
- Recall at saturation, $I_D=(k'/2)(W/L)\,(V_{GS}-V_T)^2$,
  - where $k'_n= \mu_n C_{ox}= \mu_n \varepsilon_{ox}/t_{ox}$
- Setting $I_{Dp}= -I_{Dn}$

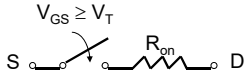
$$\frac{k'_n}{2}\frac{W_n}{L_n}(V_M - V_{Tn})^2 = \frac{k'_p}{2}\frac{W_p}{L_p}(-V_M - V_{Tp})^2$$

- Assuming $V_{Tn}=-V_{Tp}$

$$\frac{W_p/L_p}{W_n/L_n} = \frac{k'_n}{k'_p} = \frac{\mu_n}{\mu_p} = \frac{500}{180} = 2.78$$

MOS Structure Resistance

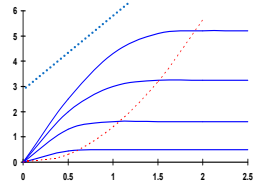
- The simplest model assumes the transistor is a switch with an infinite "off" resistance and a finite "on" resistance R_{on}

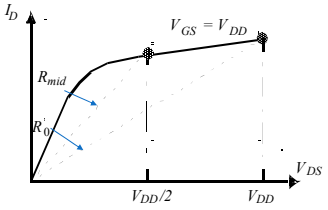


- However R_{on} is nonlinear, time-varying, and dependent on the operation point of the transistor
- *How can we determine an equivalent (constant and linear) resistance to use instead?*

MOS structure resistance

- Approximate R_{on} as the resistance found during linear operation
  - Simple to calculate but limited accuracy
- Instead use the average value of the resistances, R_{eq}, at the end-points of the transition (i.e., V_{DD} and $V_{DD}/2$)





$$R_{eq} \approx \frac{1}{2}(R_{on}(t_1) + R_{on}(t_2))$$
$$\approx \frac{1}{2}\left(\frac{V_{DD}}{I_{DSAT}} + \frac{V_{DD}/2}{I_{DSAT}}\right)$$
$$\approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}$$

Equivalent MOS Structure Resistance

$$R_{eq} = \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}$$

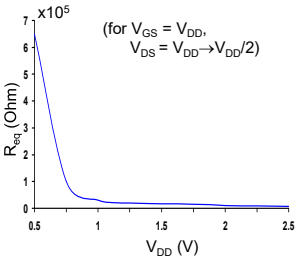
where

$$I_{DSAT} = \frac{\mu C_{ox}}{2}\frac{W}{L}\left((V_{DD} - V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2}\right)$$

so,

$$R_{eq} = \frac{3}{2\mu C_{ox}}\frac{L}{W}\frac{V_{DD}}{(V_{DD} - V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2}}$$

$$R_{eq_{NMOS}} \propto \frac{L}{\mu_n W}, \qquad R_{eq_{PMOS}} \propto \frac{L}{\mu_p W}$$

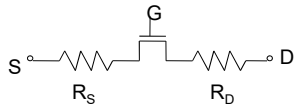


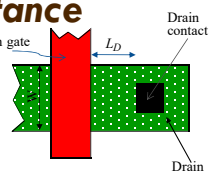
R_{eq} is essentially independent of V_{DD} as long as $V_{DD} >> V_T+V_{DSAT}/2$

Equivalent MOS Structure Resistance

- What insight can we gain from the previous slides?
  - The on resistance is proportional to L/W.
    - ➡ Doubling W halves R_{eq}
  - For $V_{DD} >> V_T+V_{DSAT}/2$, R_{eq} is independent of V_{DD}. Minor reduction in R_{eq} with increase in V_{DD} is due to channel length modulation.
  - Once the supply voltage approaches V_T , R_{eq} increases dramatically

Source and drain resistance





$R_{S,D} = (L_{S,D}/W)R_{\square}$

where

$L_{S,D}$ is the length of the source or drain diffusion

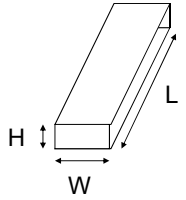
$R_{\square}$ is the sheet resistance of the source or drain diffusion (20 to 100 Ω/❐)

- Can cover with low-resistivity material (silicide) to reduce to the range of 1 to 4 Ω/❐

Contact resistance

- Transitions between routing layers (contacts through vias) add extra resistance to a wire
  - keep signal wires on a single layer whenever possible
  - avoid excess contacts or vias
- Typical contact resistances, R_C, (minimum-size)
  - 5 to 20 Ω for metal or poly to n+, p+ diffusion and metal to poly
  - 1 to 5 Ω for metal to metal contacts

Wire resistance

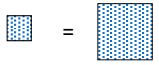


resistivity

$$R = \frac{\rho L}{A} = \frac{\rho}{H} \frac{L}{W}$$

Sheet Resistance $R_{\square}$

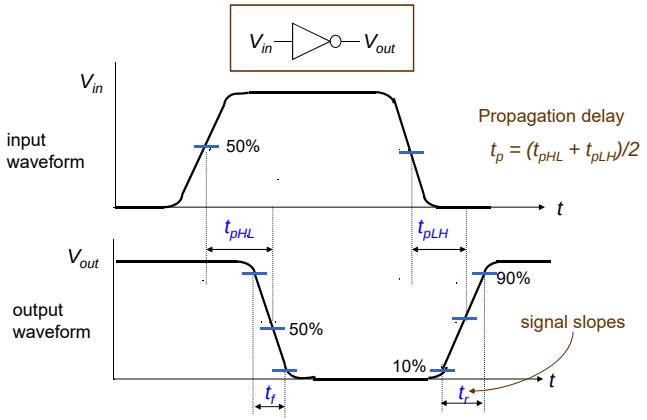
$R_{1\square} = R_{2\square}$



Material	ρ(Ω-m)
Silver (Ag)	1.6×10^{-8}
Copper (Cu)	1.7×10^{-8}
Gold (Au)	2.2×10^{-8}
Aluminum (Al)	2.7×10^{-8}
Tungsten (W)	5.5×10^{-8}

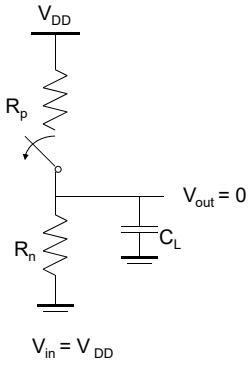
Material	Sheet Res. (Ω/❐)
n, p well diffusion	1000 to 1500
n+, p+ diffusion	50 to 150
n+, p+ diffusion with silicide	3 to 5
polysilicon	150 to 200
polysilicon with silicide	4 to 5
Aluminum	0.05 to 0.1

Delay definitions



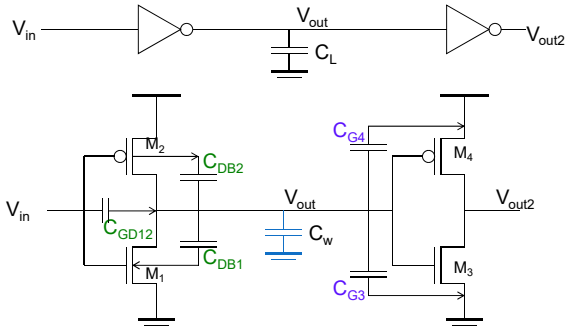
CMOS inverter: dynamic behavior

- Transient, or dynamic, response determines the maximum speed at which a device can be operated.



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

Sources of capacitance

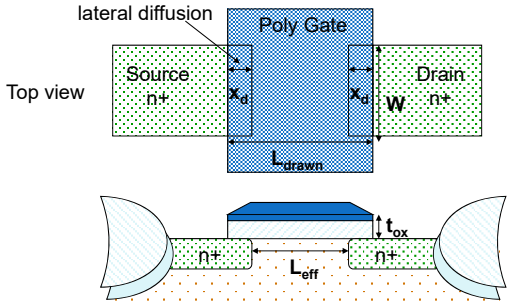


- intrinsic MOS transistor capacitances
- extrinsic MOS transistor (fanout) capacitances
- wiring (interconnect) capacitance

MOS intrinsic capacitances

- Structure capacitances
- Channel capacitances
- Diffusion capacitances

MOS structure capacitances



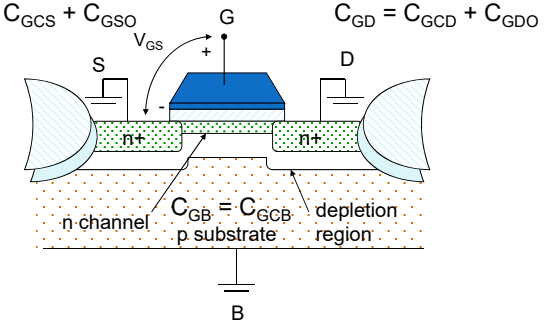
Overlap Capacitance (linear)

$$C_{GSO} = C_{GDO} = C_{ox}\, x_d\, W = C_o\, W$$

MOS channel capacitances

- The gate-to-channel capacitance depends upon the operating region and the terminal voltages

$C_{GS} = C_{GCS} + C_{GSO}$ $\quad$ $C_{GD} = C_{GCD} + C_{GDO}$



Variations in channel capacitance

- Cutoff (i.e., subthreshold) $V_{GS} \leq V_T$
  - No channel exists: C_{GC} only appears between gate and body
    - $C_{GCB} = WL\, C_{ox}$
- Strong Inversion $V_{GS} > V_T$
  - Linear (Resistive) $V_{DS} < V_{DSAT} = V_{GS} - V_T$
    - Inversion layer acts as conductor between source, drain
    - $C_{GCB} = 0$, C_{GC} distributes between S, D
      → $C_{GCS} = C_{GCD} = WL\, C_{ox}/2$
  - Saturated $V_{DS} \geq V_{DSAT} = V_{GS} - V_T$
    - $C_{GCB} = C_{GCD} = 0$, C_{GC} only appears between gate and source
      → $C_{GCS} = WL\, C_{ox}$ *(actually, it's closer to 2/3 WL C_{ox})*

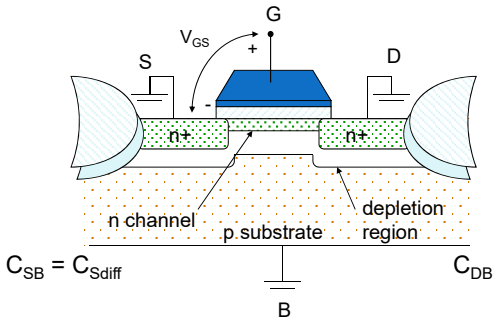
Average channel capacitance

Operation Region	C_{GCB}	C_{GCS}	C_{GCD}	C_{GC}	C_G
Cutoff	$C_{ox}WL$	0	0	$C_{ox}WL$	$C_{ox}WL + 2C_oW$
Resistive	0	$C_{ox}WL/2$	$C_{ox}WL/2$	$C_{ox}WL$	$C_{ox}WL + 2C_oW$
Saturation	0	$(2/3)C_{ox}WL$	0	$(2/3)C_{ox}WL$	$(2/3)C_{ox}WL + 2C_oW$

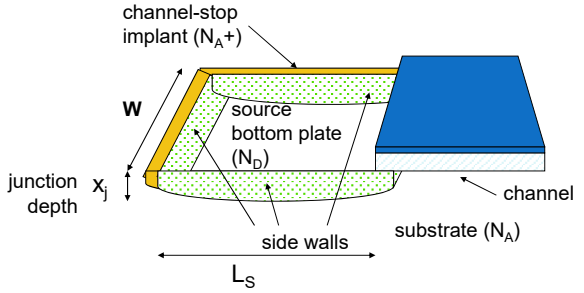
- Channel capacitance components are nonlinear and vary with operating voltage
- Most important regions are cutoff and saturation since that is where the device spends most of its time
- Gate capacitance is non-linear, but we can approximate with piecewise-linear model

MOS diffusion capacitances

- The junction (or diffusion) capacitance is from the reverse-biased source-body and drain-body pn-junctions.



Source junction view

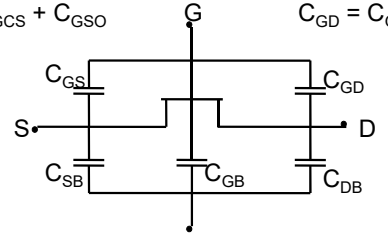


$C_{diff} = C_{bp} + C_{sw} = C_j\,AREA + C_{jsw}\,x_j\,PERIMETER$
$= C_j L_S W + C_{jsw}\,x_j(2L_S + W)$

NOTE: C_j and C_{jsw} are given in capacitance per unit area (e.g., F/m^2)

MOS Capacitance Model

$C_{GS} = C_{GCS} + C_{GSO}$ $C_{GD} = C_{GCD} + C_{GDO}$



$C_{SB} = C_{Sdiff}$ $C_{DB} = C_{Ddiff}$

$C_{GB} = C_{GCB}$

Extrinsic (fanout) capacitance

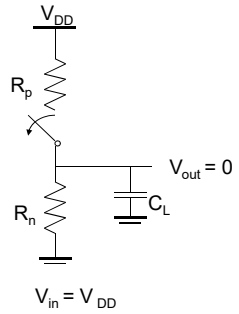
- The extrinsic, or fanout, capacitance is the total gate capacitance of the loading transistors M3 and M4.

$C_{fanout} = C_{gate}$ (NMOS) + C_{gate} (PMOS)
$= (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) +$
$(C_{GSOp} + C_{GDOp} + W_p L_p C_{ox})$

- Simplification of the actual situation
  - Assumes all the components of C_{gate} are between V_{out} and GND (or V_{DD})
  - Assumes the channel capacitances of the loading gates are constant

Inverter propagation delay

- Propagation delay is proportional to the time-constant of the network formed by the pull-down resistor and the load capacitance.



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

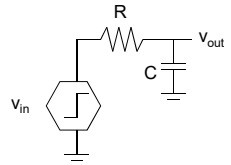
$$R_n = R_{eq} = \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}$$

$$C_L = C_{int} + C_{ext} + C_{wire}$$

Want to have equal rise/fall delays
➡ Make $R_n = R_p$

Modeling propagation delay

- Model circuit as first-order RC network



$v_{out}(t) = (1 - e^{-t/\tau})V_{in}$

where $\tau = RC$

Time to reach 50% point is
$t = \ln(2)\ \tau = 0.69\ \tau$

Time to reach 90% point is
$t = \ln(9)\ \tau = 2.2\ \tau$

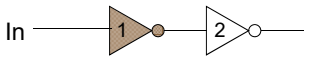
$$t_p = (t_{pHL} + t_{pLH})/2 = 0.69 C_L (R_n + R_p)/2$$

How can we improve performance?

- Reduce R_n, R_p
  - Increase W/L ratio of the transistor
    - most powerful and effective performance optimization tool for designer
    - But what happens to the intrinsic capacitance with larger W/L?
- Reduce C_L
  - Keep drain diffusions small
  - Limit interconnect capacitance
  - Limit fanout
- Increase V_{DD}
  - Trade off energy for performance
  - Increase V_{DD} above a certain level yields minimal improvements
  - Reliability concerns enforce an upper bound on V_{DD}

NMOS/PMOS ratio

- So far we have sized the PMOS and NMOS so that the R_{eq} values match (i.e., $\beta = (W/L_p)/(W/L_n) = W_p/W_n = 2$ to 2.8)
  - Symmetric 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 intrinsic capacitance
- What does this imply if we want to minimize delay for an inverter?



NMOS/PMOS ratio

- We define propagation delay as:

$$t_p = \frac{t_{pHL} + t_{pLH}}{2} = 0.69 C_L \left(\frac{R_{eqn} + R_{eqp}/\beta}{2} \right)$$

- And define

$$C_L = (C_{dp1} + C_{dn1}) + (C_{gp2} + C_{gn2}) + C_W$$

$$C_L \approx (1+\beta)(C_{dn1} + C_{gn2}) + C_W$$

- So we have

$$t_p = 0.345((1+\beta)(C_{dn1} + C_{gn2}) + C_w)(R_{eqn} + R_{eqp}/\beta)$$

$$t_p = 0.345\left((1+\beta)(C_{dn1} + C_{gn2}) + C_w\right)R_{eqn}\left(1 + \frac{r}{\beta}\right), \quad \text{where } r = \frac{R_{eqp}}{R_{eqn}}$$

- Now, optimize t_p with respect to β…

NMOS/PMOS ratio

- Given the equation for t_p:

$$t_p = 0.345\left((1+\beta)(C_{dn1} + C_{gn2}) + C_w\right)R_{eqn}\left(1 + \frac{r}{\beta}\right)$$

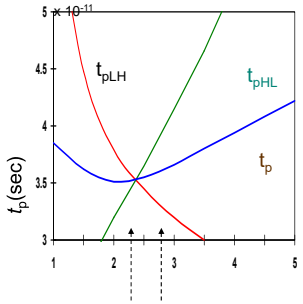
- Minimize t_p as a function of β…
- Compute the optimal value of β by setting $\partial t_p/\partial\beta = 0$

$$\beta_{opt} = \sqrt{r\left(1 + \frac{C_W}{C_{dn1} + C_{gn2}}\right)}$$

Where $r = R_{eqp}/R_{eqn}$ = resistance ratio for identically sized PMOS, NMOS

$$\beta_{opt} = \sqrt{r} \text{ when } C_W \text{ is negligible}$$

PMOS/NMOS ratio effects



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

β of 2.4 gives symmetrical response

β of 1.6 to 1.9 gives optimal performance

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

Device sizing for performance

- Divide capacitive load, C_L, into
  - C_{int} : intrinsic → diffusion (and gate-drain overlap cap (Miller cap))
  - C_{ext} : extrinsic → fanout (gate-channel cap and wiring)

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

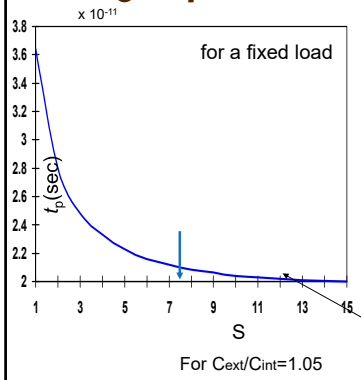
$$t_p = 0.69\ R_{ref}\ C_{iref}\ (1 + C_{ext}/(SC_{iref})) = t_{p0}(1 + C_{ext}/(SC_{iref}))$$

Device sizing for performance

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

- What can we conclude from this?
  - 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



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

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

- Extrinsic capacitance, C_{ext} , is a function of the gates being driven by the gate under question (i.e. the fanout)

  larger fanout ➡ larger external load.

- Re-express the intrinsic capacitance (C_{int}) in terms of input gate capacitance:

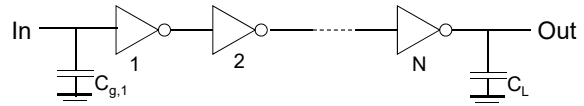
$$C_{int} = \gamma C_g\ , \quad \text{where } \gamma \approx 1$$

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

$f = C_{ext}/C_g$ is the effective fanout

Inverter chain

- 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 $t_p = t_{p,1} + t_{p,2} + \ldots + t_{p,N}$

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

- If C_L and $C_{g,1}$ are given, we have 2 different optimizations
  - How should the inverters be sized to minimize delay?
  - How many stages are needed to minimize the delay?

Sizing the inverters in the chain

- After a bit of calculus, we find that for minimum delay:

$$C_{g,j+1}/C_{g,j} = C_{g,j}/C_{g,j-1} \quad \text{for } j=2\ldots N$$

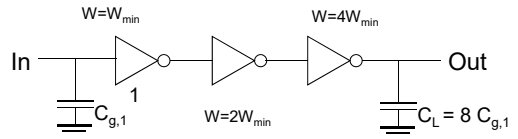
- What does this imply?
  - All gates have the same effective fanout, *f*
  - Each gate should be scaled up by the same factor w.r.t. its preceding gate
- What is the effective fanout for a gate given C_L and $C_{g,1}$?
  - With a bit of algebra and inductive reasoning we find that:

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

  - $F = C_L/C_{g,0}$ is the overall effective fanout
- What is the minimum delay through the chain?

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

Example: inverter chain sizing



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

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

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

- Assuming L remains unchanged for all inverters, *f* is obtained by adjusting W relative to the previous stage (i.e., scale up by factor of 2 relative to the previous gate).

Optimal number of inverters

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

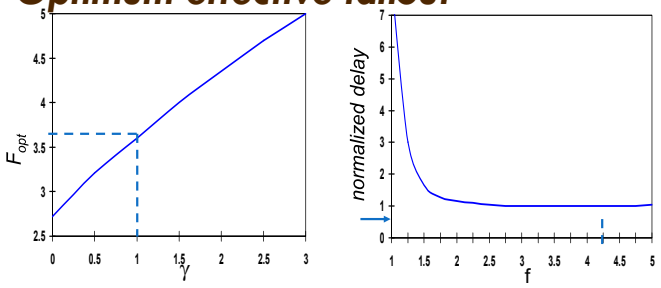
$$t_p = Nt_{p0}(1 + \sqrt[N]{F}/\gamma)$$

$$\partial t_p / \partial N = \gamma + \sqrt[N]{F} - \frac{\sqrt[N]{F}\ln F}{N} = 0$$

$$\rightarrow f = e^{(1+\gamma/f)}$$

- For $\gamma = 0$ (ignoring self-loading) $N = \ln(F)$ and the effective-fan out (tapering factor) becomes $f = e = 2.718$
- For $\gamma = 1$ (the typical case) the optimum effective fan-out can be solved numerically and turns out to be close to 3.6

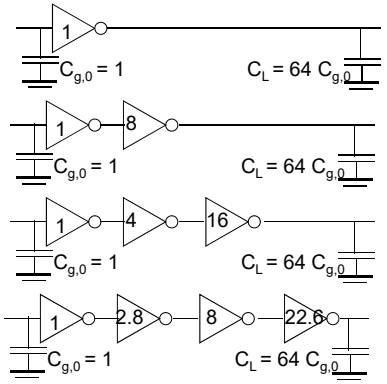
Optimum effective fanout



- Too many stages has a substantial negative impact on delay
- Choosing f slightly larger than optimum has little effect on delay and reduces the number of stages (and area).
  - Common practice to use f = 4 (for $\gamma = 1$)

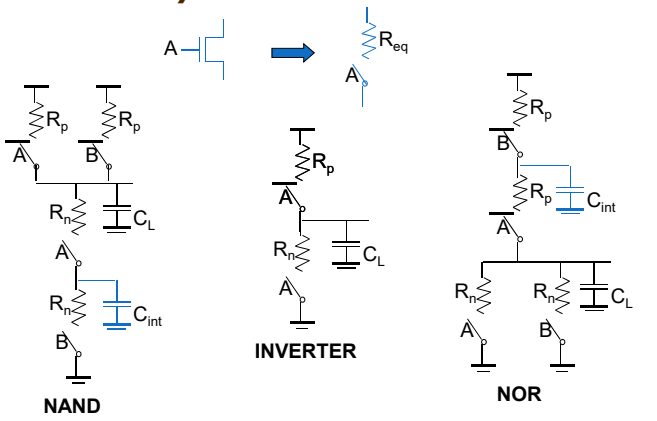
➡ Fanout of 4 (FO4) rule of thumb delay metric is based on this result

Example of inverter (buffer) staging

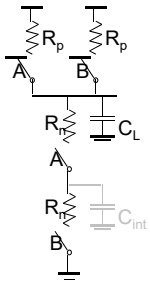


N	f	t_p
1	64	65
2	8	18
3	4	15
4	2.8	15.3

Switch delay model



Input pattern effects on delay

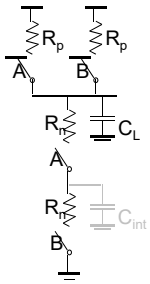


- Delay is dependent on the pattern of inputs
- 1st order approximation of delay:

  $t_p \approx 0.69\ R_{eff}\ C_L$

- R_{eff} depends on the input pattern

Input pattern effects on delay



- 0→1 transition on output: 2 possibilities
  - one input goes low: what is R_{eff} ?
    - delay is 0.69 R_p C_L
  - both inputs go low: what is R_{eff} ?
    - delay is 0.69 $R_p/2$ C_L since two p-resistors are on in parallel
- 1→0 transition on output: 1 possibility
  - both inputs go high
    - delay is 0.69 $2R_n$ C_L

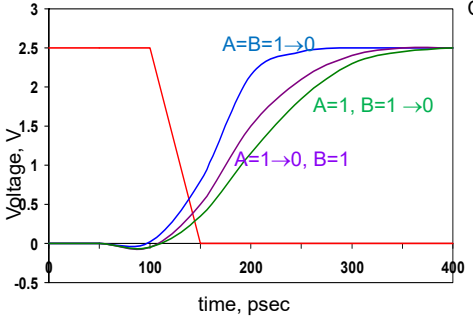
$t_p \approx 0.69\ R_{eff}\ C_L$

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

BROWN

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

BROWN

Transistor sizing

- How should NMOS and PMOS devices be sized relative to an inverter with equal rise/fall times?