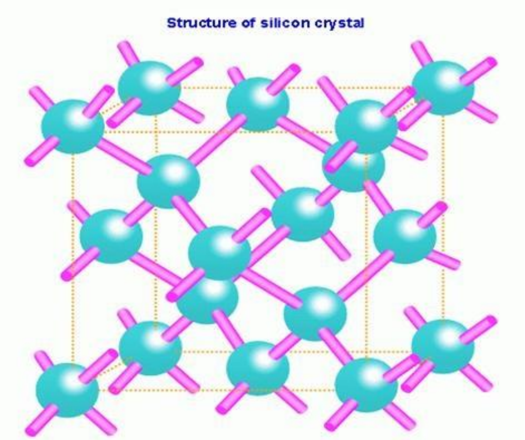
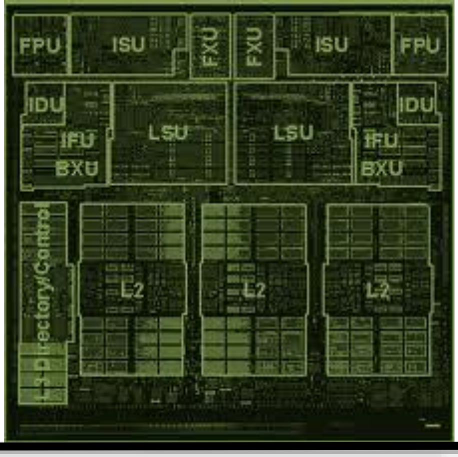
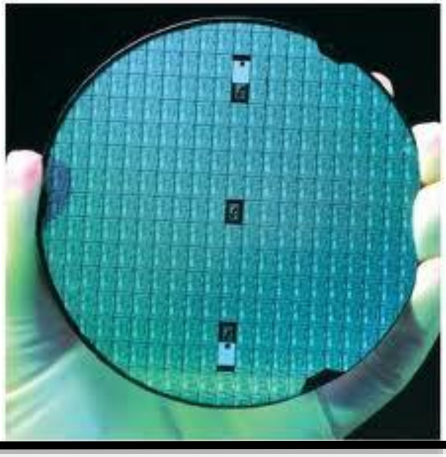






ECE 122A

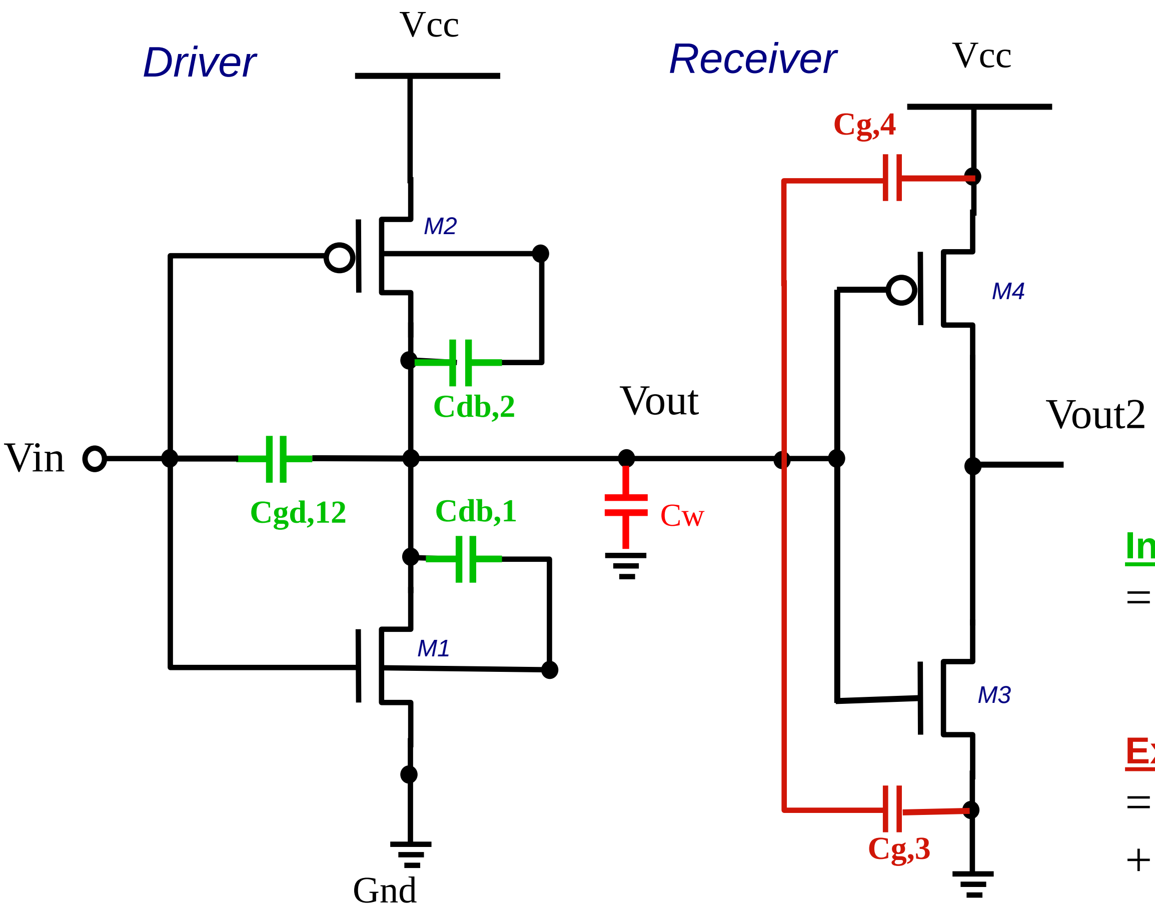
VLSI Principles

Lecture 10

Prof. Kaustav Banerjee
Electrical and Computer Engineering
University of California, Santa Barbara
E-mail: kaustav@ece.ucsb.edu

Inverter Sizing

Load capacitances



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

Internal Caps of Driver (Cint):

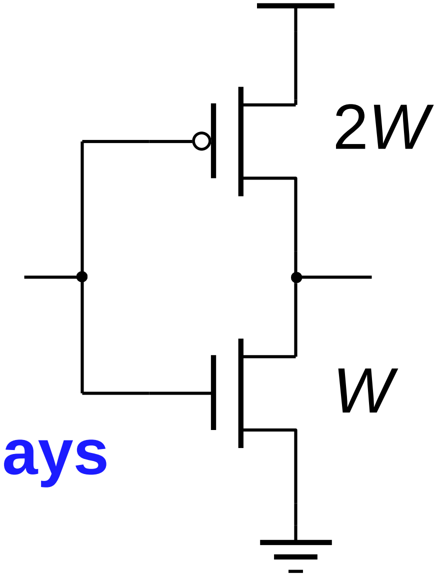
= Junction caps: $C_{db,12}$ +
Gate caps: $C_{gd,12}$ (including Miller Caps.)

External Caps (Cext):

= Interconnect cap: C_w
+ Receiver gate caps: $C_{g,43}$

Inverter Delay

- Minimum length devices, L=0.25µm
- Assume that for $W_P = 2W_N = 2W$
  - same pull-up and pull-down currents
  - approx. equal resistances $R_N = R_P$
  - approx. **equal rise t_{pLH} and fall t_{pHL} delays**
- Analyze as an RC network

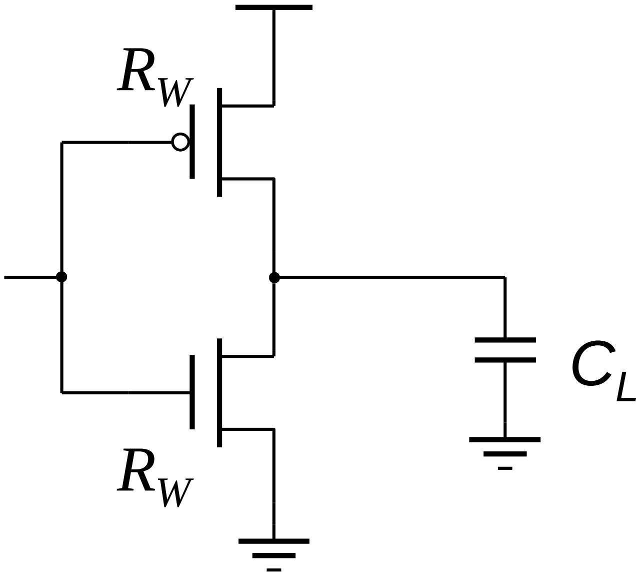


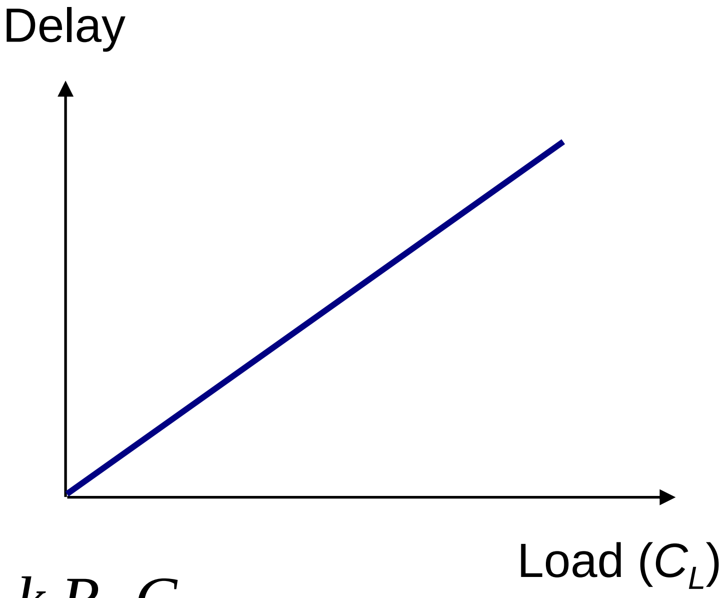
Delay (*D*): $\boxed{t_{pHL} = (\ln 2)\, R_N C_L}$ $\boxed{t_{pLH} = (\ln 2)\, R_P C_L}$

Load for previous stage: $C_{gin} = 3\dfrac{W}{W_{unit}} C_{unit}$

W_{unit} and C_{unit} correspond to an unit size (minimum size) device…

Inverter with Load





$W_{unit} = 1$

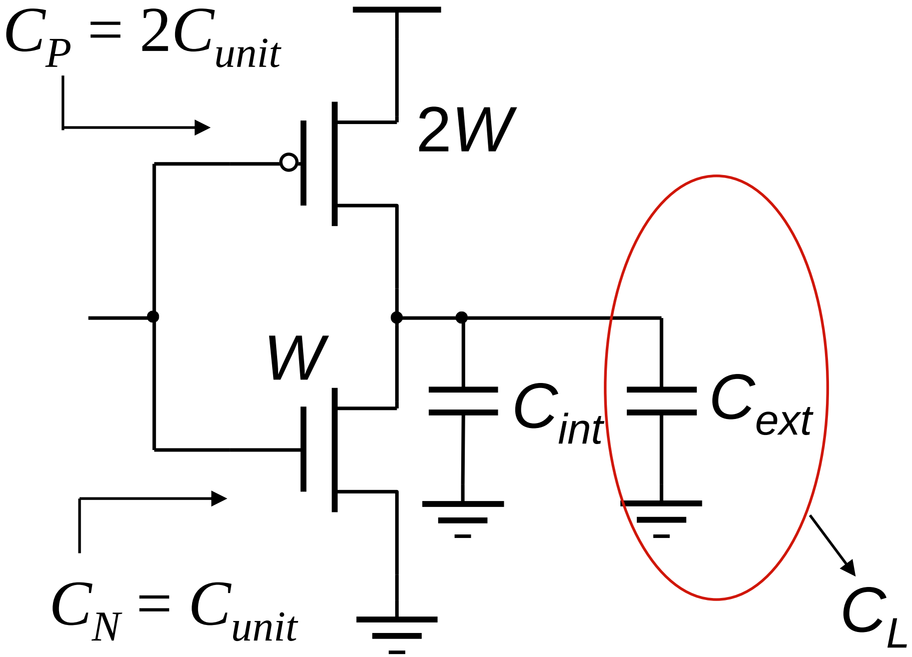
$$t_p = (t_{pHL} + t_{pLH})/2 = k\,R_W C_L$$

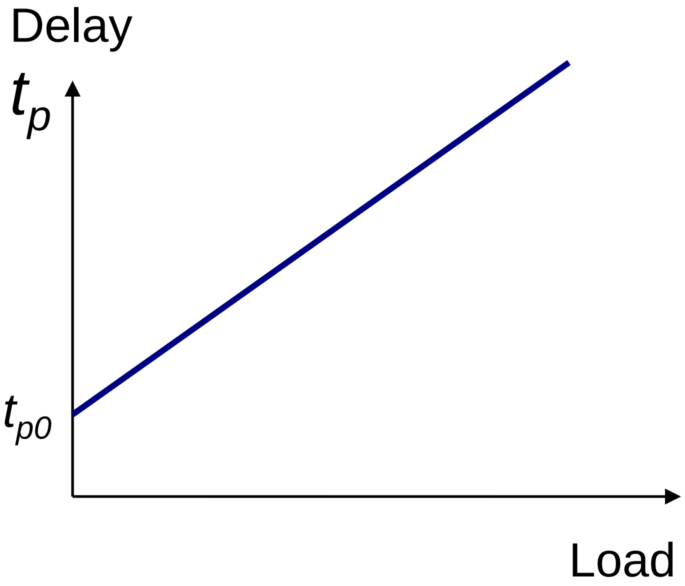
k is a constant, equal to 0.69

Note: $R_p = R_n = R_W$
Hence, $(R_p + R_n)/2 = R_W$

Assumptions: no load → zero delay?

Inverter with Load





Delay (t_p) = $kR_W(C_{int} + C_{ext}) = kR_WC_{int} + kR_WC_{ext} = \underbrace{kR_W C_{int}}_{t_{p0}\ \textit{(intrinsic delay)}}(1+ C_{ext} /C_{int})$

This is the net internal cap.

Intrinsic delay of CMOS inverter

Let R_{eq} be the equivalent resistance of the gate (inverter), then delay (t_p) is defined as:

$$
\begin{aligned}
t_p &= 0.69\, R_{eq} \left(C_{int} + C_{ext}\right) \\
&= 0.69\, R_{eq}\, C_{int} \left(1 + \frac{C_{ext}}{C_{int}}\right) \\
&= t_{p0} \left(1 + \frac{C_{ext}}{C_{int}}\right)
\end{aligned}
$$

t_{p0} is the intrinsic delay

Impact of sizing on gate delay

Let S be the sizing factor

R_{ref} be the resistance of a reference gate (usually a minimum size gate)

C_{iref} be the internal capacitance of the reference gate

$$C_{\mathrm{int}} = S\,C_{iref}\ , \quad R_{eq} = \frac{R_{ref}}{S}$$

$$t_p = 0.69\left(\frac{R_{ref}}{S}\right)\left(S\,C_{iref}\right)\left(1+\frac{C_{ext}}{SC_{iref}}\right)$$

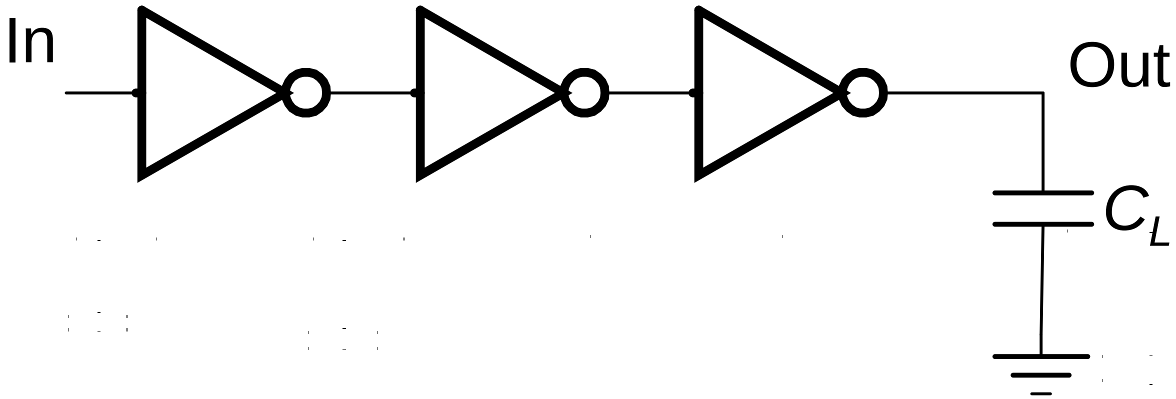
$$= 0.69\,R_{ref}\ C_{iref}\left(1+\frac{C_{ext}}{SC_{iref}}\right)$$

$$= t_{p0}\left(1+\frac{C_{ext}}{SC_{iref}}\right)$$

Hence:

1. *Intrinsic delay is independent of gate sizing, and is determined only by technology and inverter layout*
2. *If S is made very large, gate delay approaches the intrinsic value but increases the area significantly*

Inverter Chain

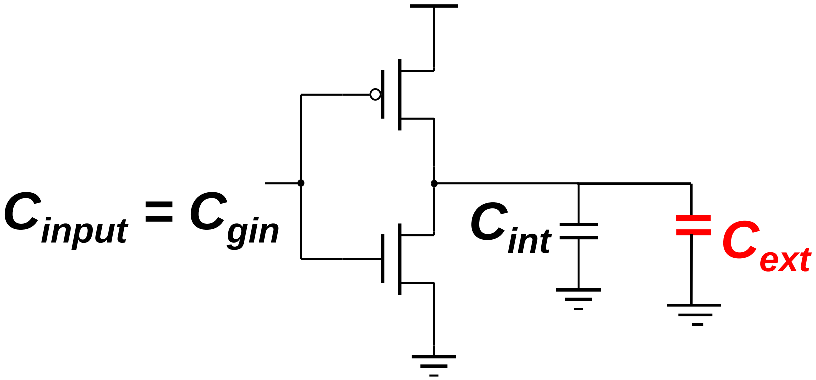


If C_L is given:

- How many stages are needed to minimize the delay?
- How to size the inverters?

May need some additional constraints….

Delay Formula: inverter chain



Let $C_{int} = \gamma C_{gin}$ with $\gamma \approx 1$

$f = C_{ext}/C_{gin}$ - effective fanout

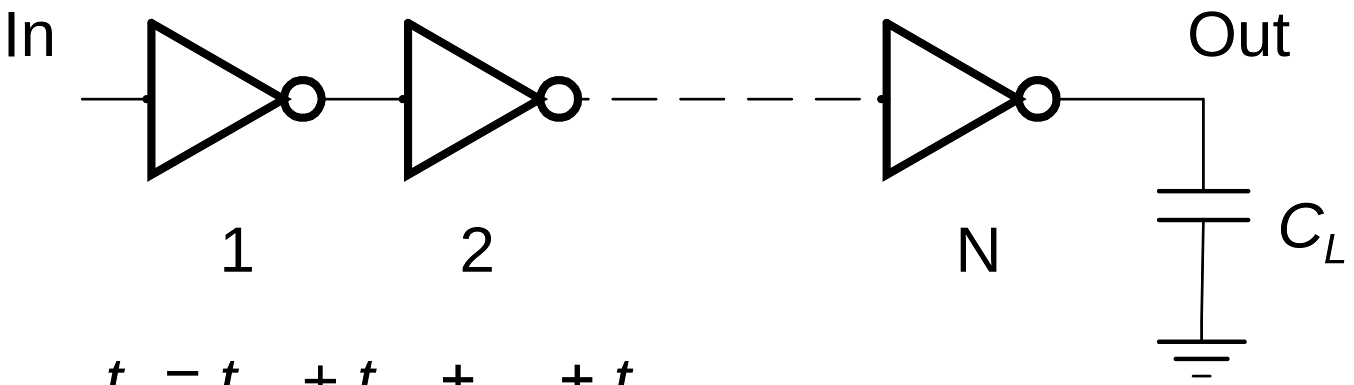
$$Delay \sim R_{eq}\left(C_{int} + C_{ext}\right)$$

Inverter delay is only a function of the RATIO between C_{ext} and C_{input}

$$t_p = \underbrace{0.69 R_{eq} C_{int}}_{t_{p0}}\left(1 + C_{ext}/\gamma C_{gin}\right) = t_{p0}\left(1 + f/\gamma\right)$$

relates the input gate cap. (C_{gin}) and the intrinsic output cap. (C_{int}) of the inverter…

Apply to Inverter Chain



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

$$t_{p,j} = t_{p0}\left(1 + \frac{C_{gin,j+1}}{\gamma C_{gin,j}}\right)$$

This is *Cext* for the *j*th gate

This is *Cint* for the *j*th gate

$$t_p = \sum_{j=1}^{N} t_{p,j} = t_{p0}\sum_{j=1}^{N}\left(1 + \frac{C_{gin,j+1}}{\gamma C_{gin,j}}\right), \quad C_{gin,N+1} = C_L$$

Optimal Tapering for Given N

Delay equation has N - 1 unknowns, $C_{g,2}$ …. $C_{g,N}$

Minimize the delay, find N - 1 partial derivatives and equate them to zero, or $\left(\partial t_p / \partial C_{g,j}\right) = 0$

Result: $C_{g,j+1}/C_{g,j} = C_{g,j}/C_{g,j\text{-}1}$ With j = 2,…..,N

Size of each stage is the geometric mean of two neighbors

$$C_{g,j} = \sqrt{C_{g,j-1}C_{g,j+1}}$$

- each stage has the same effective fanout ($f_j = f = C_{ext}/C_{g,j}$)
- hence, each stage has the same delay: $t_p = t_{p0}\ (1 + f/\gamma)$

Optimum Delay and Number of Stages

When each stage is sized by *f* and has same eff. fanout *f*:

$$\frac{C_L}{C_{g,N}} = \frac{C_{g,N}}{C_{g,N-1}} = \cdots\cdots\cdots\cdots = \frac{C_{g,2}}{C_{g,1}} = f$$

(multiplying all the terms)

$$Hence,\ f^N = C_L / C_{g,1} = F$$

F is the overall effective fanout of the circuit

Effective fanout of each stage: $f = \sqrt[N]{F}$

If C_L and $C_{g,1}$ are known….

Minimum path delay:

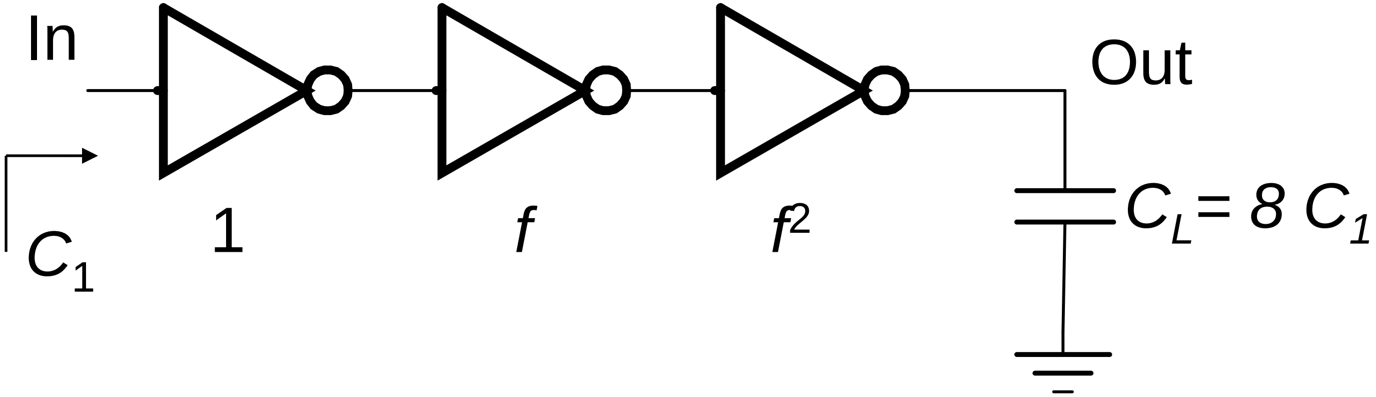
$$t_p = N t_{p0} \left(1 + \sqrt[N]{F} / \gamma\right)$$

If N is too large, intrinsic delay of stages dominate, while if N is small, effective fanout of each stage (f) is large and the second term dominates

How to choose N?

Example

If N is given….



C_L/C_1 has to be evenly distributed across $N = 3$ stages:

$F = (8C_1)/C_1 = 8$

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

Optimum Number of Stages

For a given load, C_L and given input capacitance C_{in}
Find optimal sizing f

$$C_L = F \cdot C_{in} = f^N C_{in} \quad with \quad N = \frac{\ln F}{\ln f}$$

$$t_p = N t_{p0} \left(F^{1/N} / \gamma + 1 \right) = \frac{t_{p0} \ln F}{\gamma} \left(\frac{f}{\ln f} + \frac{\gamma}{\ln f} \right)$$

$$\frac{\partial t_p}{\partial f} = \frac{t_{p0} \ln F}{\gamma} \cdot \frac{\ln f - 1 - \gamma / f}{\ln^2 f} = 0$$

If self-loading is ignored….

For γ = 0, *f* = e = 2.718, *N* = ln*F*

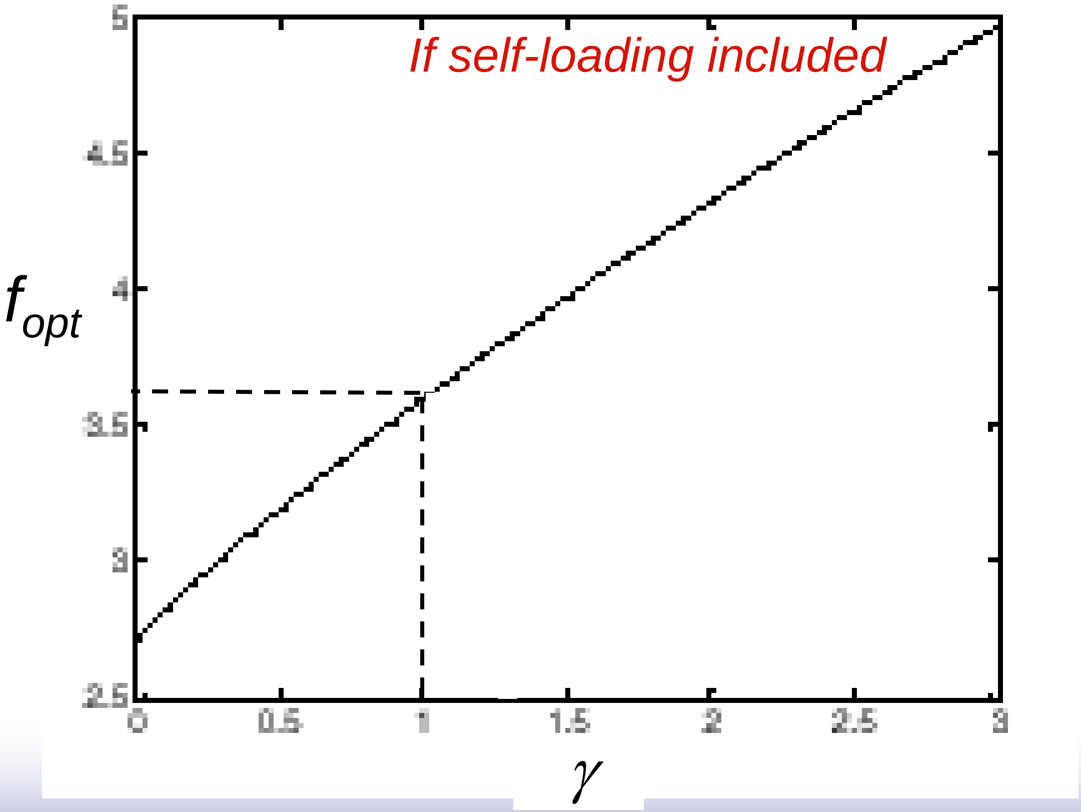
Otherwise….

$$f = \exp\left(1 + \gamma / f\right)$$

Optimum Effective Fanout f

Optimum *f* for given process defined by γ

$$f = exp\left(1 + \gamma / f\right)$$

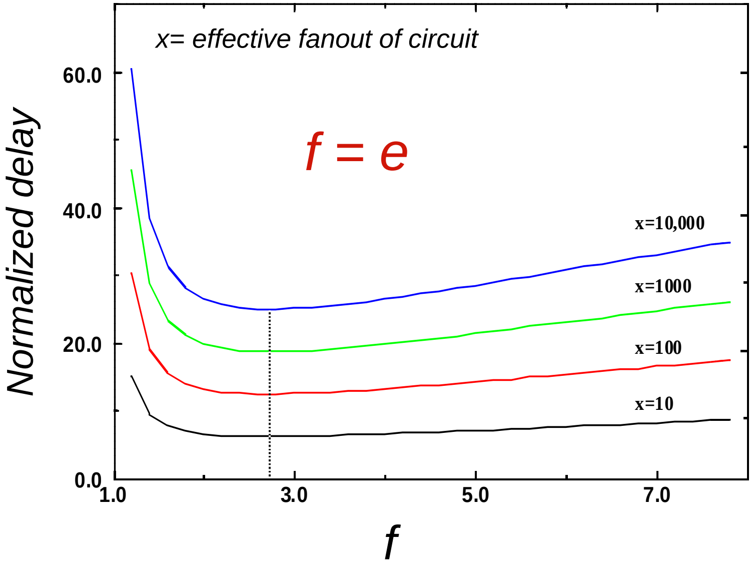


Optimum tapering factor:

f_{opt} = 3.6
for γ=1 (typical case)

Impact of Self-Loading on tp

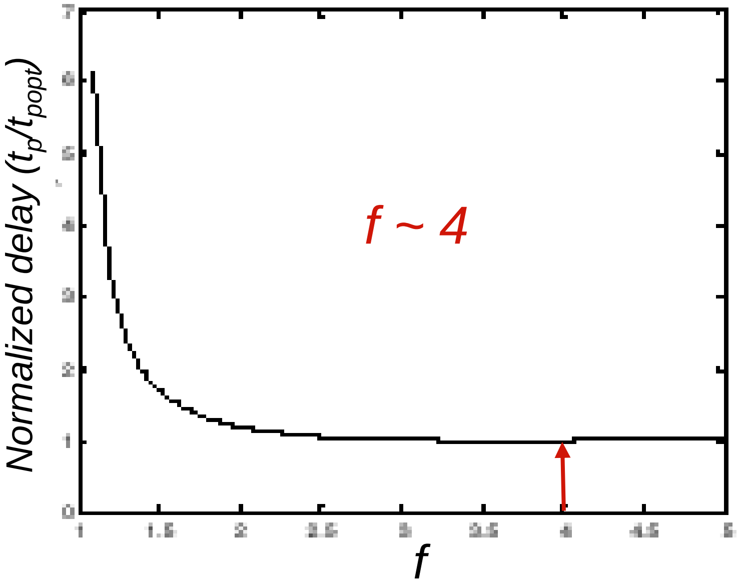
No Self-Loading, γ=0



f = e

Optimal number of stages, N= ln(F)

With Self-Loading γ=1



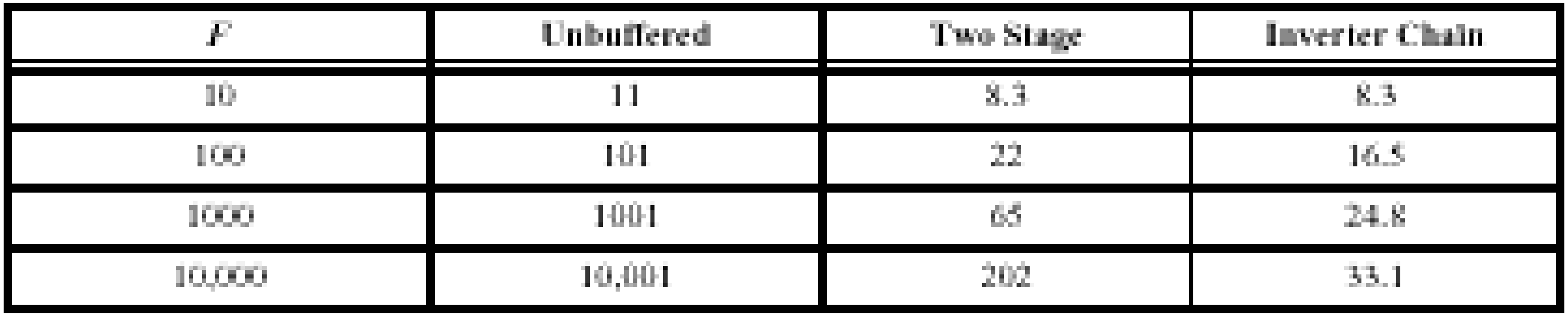
f ~ 4

If $f<f_{opt}$ (too many stages) will result in delay to increase

Normalized delay function of F

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

t_{popt}/t_{p0} for γ=1

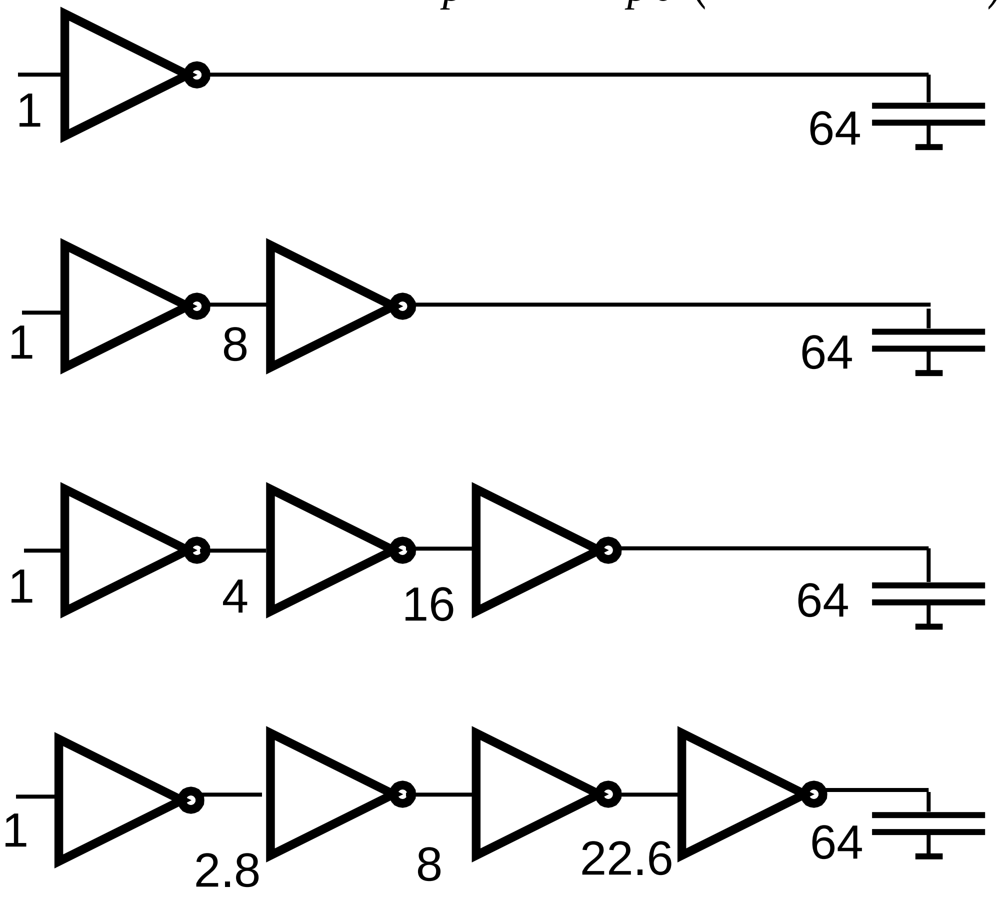


F	Unbuffered	Two Stage	Inverter Chain
10	11	8.3	8.3
100	101	22	16.5
1000	1001	65	24.8
10,000	10,001	202	33.1

As F increases, the differences between the unbuffered case (or two-stage buffer case) and the case of inverter chain increases…..

Buffer Design

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

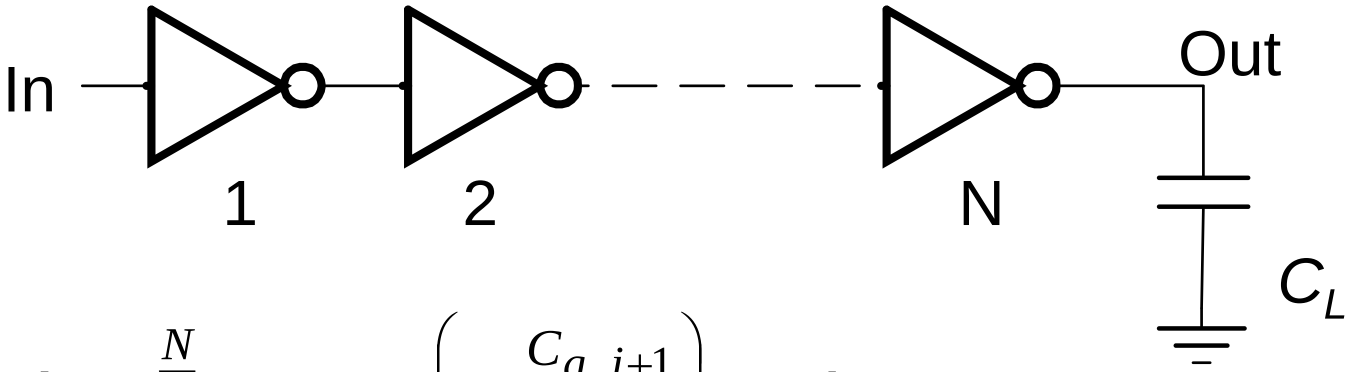


	$f = F^{1/N}$	t_{popt}/t_{p0} for γ=1
N	**f**	$\mathbf{t_p}$
1	**64**	**65**
2	**8**	**18**
3	**4**	**15**
4	**2.8**	**15.3**

Sizing Logic Paths for Speed

- Frequently, input capacitance of a logic path is constrained
- Logic also has to drive some capacitance
- Example: ALU load in an Intel's microprocessor is 0.5pF
- How do we size the ALU datapath to achieve maximum speed?
- We have already solved this for the inverter chain – can we generalize it for any type of logic?

Buffer Example



$$Chain\ Delay = \sum_{j=1}^{N} t_{p,j} = t_{p0}\left(1 + \frac{C_{g,j+1}}{\gamma C_{g,j}}\right), \quad with\ C_{g,N+1} = C_L$$

(in units of τ_{inv})

For given N: $C_{g,\ j+1}/C_{g,j} = C_{g,j}/C_{g,\ j-1}$

Optimal fanout (f): $C_{g,\ j+1}/C_{g,\ j} \sim 4$

How to generalize this to any logic path?

Minimizing Delay in Complex Logic Networks

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

$$= t_{p0}\left(p+\frac{g\cdot f}{\gamma}\right)(Complex\ gate)$$

Everything Normalized w.r.t an inverter:
g_{inv} =1, p_{inv} = 1

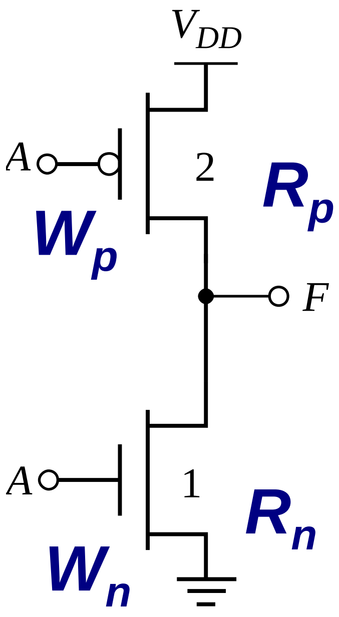
f – effective fanout *(ratio of external load and input cap. of gate)*
p – ratio of intrinsic delays of complex gate and inverter
(value increases with complexity of gate)
g – logical effort: how much more input capacitance is presented by the complex gate to deliver the same output current as an inverter (depends only on circuit topology)

Logical Effort

- Inverter has the smallest logical effort and intrinsic delay of all static CMOS gates
- Logical effort of a gate is the ratio of its input capacitance to the inverter capacitance when sized to deliver the same current
- Logical effort increases with gate complexity

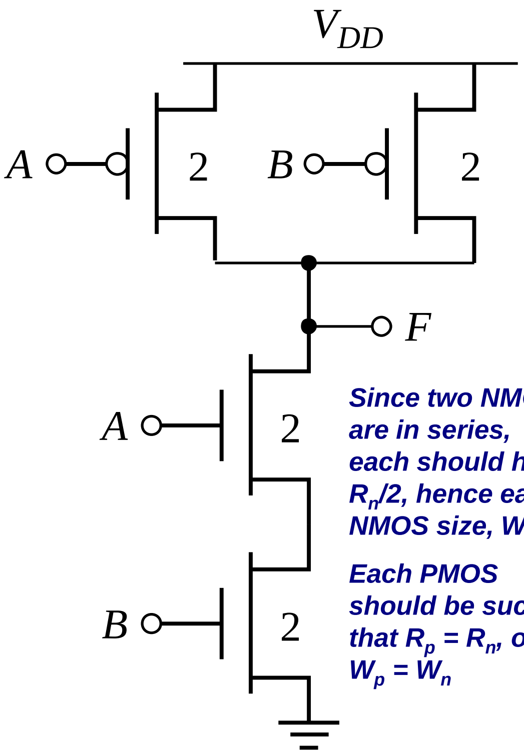
Logical Effort

Logical effort is the ratio of input capacitance of a gate to the input capacitance of an inverter with the same output current



Inverter

***g* = 1**

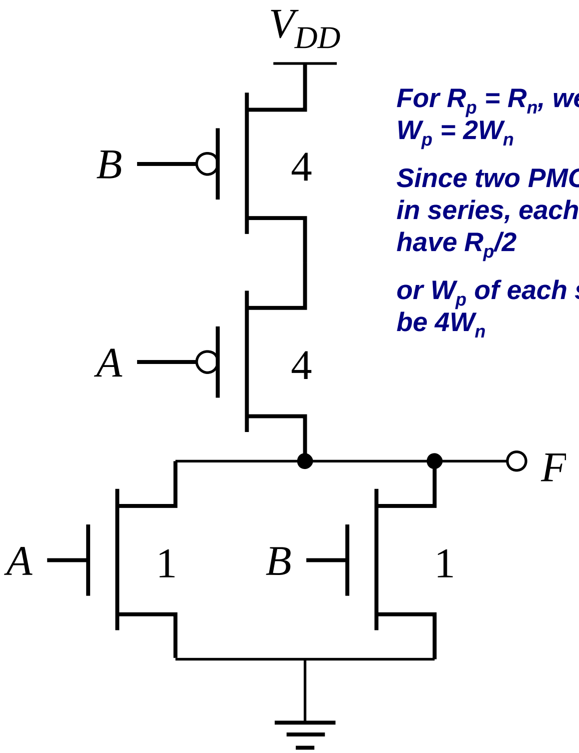


Since two NMOS are in series, each should have $R_n/2$, hence each NMOS size, W_n =2

Each PMOS should be such that $R_p = R_n$, or $W_p = W_n$

2-input NAND

***g* = 4/3**



For $R_p = R_n$, we need $W_p = 2W_n$

Since two PMOS are in series, each should have $R_p/2$

or W_p of each should be $4W_n$

2-input NOR

***g* = 5/3**

Delay in a Logic Gate

Gate delay:

$$d = h + p$$

effort delay **intrinsic delay**

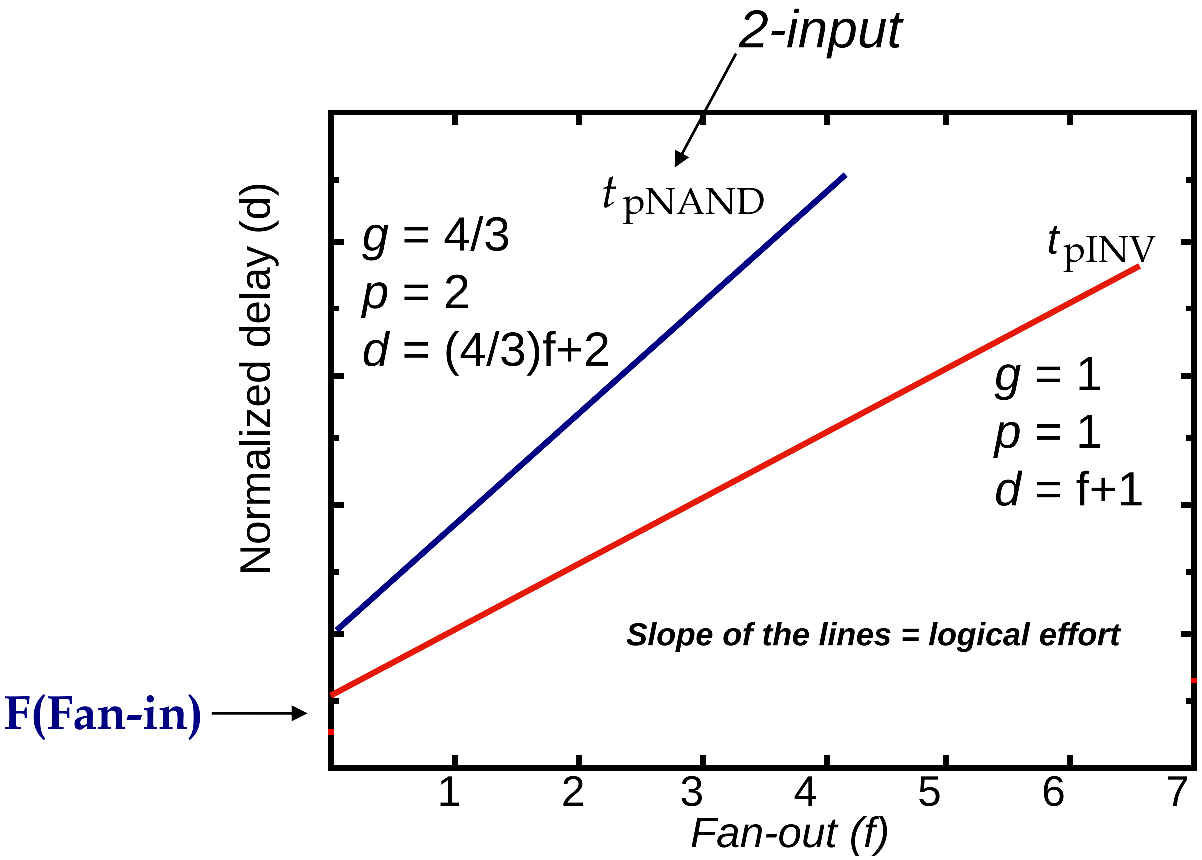
Effort delay (or gate effort):

$$h = g\, f$$

logical effort **effective fanout** $= C_{ext}/C_{in}$

- **Logical effort is a function of topology, independent of sizing**
- **Effective fanout (electrical effort) is a function of load/gate size**

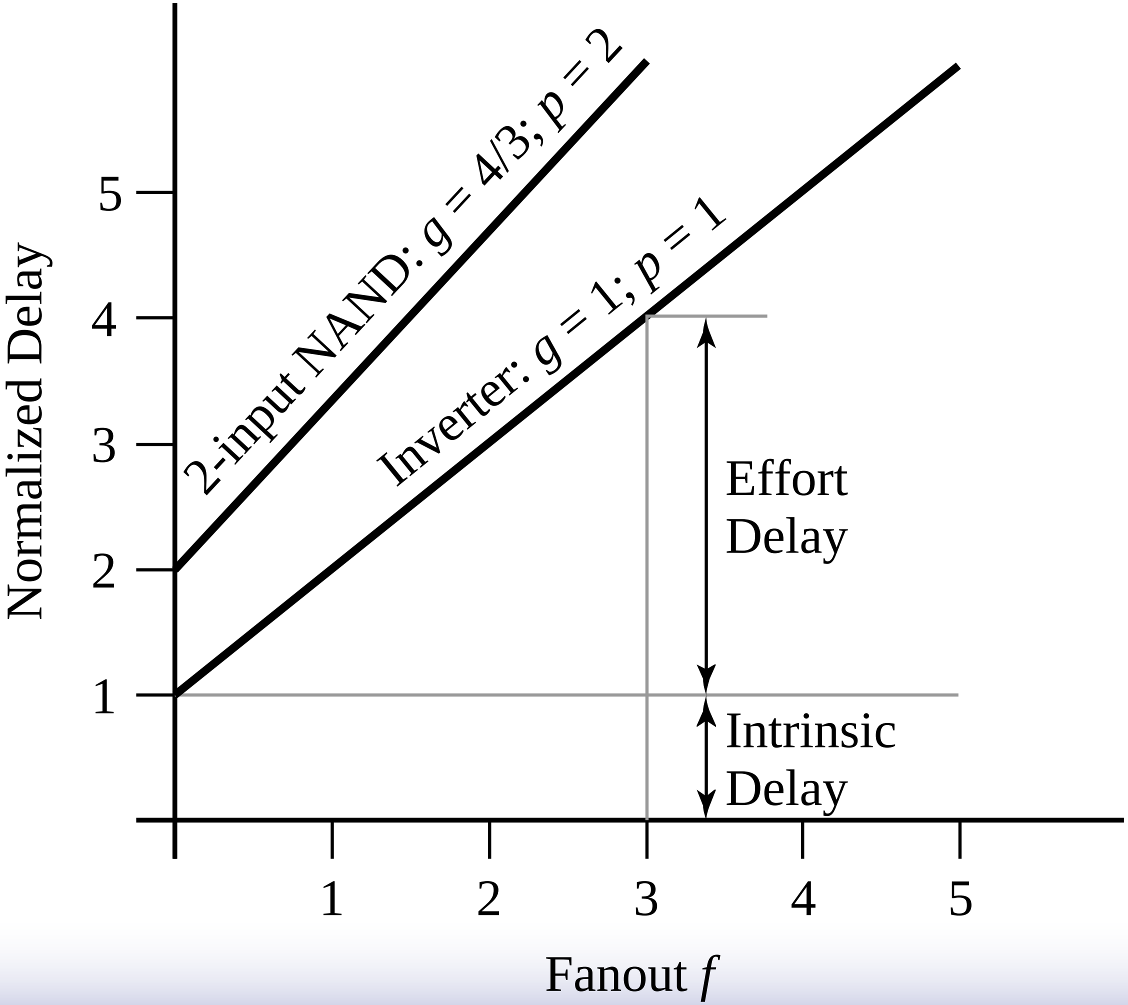
Logical Effort of Gates



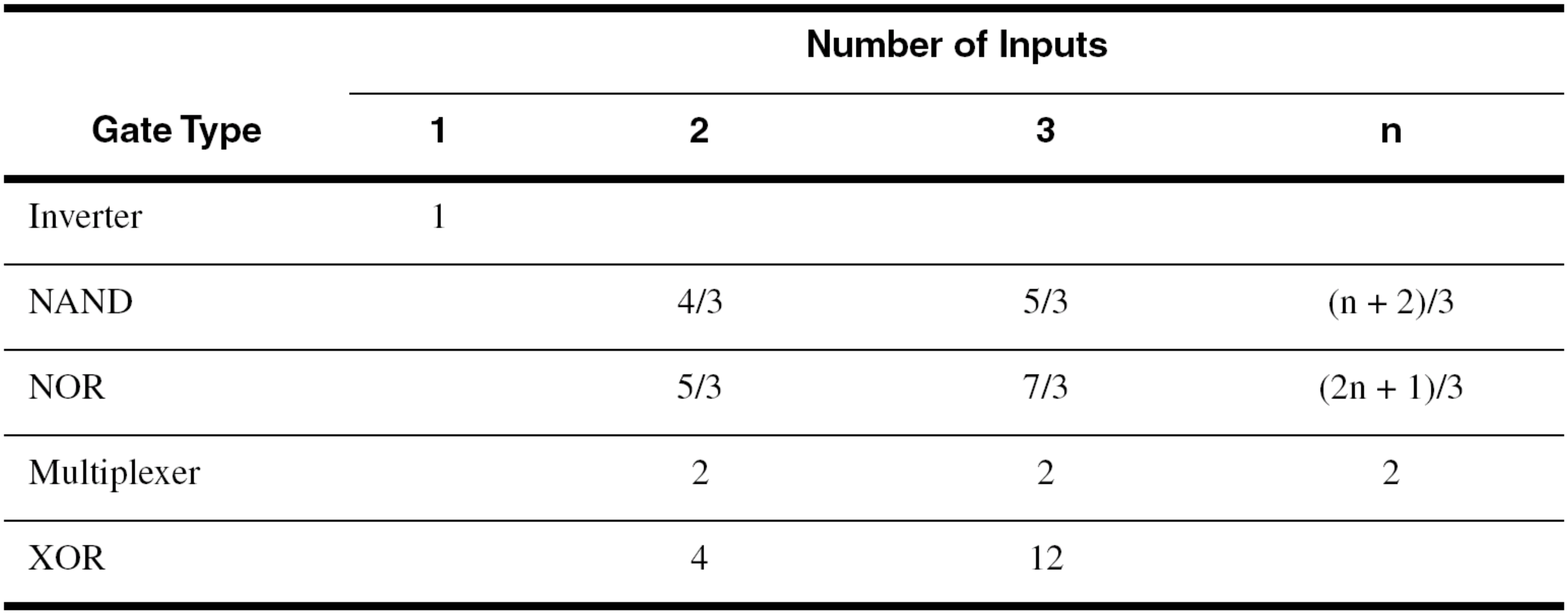
$$d = p + g\,f$$

- **Delay can be adjusted by**:
  - *transistor sizing that changes the effective fanout*
  - *Choosing a gate with different g*

Logical Effort of Gates



Logical Effort



Gate Type	Number of Inputs			
	1	2	3	n
Inverter	1			
NAND		4/3	5/3	(n + 2)/3
NOR		5/3	7/3	(2n + 1)/3
Multiplexer		2	2	2
XOR		4	12	

From Sutherland, Sproull

Total delay through a combinational logic block

$$t_p = \sum_{j=1}^{N} t_{p,j} = t_{p0} \sum_{j=1}^{N} \left(p_j + \frac{f_j g_j}{\gamma} \right)$$

Similar to inverter chain delay….find N-1 partial derivatives and equate them to zero….

For minimal delay : $g_1 f_1 = g_2 f_2 = \ldots = g_N f_N$ (each stage should have the same gate effort, h)

$$\mathit{Path\ Logic\ Effort} = G = \prod_{1}^{N} g_i$$

Note: In the text book, this is defined as H

$$\mathit{Path\ Effective\ Fanout} = F = \frac{C_L}{C_{g1}}$$

(or Path Electrical Effort)

Branching Effort

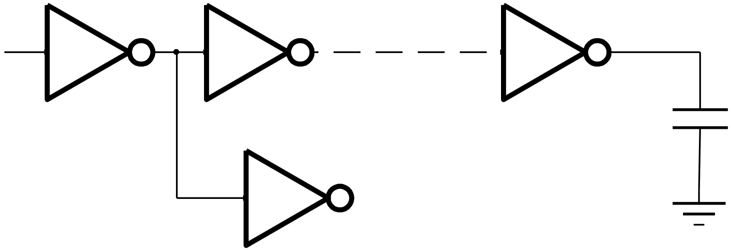
To relate F to the effective fanouts of the individual gates, one must account for the logical fanout within the network

When fanout occurs at the output of a node, some of the available drive current is directed along the path being analyzed

Branching effort of a logic gate:

$$b = \frac{C_{on-path} + C_{off-path}}{C_{on-path}}$$

Load capacitance of the gate along the path under study



$$Path\,Branching\,Effort = B = \prod_{1}^{N} b_i$$

Total Path Effort

- Path electrical effort can be related to the electrical and branching efforts of the individual gates:

$$F = \prod_{1}^{N} \frac{f_i}{b_i} = \frac{\prod f_i}{B}$$

- Total path effort can be defined as:

$$H = \prod_{1}^{N} h_i = \prod_{1}^{N} g_i \; f_i = GFB$$

Note: In the text book, H and F have been swapped…

- Gate effort that minimizes the path delay = ?

- Minimum delay through path = ?

Multistage Networks

$$Delay = \sum_{i=1}^{N} \left(p_i + g_i \cdot f_i\right)$$

Gate effort: $h_i = g_i f_i$

Path electrical effort: $F = C_L / C_{gin}$

Path logical effort: $G = g_1 g_2 \ldots g_N$

Path branching effort: $B = b_1 b_2 \ldots b_N$

*Path effort: $H = GFB$

Path delay $D = \Sigma d_i = \Sigma p_i + \Sigma h_i$

*** Note: In the text book, this is defined as: F = GHB**

Optimal Number of Stages

For a given load,
and given input capacitance of the first gate
Find optimal number of gates and optimal sizing

$$D = NH^{1/N} + Np_{inv}$$

$$\frac{\partial D}{\partial N} = -H^{1/N}\ln\left(H^{1/N}\right) + H^{1/N} + p_{inv} = 0$$

Substitute ‘best gate effort’: $h = H^{1/N}$ → *Gate effort that minimizes path delay*

A path achieves least delay by using $N = \log_4 H$ stages

Optimum Effort per Stage

When each stage bears the same effort:

$$h^N = H$$

$$h = \sqrt[N]{H}$$

gate efforts: $g_1 f_1 = g_2 f_2 = \ldots = g_N f_N$

Effective fanout of each gate: $f_i = h / g_i$

Minimum path delay:

$$D = t_{p0}\left(\sum_{j=1}^{N} p_j + \frac{N\left(\sqrt[N]{H}\right)}{\gamma}\right)$$

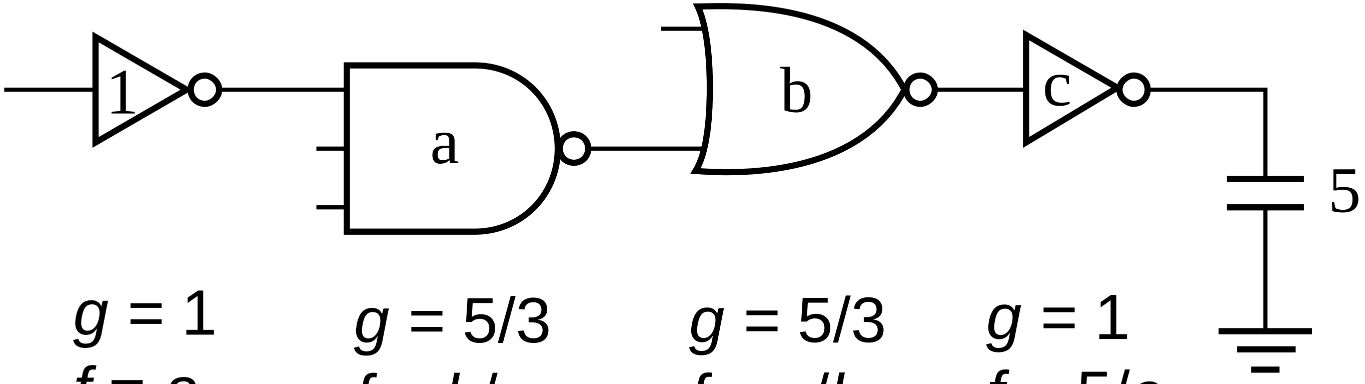
Sizing of Chain of Gates

- Consider chain s_i
- Sizing factors for each gate in the chain can be derived by working out from front to end (or vice versa).
- Assume that a unit-size gate has a driving capability equal to a minimum-size inverter
- Hence, $C_{gin} = g\, C_{in\_ref}$
- If s_1 is the sizing factor for gate 1:
  - $C_{g1} = s_1\, g_1\, C_{in\_ref}$
  - Input capacitance of gate 2 is larger by f_1/b_1:

    That is, $C_{g2} = f_1/b_1\, C_{g1} = s_2\, g_2\, C_{in\_ref}$
  - For gate i in the chain:

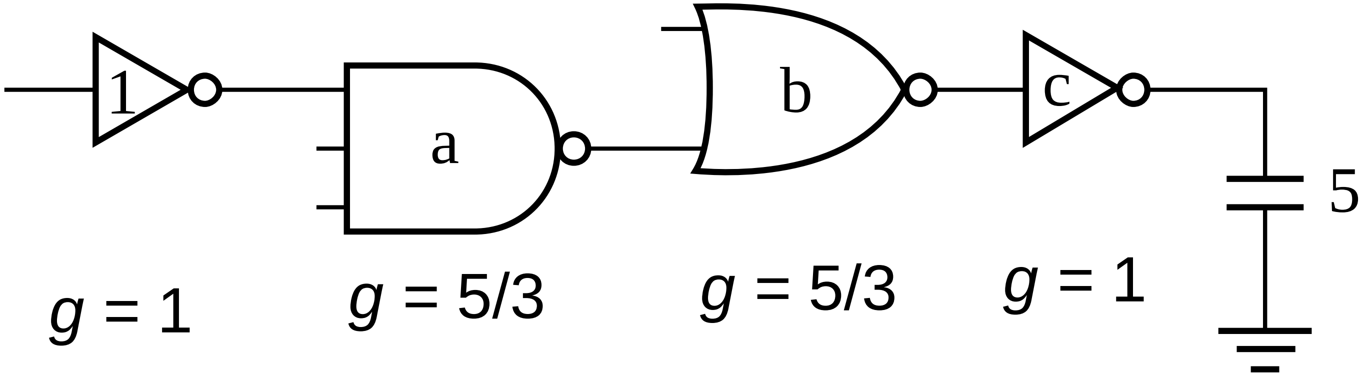
$$s_i = \left(\frac{g_1 s_1}{g_i}\right) \prod_{j=1}^{i-1} \left(\frac{f_j}{b_j}\right)$$

Example: Optimize Path



$g = 1$	$g = 5/3$	$g = 5/3$	$g = 1$
$f = a$	$f = b/a$	$f = c/b$	$f = 5/c$

Example: Optimize Path



Effective fanout, F = 5/1 = 5

$G = 1 \times 5/3 \times 5/3 \times 1 = 25/9$

B=1 (no branching)

$H = GFB = 125/9 = 13.9$

$h = H^{1/4} = 1.93$ (optimal gate effort)

Derive Fanout Factors (taking gate types into account):

f1 = 1.93 (since h=gf)

f2 = 1.93 (*3/5)* = 1.16

f3 = 1.16

f4=1.93

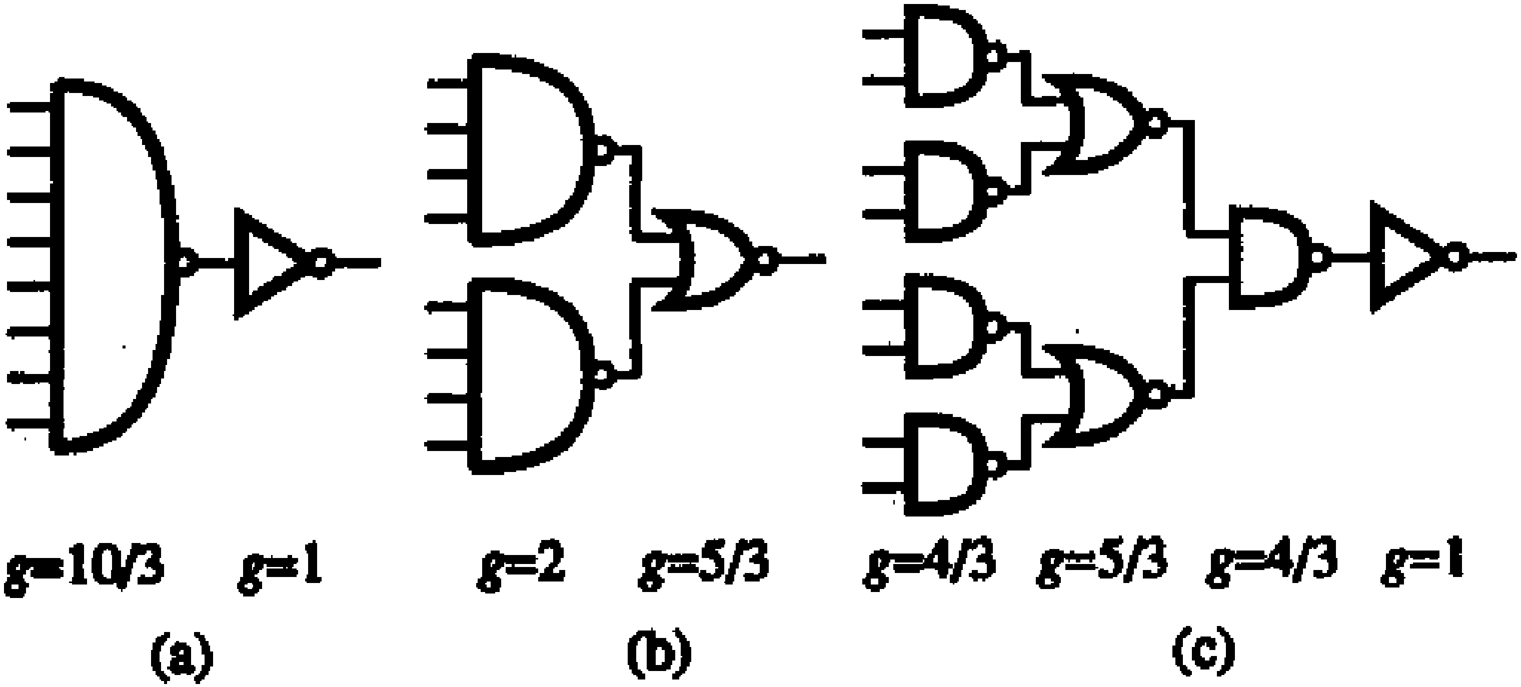
Derive Gate Sizes:

$a = f_1 g_1 / g_2 = 1.16$

$b = f_1 f_2 g_1 / g_3 = 1.34$

$c = f_1\, f_2\, f_3\, g_1 / g_4 = 2.6$

Example – 8-input AND



Method of Logical Effort

- □ Compute the path effort: $H = GFB$
- □ Find the best number of stages $N \sim \log_4 H$
- □ Compute the stage effort $h = H^{1/N}$
- □ Sketch the path with this number of stages
- □ Work from either end, find sizes:
  $C_{in} = C_{out} * g/h$

Reference: Sutherland, Sproull, Harris, “Logical Effort, Morgan-Kaufmann 1999.

Summary

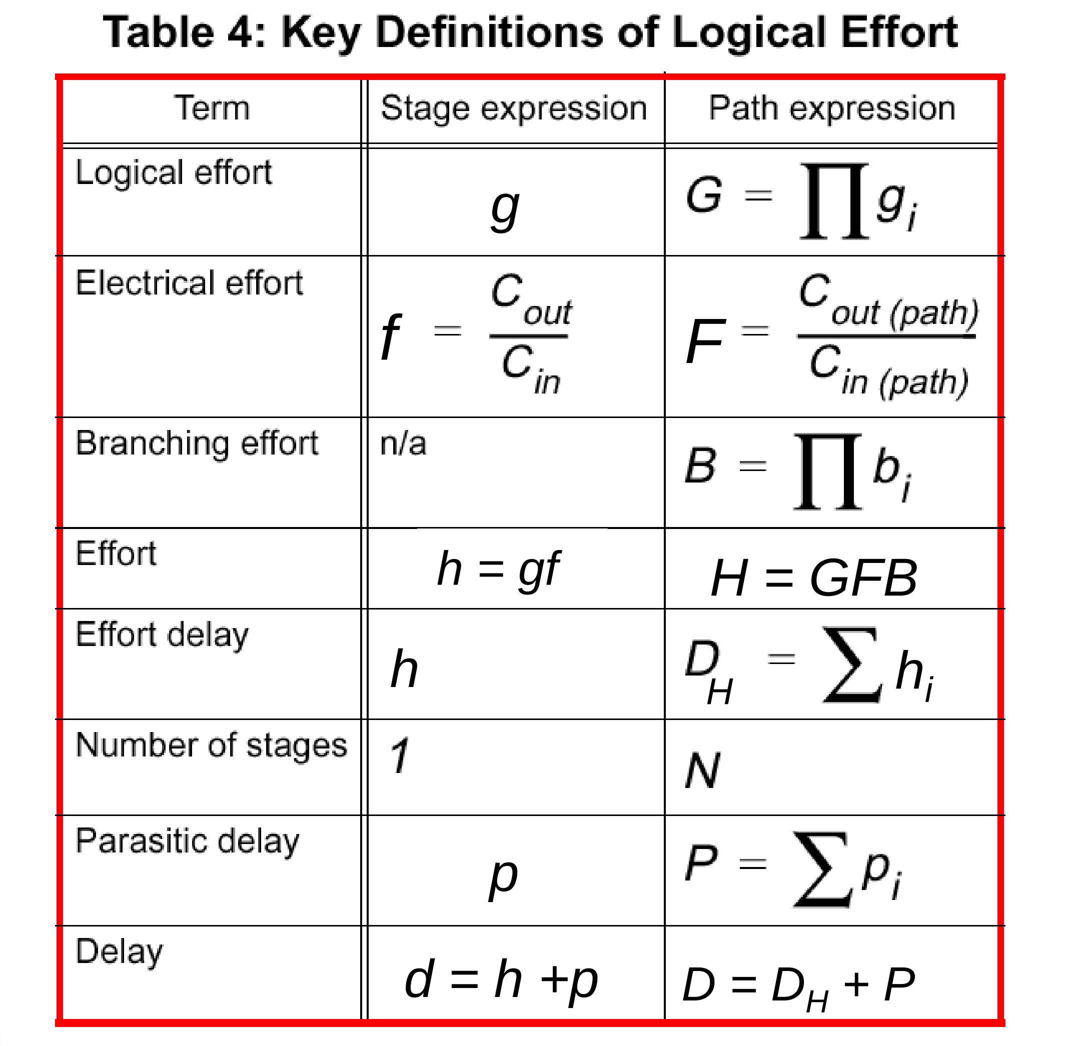


Table 4: Key Definitions of Logical Effort

Term	Stage expression	Path expression
Logical effort	g	$G = \prod g_i$
Electrical effort	$f = \frac{C_{out}}{C_{in}}$	$F = \frac{C_{out\ (path)}}{C_{in\ (path)}}$
Branching effort	n/a	$B = \prod b_i$
Effort	$h = gf$	$H = GFB$
Effort delay	h	$D_H = \sum h_i$
Number of stages	1	N
Parasitic delay	p	$P = \sum p_i$
Delay	$d = h + p$	$D = D_H + P$

Sutherland, Sproull and Harris