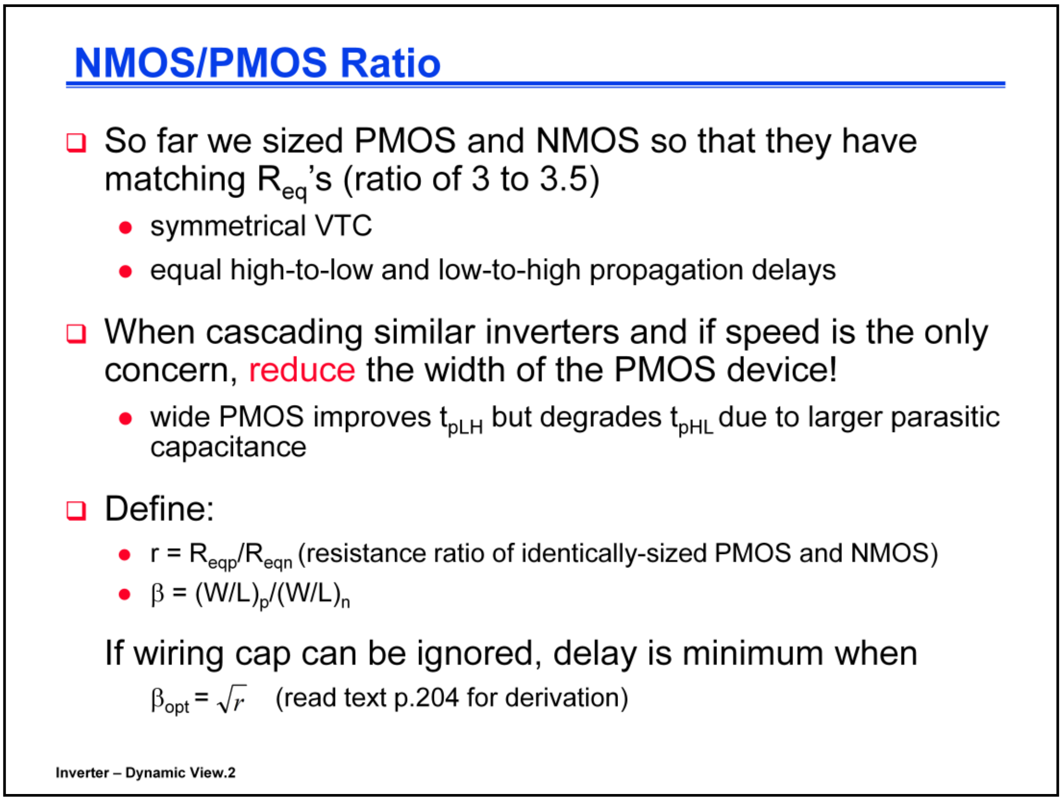
VLSI Circuits and Design Inverter and Dynamic View



NMOS/PMOS Ratio

- ❑ So far we sized PMOS and NMOS so that they have matching R_{eq}'s (ratio of 3 to 3.5)
  - symmetrical VTC
  - equal high-to-low and low-to-high propagation delays
- ❑ When cascading similar inverters and if speed is the only concern, reduce the width of the PMOS device!
  - wide PMOS improves t_{pLH} but degrades t_{pHL} due to larger parasitic capacitance
- ❑ Define:
  - $r = R_{eqp}/R_{eqn}$ (resistance ratio of identically-sized PMOS and NMOS)
  - $\beta = (W/L)_p/(W/L)_n$

If wiring cap can be ignored, delay is minimum when

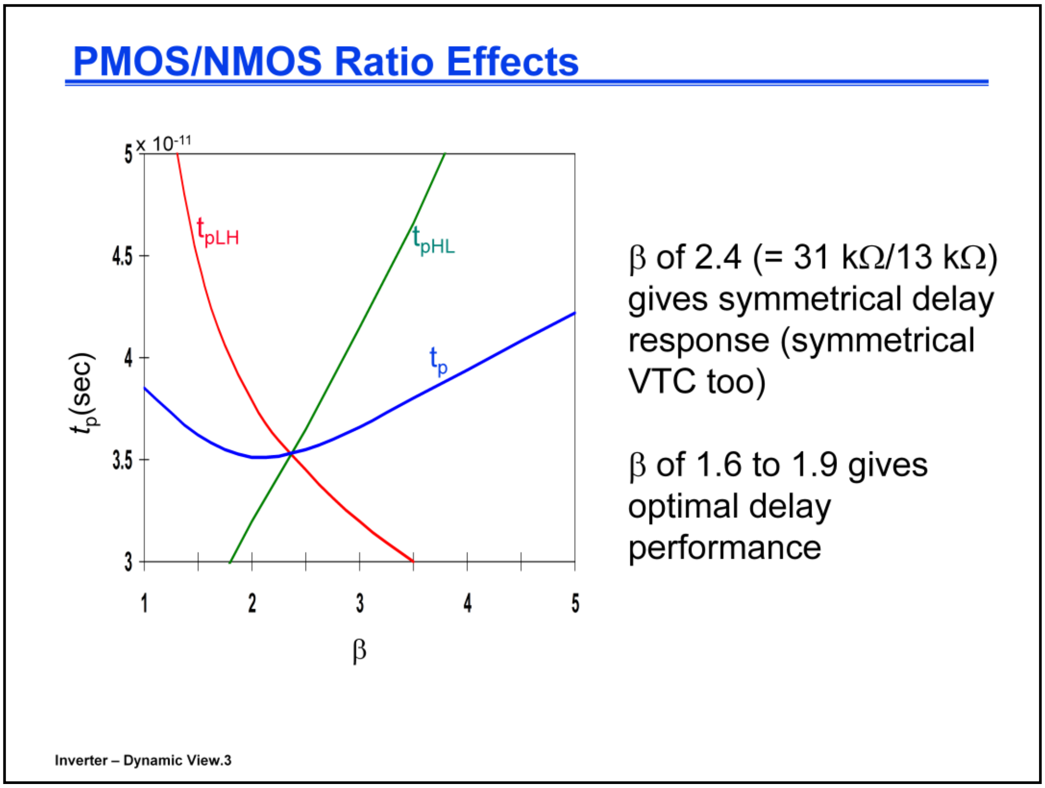
$\beta_{opt} = \sqrt{r}$ (read text p.204 for derivation)

In slide "CMOS Inverter", the VTC simulation on p. 20 was obtained using $(W/L)_p/(W/L)_n = 3.4$. For this inverter, $V_M \sim 1.25$ V.

When we use smaller PMOS – get better speed at the cost of VTC symmetry and noise margin.

Widening PMOS of course improves t_{pLH} because it improves R_{eq} of the PMOS.

If wiring capacitance is not negligible, larger values of β should be used. The surprising result is that smaller device sizes (and hence smaller area) yield a faster design at the expense of VTC symmetry and noise margin.



PMOS/NMOS Ratio Effects

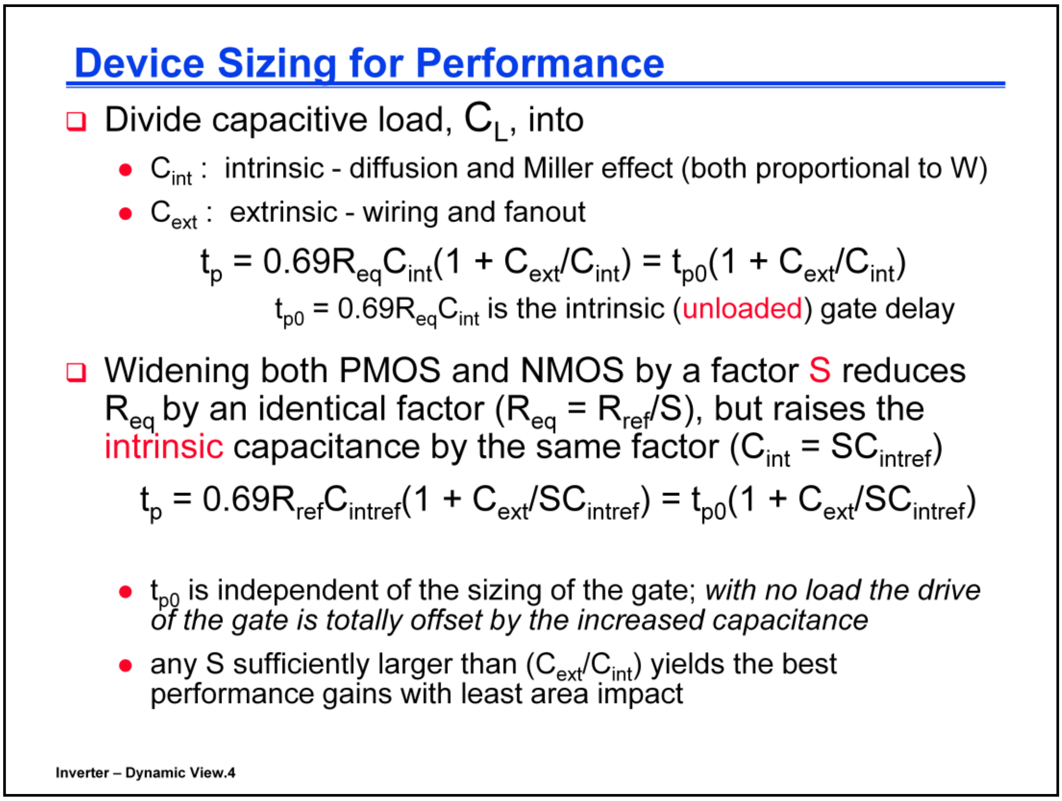
β of 2.4 (= 31 kΩ/13 kΩ) gives symmetrical delay response (symmetrical VTC too)

β of 1.6 to 1.9 gives optimal delay performance

From slide “MOS Transistor” p. 23, R_{eq} for PMOS and NMOS are 31 kΩ and 13 kΩ. For both transistors, (W/L) = 1.

For symmetrical characteristics, we want $R_{pu} = R_{pd}$. Hence we make $(W/L)_p = 2.4(W/L)_n$ i.e. β = r = 2.4. From graph, delay is indeed symmetrical for this value of β.

In this example r = 31/13 = 2.4. Theoretically, for min delay $\beta_{opt} = 2.4^{0.5} = 1.5$. From plot above, min delay is actually at β = 1.9.



Device Sizing for Performance

- Divide capacitive load, C_L, into
  - C_{int} : intrinsic - diffusion and Miller effect (both proportional to W)
  - C_{ext} : extrinsic - wiring and fanout

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

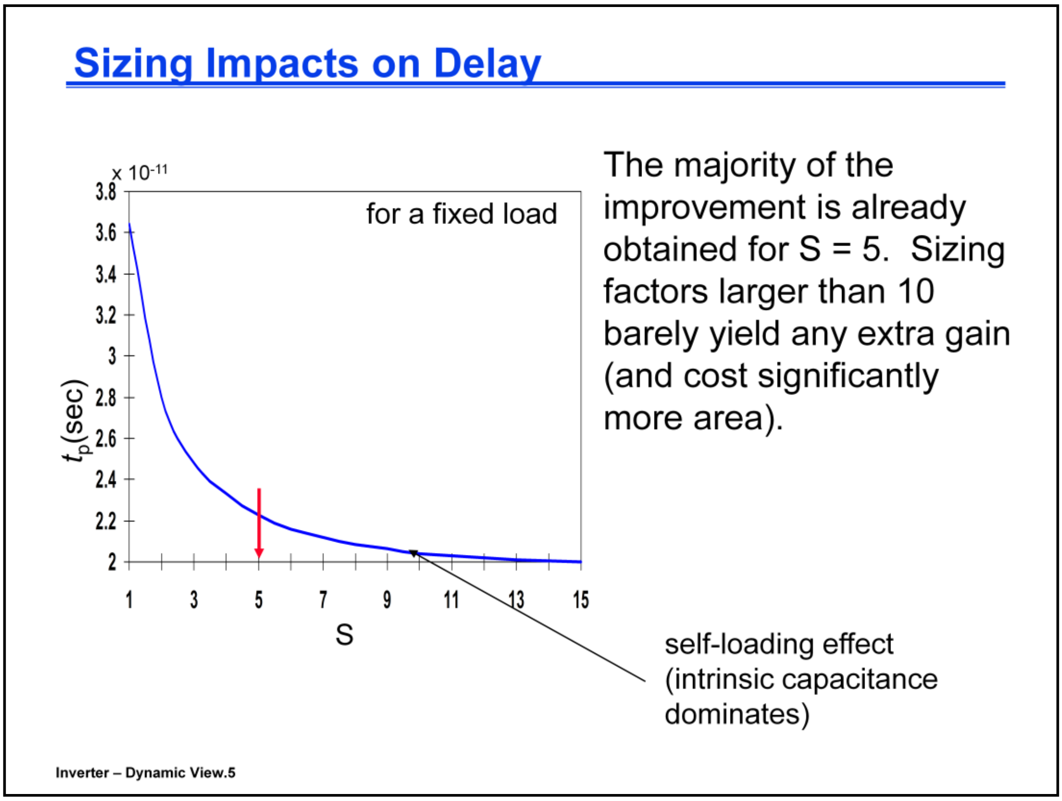
$t_{p0} = 0.69R_{eq}C_{int}$ is the intrinsic (unloaded) gate delay

- 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 ($C_{int} = SC_{intref}$)

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

  - 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

Making S infinitely large yields the maximum obtainable performance gain, i.e. min delay but of course area is also infinitely large too.

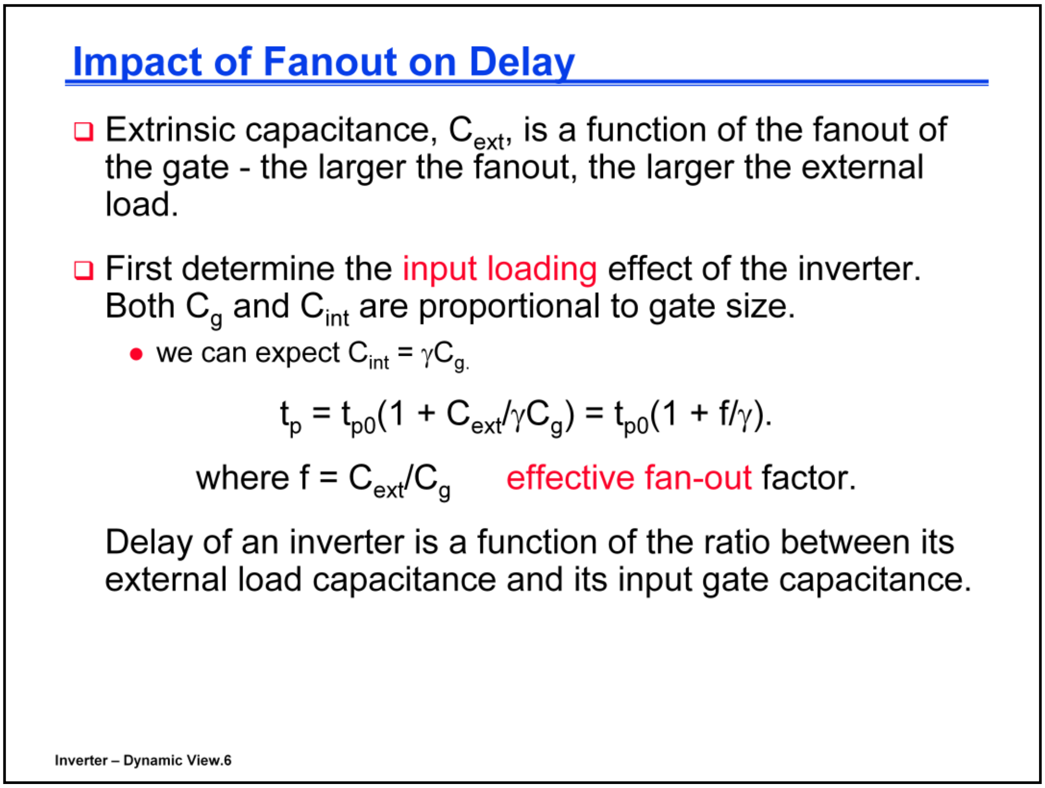


Sizing Impacts on Delay

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)

While sizing up an inverter reduces its delay, it also increases its input capacitance – impacting the delay of the driving gate!



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 gate size.
  - ● we can expect $C_{int} = \gamma C_{g.}$

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

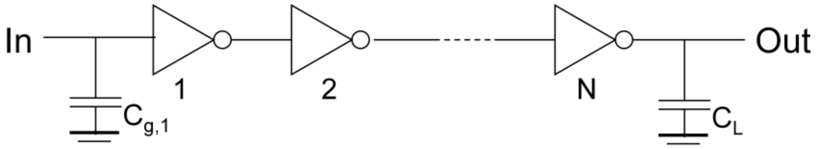
where $f = C_{ext}/C_g$ effective fan-out factor.

Delay of an inverter is a function of the ratio between its external load capacitance and its input gate capacitance.

γ ≈ 1 for most submicron processes.

Inverter Chain

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



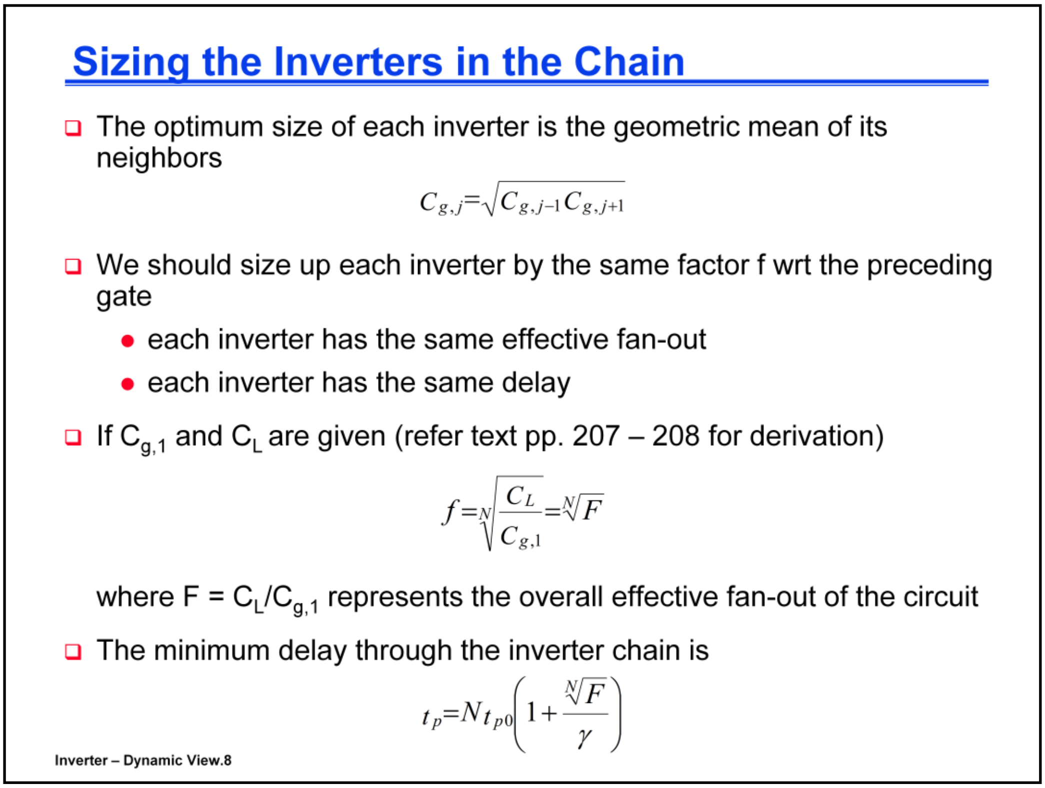
the delay of the j-th inverter stage is (ignore wire cap)

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

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

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

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

- We should size up each inverter by the same factor f wrt the preceding gate
  - each inverter has the same effective fan-out
  - each inverter has the same delay
- If $C_{g,1}$ and C_L are given (refer text pp. 207 – 208 for derivation)

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

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

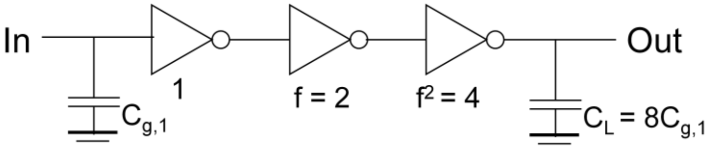
- The minimum delay through the inverter chain is

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

Next question is, “what is the best N to minimize the delay for a given F?”

Before that, an example of inverter chain sizing.

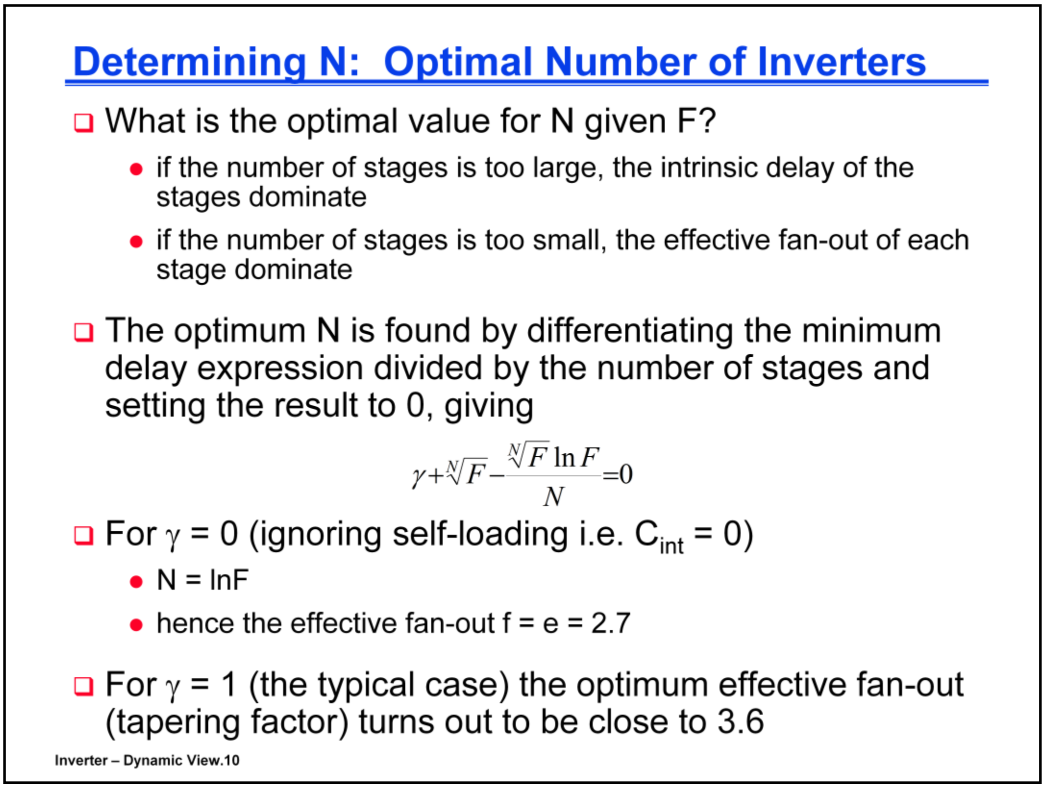
Example of Inverter Chain Sizing



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

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

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



Determining N: Optimal Number of Inverters

- ❑ What is the optimal value for N given F?
  - ● if the number of stages is too large, the intrinsic delay of the stages dominate
  - ● if the number of stages is too small, the effective fan-out of each stage 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

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

- ❑ For $\gamma = 0$ (ignoring self-loading i.e. $C_{int} = 0$)
  - ● N = lnF
  - ● hence the effective fan-out f = e = 2.7
- ❑ For $\gamma = 1$ (the typical case) the optimum effective fan-out (tapering factor) turns out to be close to 3.6

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

$$\frac{d}{dN}t_p = t_{p0} + \frac{t_{p0}}{\gamma}\frac{d}{dN}NF^{1/N}$$

$$= t_{p0} + \frac{t_{p0}}{\gamma}\left(F^{1/N}\frac{d}{dN}N + N\frac{d}{dN}F^{1/N}\right)$$

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

$$= t_{p0} + \frac{t_{p0}}{\gamma}\left(F^{1/N} - N(\ln F)F^{1/N}\frac{1}{N^2}\right)$$

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

Set $dt_p/dN = 0$, we get $\gamma + F^{1/N} - \frac{1}{N}(\ln F)F^{1/N} = 0$.

Derivation for optimum fan-out:

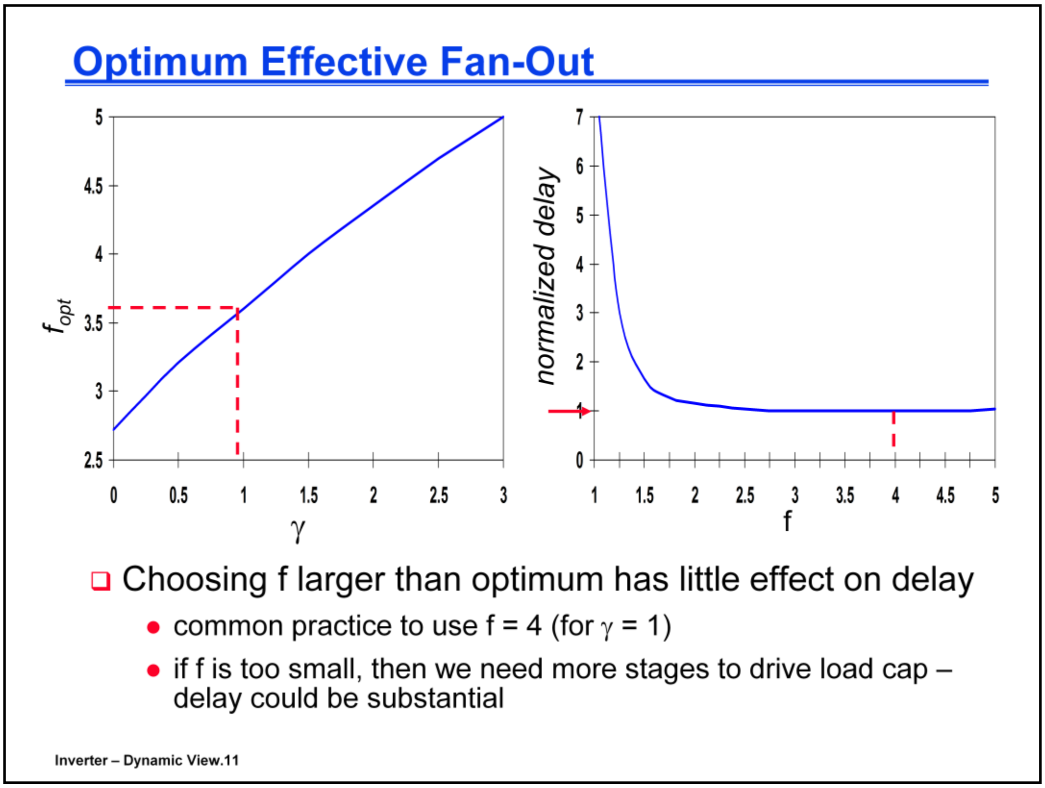
N = ln F

∴ $e^N = F$

But effective fan-out $f = F^{1/N}$, or $f^N = F$

∴ $e^N = f^N$

∴ f = e for optimum delay.



Optimum Effective Fan-Out

- ❑ Choosing f larger than optimum has little effect on delay
  - ● common practice to use f = 4 (for γ = 1)
  - ● if f is too small, then we need more stages to drive load cap – delay could be substantial

Inverter – Dynamic View.11

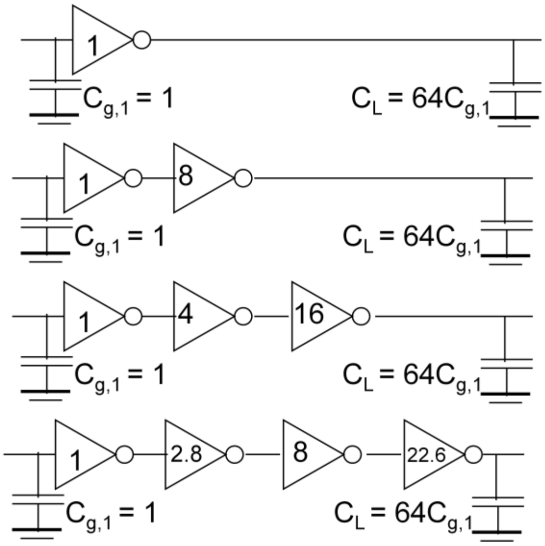
Rewrite $\gamma + F^{1/N} - \frac{1}{N}(\ln F)F^{1/N} = 0$ in terms of f, where f = $F^{1/N}$

$$\gamma + f - \frac{1}{N}(N \ln f) f = 0$$

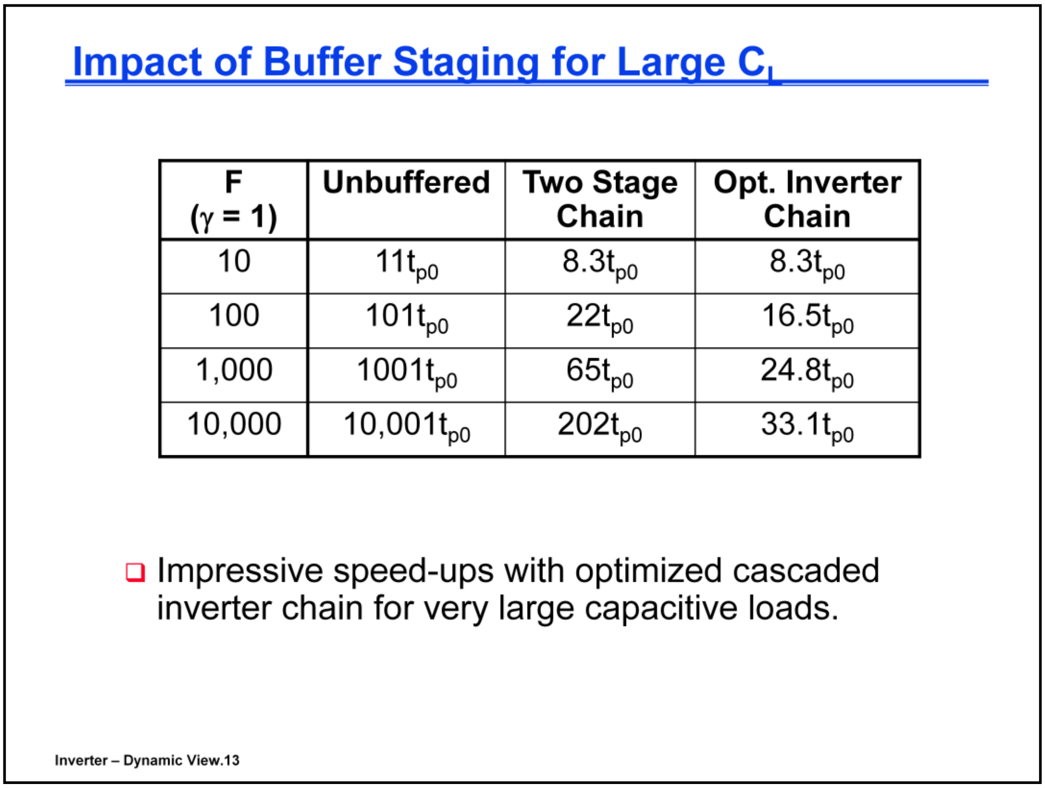
$$\gamma + f - f \ln f = 0$$

This equation has to be solved numerically to find optimum f.

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	$11t_{p0}$	$8.3t_{p0}$	$8.3t_{p0}$
100	$101t_{p0}$	$22t_{p0}$	$16.5t_{p0}$
1,000	$1001t_{p0}$	$65t_{p0}$	$24.8t_{p0}$
10,000	$10,001t_{p0}$	$202t_{p0}$	$33.1t_{p0}$

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

Exercise:

How many stages is needed to obtain minimum delay for F = 1,000?

How many stages is needed to obtain minimum delay for F = 10,000?