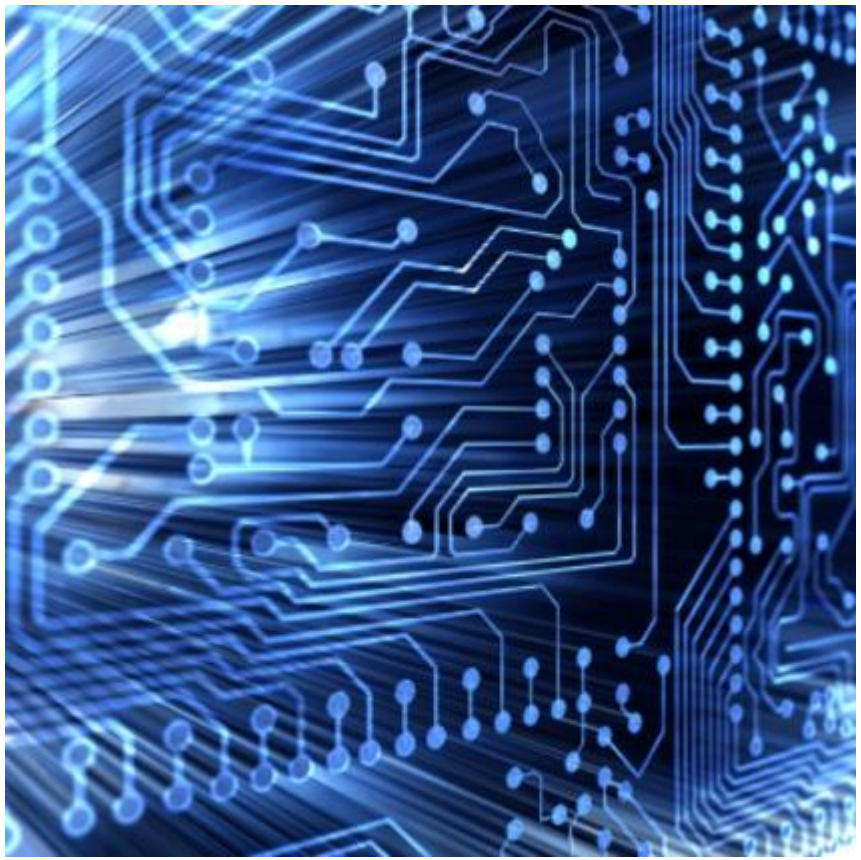
Digital Microelectronic Circuits (361-1-3021)



Presented by: Dr. Alex Fish

Lecture 5:

Parasitic Capacitance and Driving a Load

Motivation

- Thus far, we have learned how to model our essential building block, the MOSFET transistor, and how to use these building blocks to create the most popular logic family, Static CMOS.
- We analyzed the characteristics of a static CMOS inverter, including its Static and Dynamic Properties.
- We saw that both the delay and the power consumption of a CMOS gate depend on the load capacitance of the gate.

$$t_{pd} = 0.69 R_{eq} C_{Load} \qquad P_{dynamic} = f \cdot C \cdot V_{DD}^2$$

What will we learn today?

- Today, we will go back to our MOSFET transistor to try and understand what parasitic capacitances are inherent to its structure.
- Then, we will develop a model for equivalent capacitance estimation for delay calculation of a CMOS inverter.
- Accordingly, we will examine the optimal sizing of a CMOS gate.
- And finally, we will develop a methodology for sizing a chain of inverters to drive a large load.

What will we learn today?

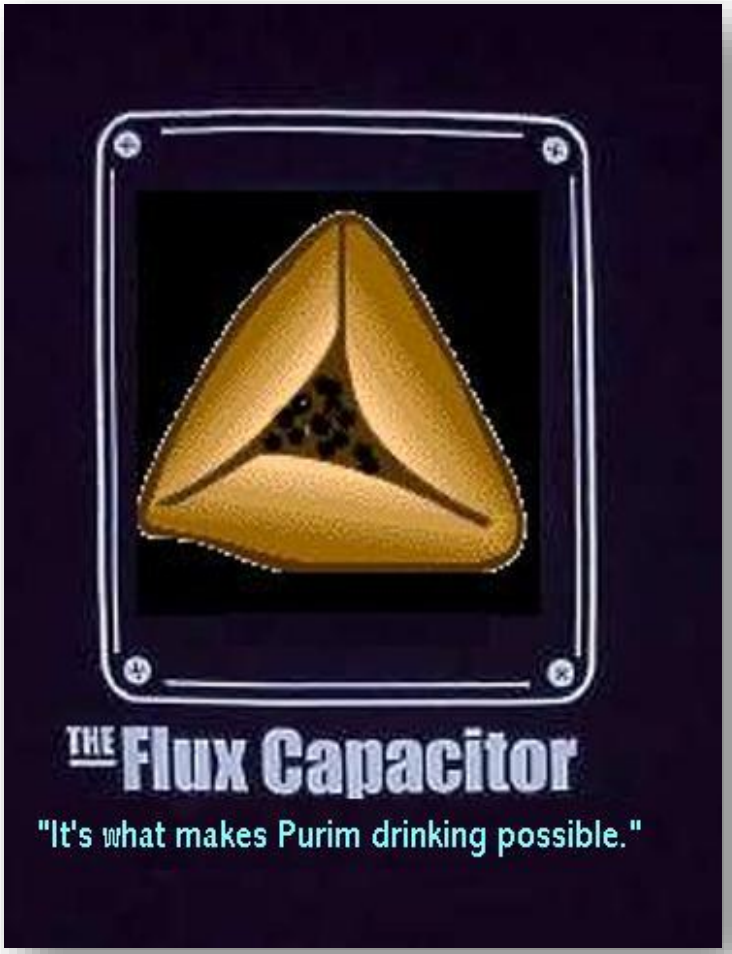
- 5.1 MOSFET Capacitance
- 5.2 Inverter Delay Capacitance Model
- 5.3 Driving a Load

5.2

5.1 MOSFET Capacitance

5.2 Inverter Delay Capacitance Model

5.3 Driving a Load

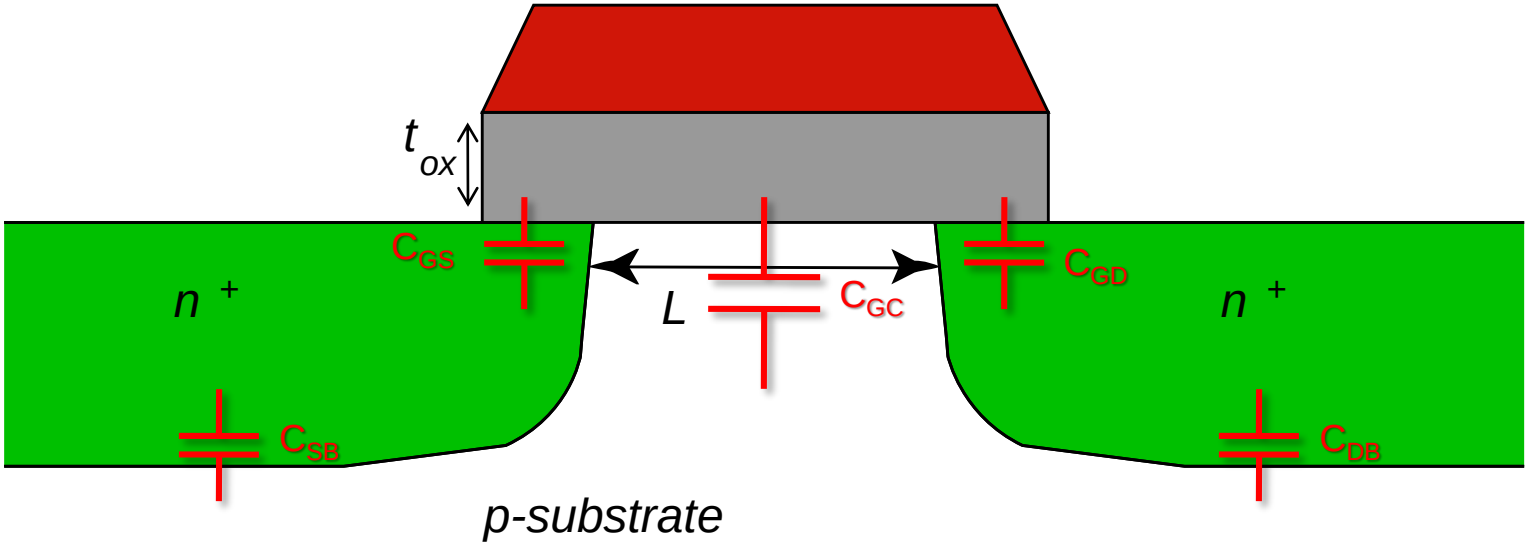


So back to our device, let's see what parasitic capacitances we have:

MOSFET CAPACITANCE

MOSFET Capacitance

- One of the important parameters of a ***MOS Transistor*** is its ***capacitance***.
- The ***MOSFET*** has two major categories of ***capacitance***:
  - ***Gate/Channel Capacitance*** – capacitance caused by the insulating oxide layer under the gate.
  - ***Junction Capacitance*** – pn-Junction capacitance between the diffusions and the substrate.



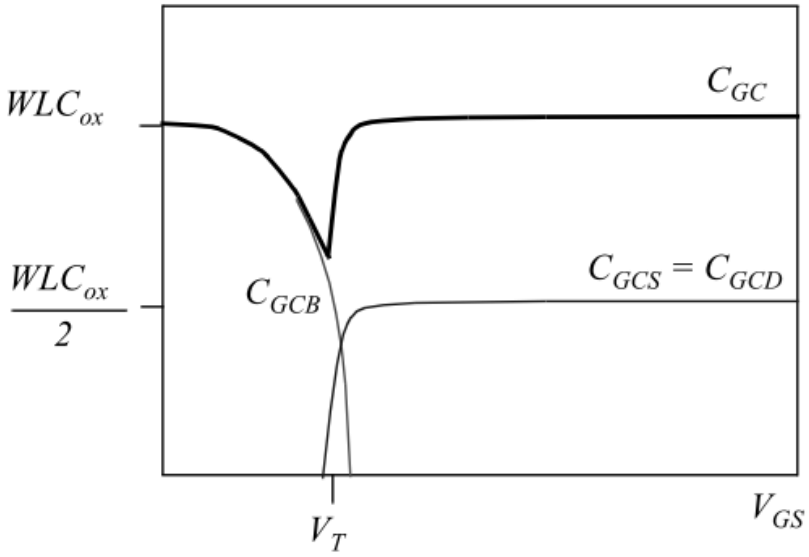
Gate Capacitance

- The ***Gate Capacitance*** includes:
  - » *Gate to Channel Capacitance*, C_{GC}:
    The main capacitance that is dependent on the region of operation.
    In general: $$C_{GC} = WL\frac{\varepsilon_{ox}}{t_{ox}} = C_{ox}WL$$
  - » *Gate Overlap Capacitance*, C_{GDO}, C_{GSO}:
    A constant (small) capacitance caused by gate overlap of the diffusions.
    $$C_{GSO} = C_{GDO} = WC_{Overlap}$$

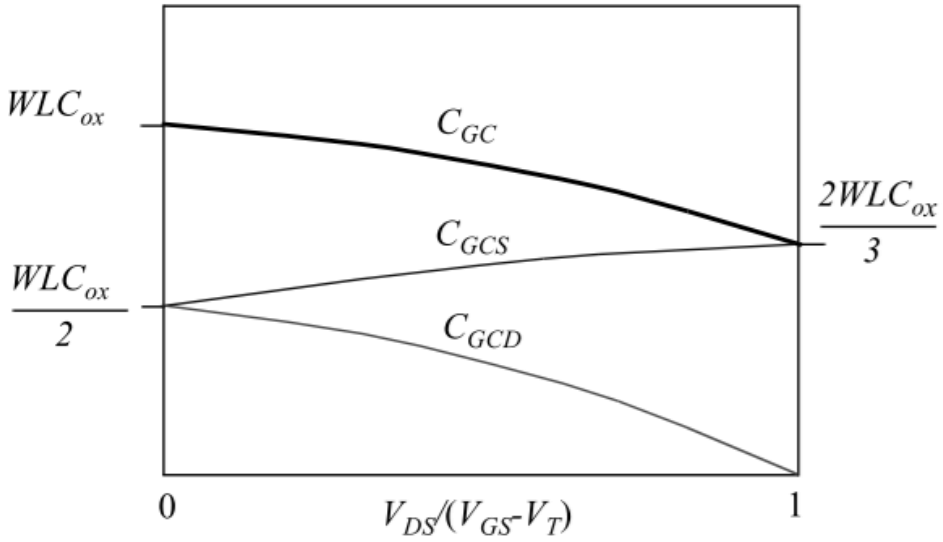
Gate Capacitance

- Looking at gate capacitance as a function of biasing shows how it changes.
  - In *accumulation*, the capacitance is across the oxide.
  - As V_{GS} grows, the depletion layer *decreases* the capacitance (as if the dielectric gets longer)
  - Once the channel is formed, the capacitance jumps.
  - At pinch-off, the drain capacitance drops to zero.

Gate Capacitance



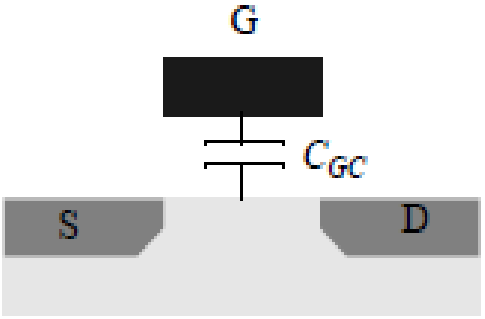
(a) C_{GC} as a function of V_{GS} (with V_{DS}=0)



(b) C_{GC} as a function of the degree of saturation

Gate Capacitance

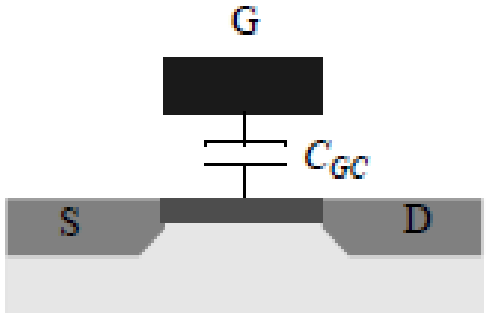
- To model this non-linear behavior, we will use the following approximations:



(a) cut-off

All capacitance is towards substrate

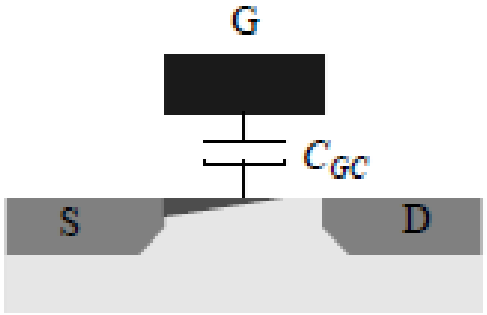
$$C_{GB} = C_{ox}WL$$



(b) resistive

Capacitance symmetrically divided between source and drain

$$C_{GCS} = C_{GCD} = \frac{C_{ox}WL}{2}$$



(c) saturation

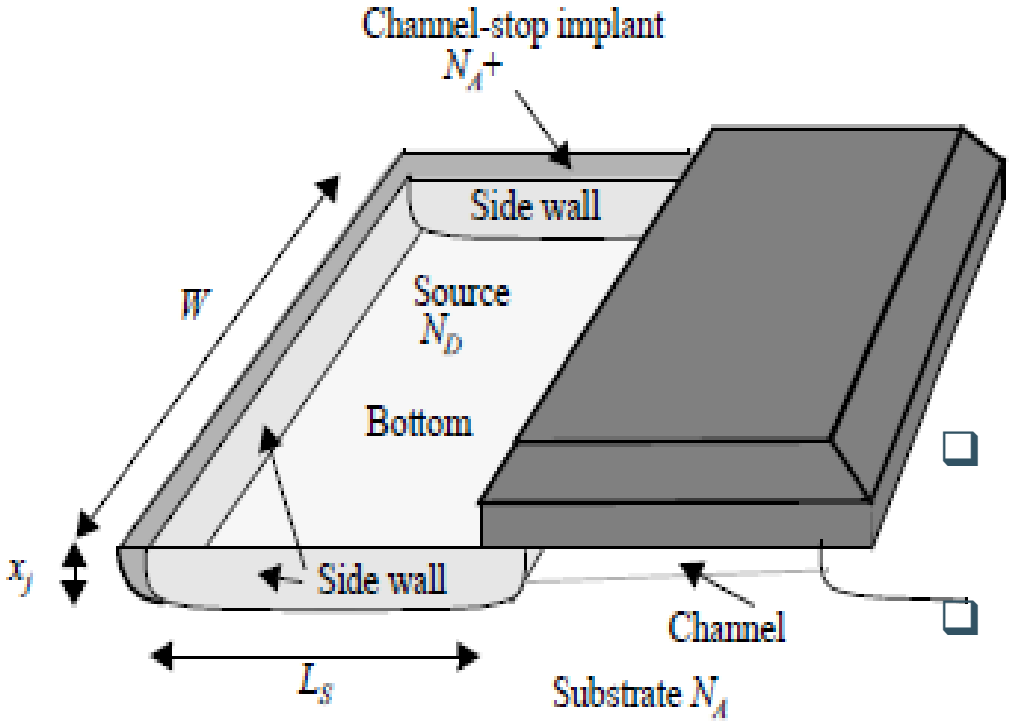
All capacitance to Source

$$C_{GCS} = \frac{2}{3}C_{ox}WL$$

Question: How do we relate to Velocity Saturation?

Junction (Diffusion) Capacitance

- The ***Junction Capacitance*** is the diffusion capacitance of the ***MOSFET***.
- This is measured according to fabrication parameters



$$C_{diff} = C_{bottom} + C_{side\,walls} =$$
$$= C_j \cdot Area + C_{jsw} \cdot Perimiter =$$
$$= C_j \cdot L_{diff} \cdot W + C_{jsw}\left(2L_{diff} + W\right)$$

- Diffusion cap is non-linear and voltage dependent.
- For simplicity, we will take it as constant in this course.

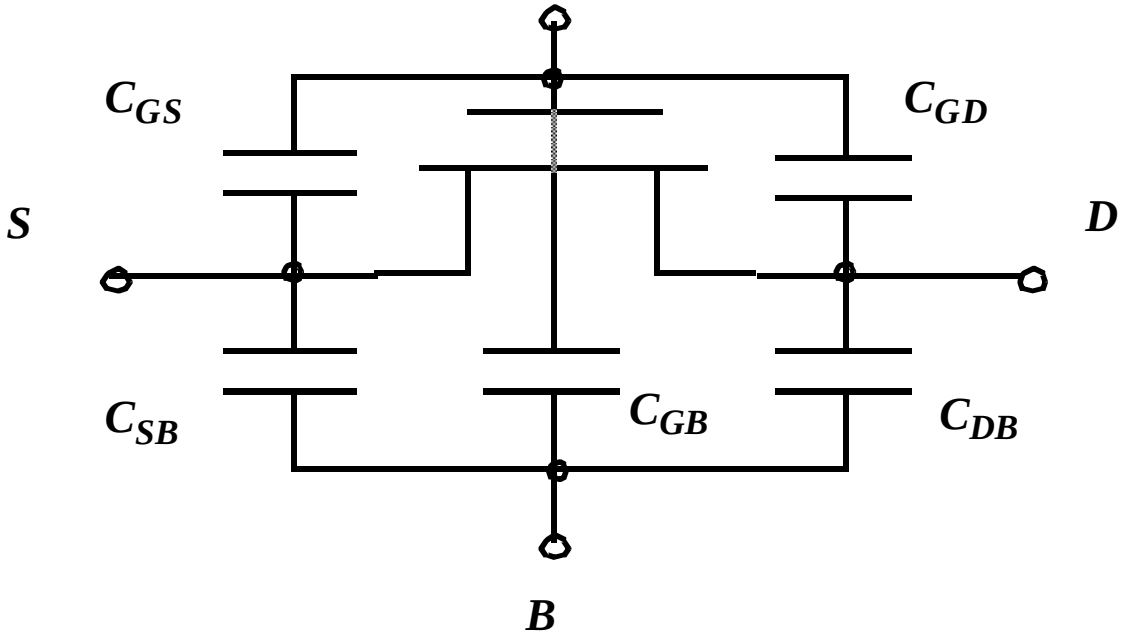
MOSFET Capacitance Summary

- Dependence of MOS capacitances on W and L:

MOSFET Capacitance Summary

Table 3.4 Average distribution of channel capacitance of MOS transistor for different operation regions.

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	$(2/3)C_{ox}WL$	$(2/3)C_{ox}WL+2C_oW$



$$C_{GS} = C_{GCS} + C_{GSO}$$

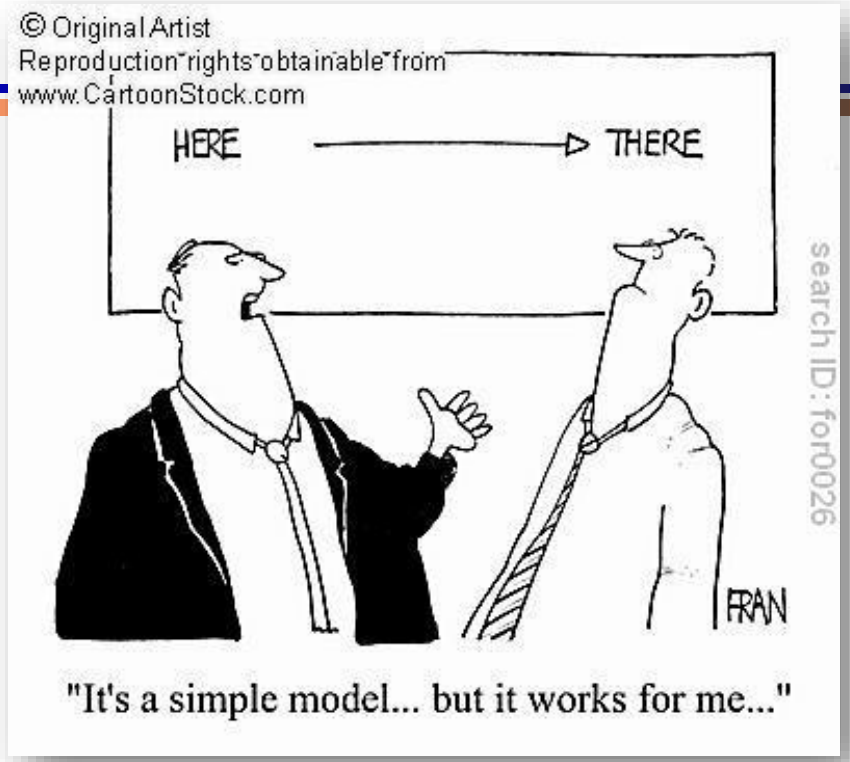
$$C_{GD} = C_{GCD} + C_{GDO}$$

$$C_{GB} = C_{GCB}$$

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

5.2

- 5.1 MOSFET Capacitance
- 5.2 Inverter Delay Capacitance Model
- 5.3 Driving a Load



OK, so we saw that the MOSFET has a bunch of non-linear parasitic capacitances, which makes them tough to use. To simplify life we'll now develop an:

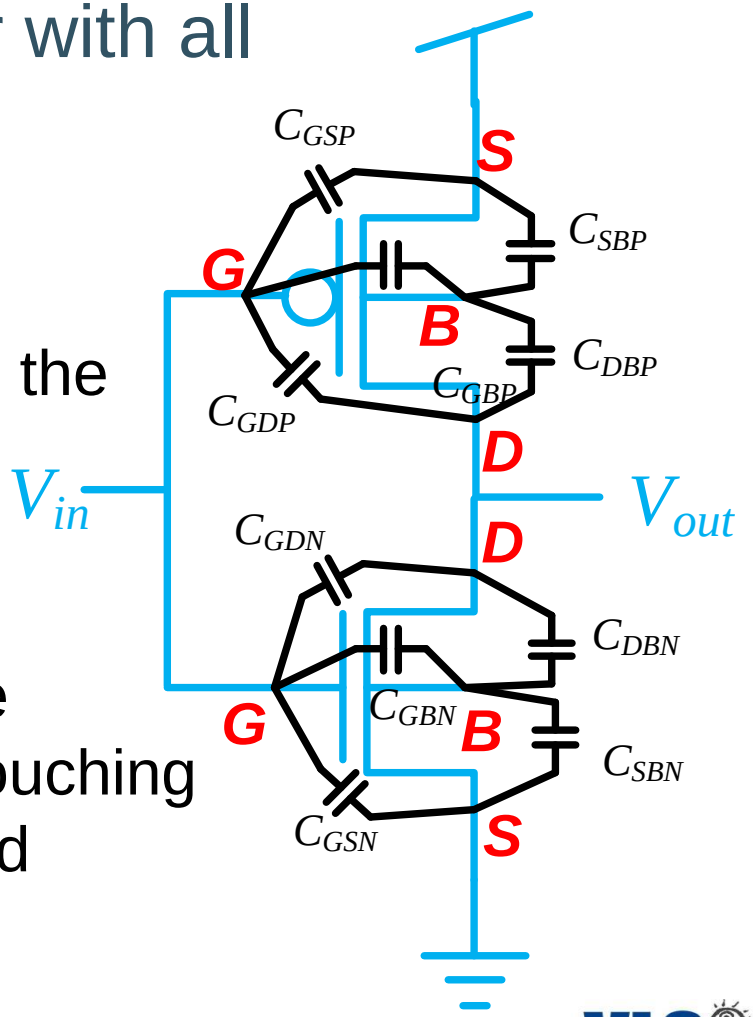
INVERTER DELAY CAPACITANCE MODEL

Capacitance Modeling

- As we saw, MOSFET capacitances are non-constant and non-linear.
- Therefore, it is hard to solve a general equation for an arbitrary transition/operation.
- Instead, we will develop a simple model that will approximate the capacitances during a specific transition that interests us.
- In this case, we are looking for the *Load Capacitance* to use when finding the *gate delay*.
- Therefore, we will apply a step function to the input of an inverter and approximate the capacitances according to the MOSFET parasitics we just learned.

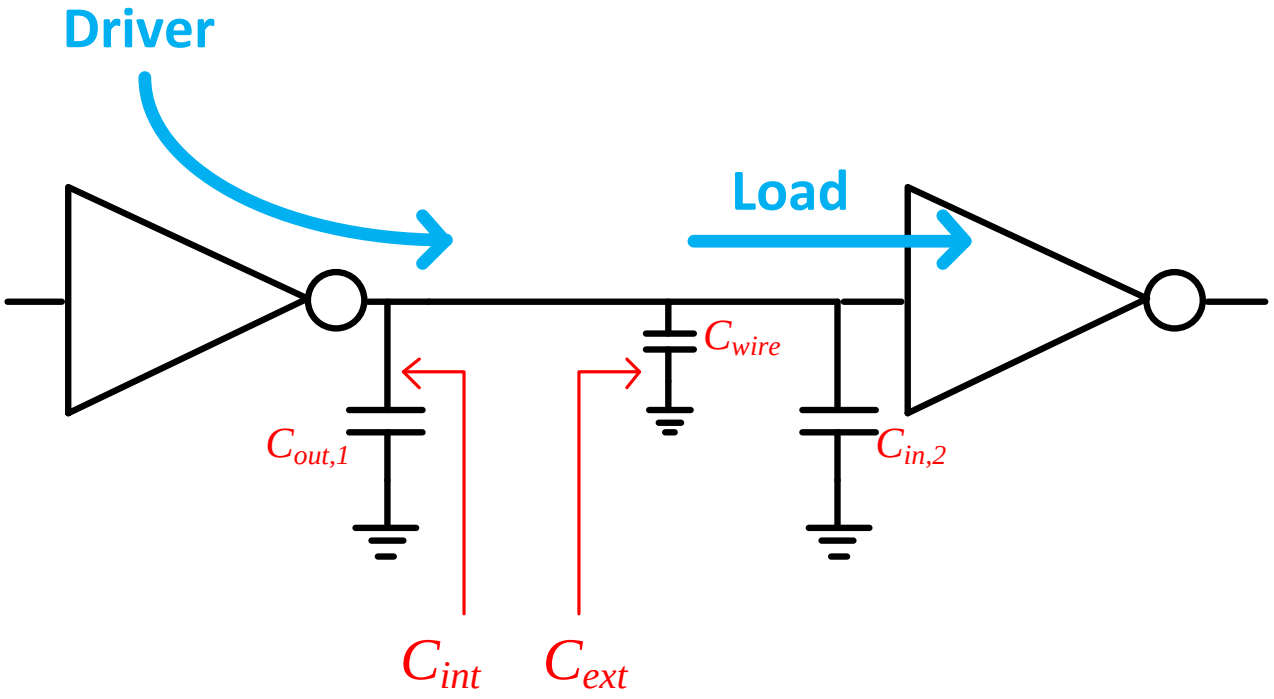
Capacitance Modeling

- Let's look at a CMOS inverter with all its parasitic capacitances:
  - » Considering the *Gates* of the transistors are the inputs to the inverter, any capacitor touching the gate should be considered *input capacitance.*
  - » Considering the *Drains* of the transistors are connected to the inverter output, any capacitor touching the Drains should be considered *output capacitance.*



Capacitance Modeling

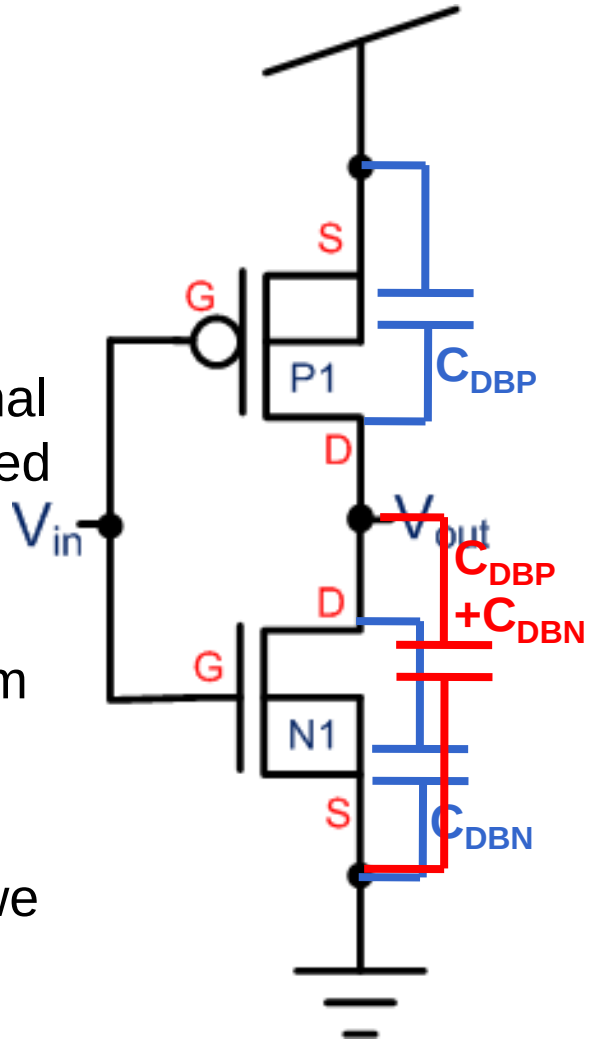
- We have to differentiate between output (intrinsic) capacitances and load (extrinsic) capacitances.



- Our total load capacitance is $C_{Load} = C_{out,1} + C_{wire} + C_{in,2}$

Intrinsic (Output) Capacitance

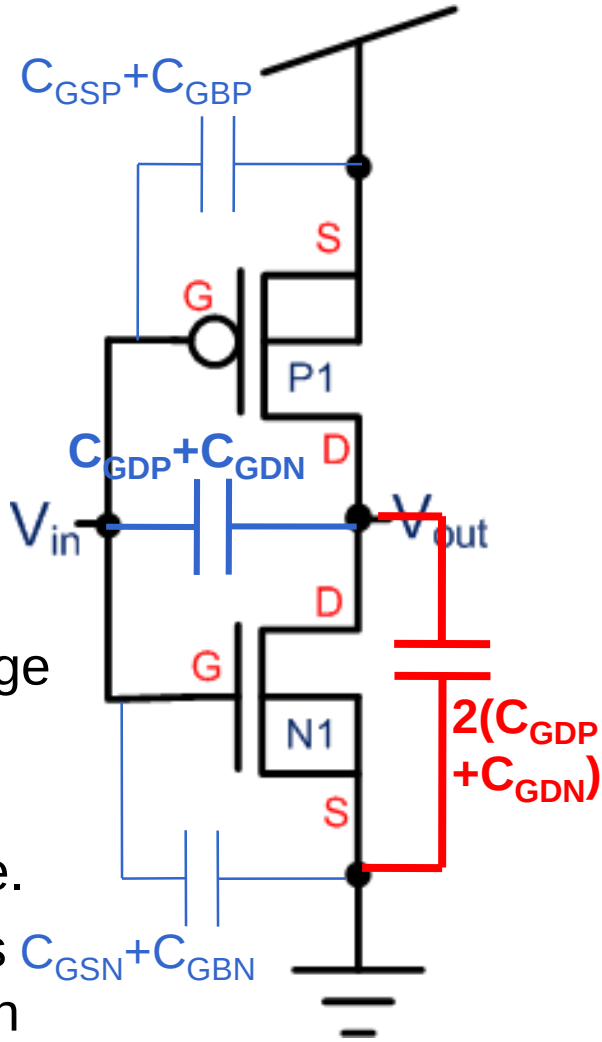
- We'll now look at what makes up the intrinsic output capacitance of the driver.
- This is primarily made up of ***diffusion*** capacitances:
  - » Both ***drain-to-body*** capacitances have a terminal with a constant voltage and the other connected to the output.
  - » For a simple computation, we will replace them with an ***equivalent capacitance to ground***.
  - » These capacitances are very non-linear and we will not go into their calculation in this course.



Intrinsic (Output) Capacitance

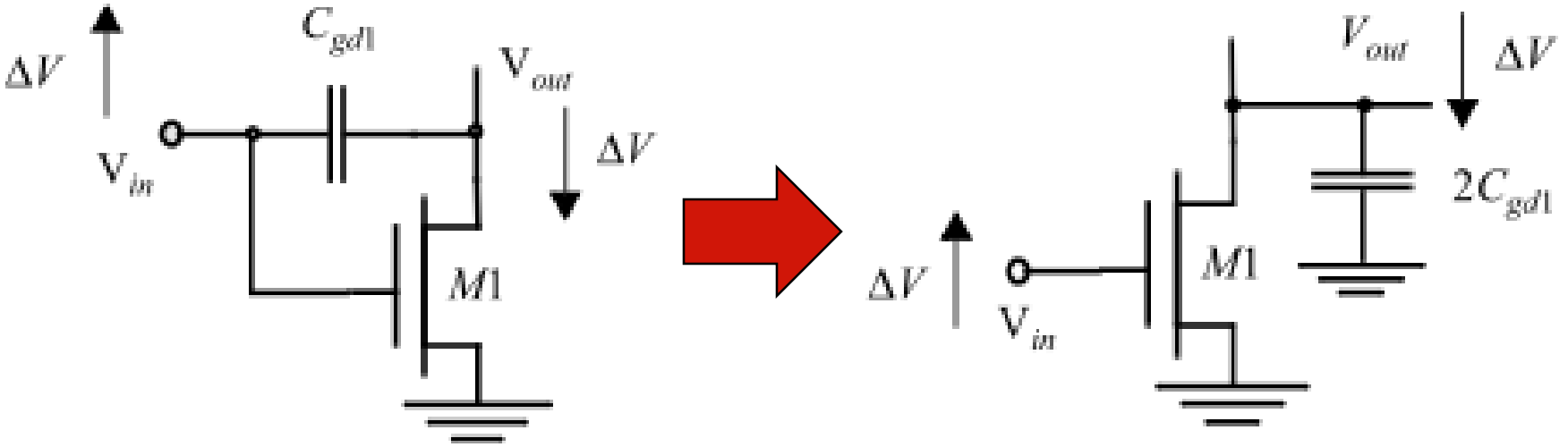
- How about *feedthrough* capacitance?
  - Taking the input step as ideal, the ***gate-to-source*** and ***gate-to-body*** capacitances don't contribute to the propagation delay.
  - The ***source-to-body*** capacitance is shorted to the supply, so it doesn't switch.

- What about the ***gate-to-drain*** capacitance?
  - While the gate voltage rises (V_{in}), the drain voltage drops (V_{out}) and vice versa.
  - According to the ***Miller Effect***, we can move this capacitance relative to ground, doubling its value.
  - This can be regarded as overlap capacitance, as for the majority of the transition the devices are in cutoff or saturation.



The Miller Effect

- A capacitor experiencing identical, but opposite voltage swings at both its terminals can be replaced by a capacitor to ground, whose value is twice the original value.



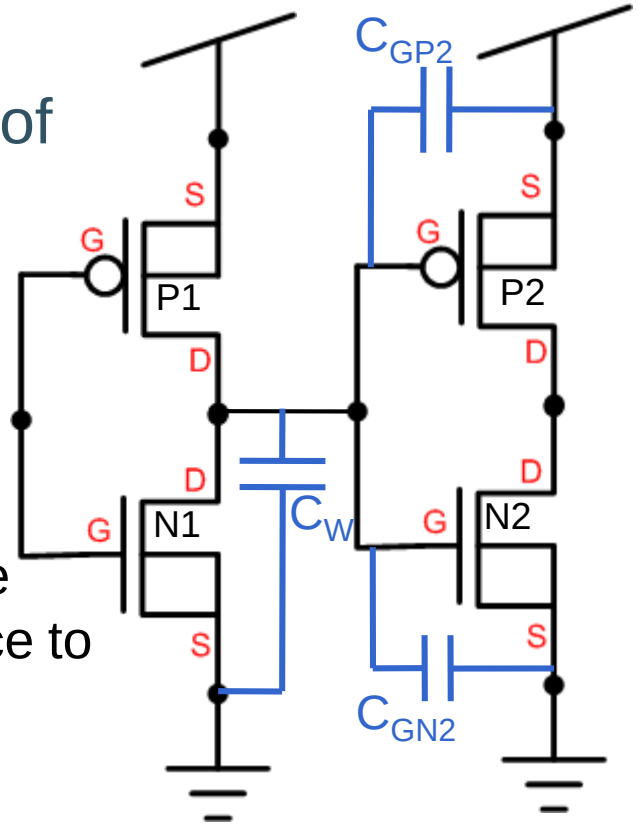
Summary of Intrinsic Output Cap

External Capacitance

- Now, to annotate the parasitics during switching, we will cascade another inverter after the first.
- We first add the ***Wire Capacitance***.
- Then we add the ***gate capacitances*** of the second inverter.
  - These are approximately the ***oxide capacitance*** times the ***area***:

$$C_{GN2} + C_{GP2} \approx C_{OX}\left(W_{N2}L_{N2} + W_{P2}L_{P2}\right)$$

  - Again, we can just add the ***pMOS*** gate capacitance to the general capacitance to ground.



External Capacitance

- What happened to overlap capacitance and the Miller Effect on C_{GD2}?
  - Remember that this is an approximate model, but…
  - $L=L_{eff}+2*L_{ov}$, so $C_G=C_{ox}*W*L_{drawn}$.
  - During the transient, for most of the time the load gate's transistors are in cutoff or in linear.
  - Miller effect won't appear, because the second gate won't switch until t_{pd1} is over.
- Therefore:
  - YES, C_{GD2} and C_{GS2} contribute to the load capacitance.
  - BUT, a good approximation is just $C_{G2}=C_{OX}*W*L$

Summary of Next Stage Input Cap

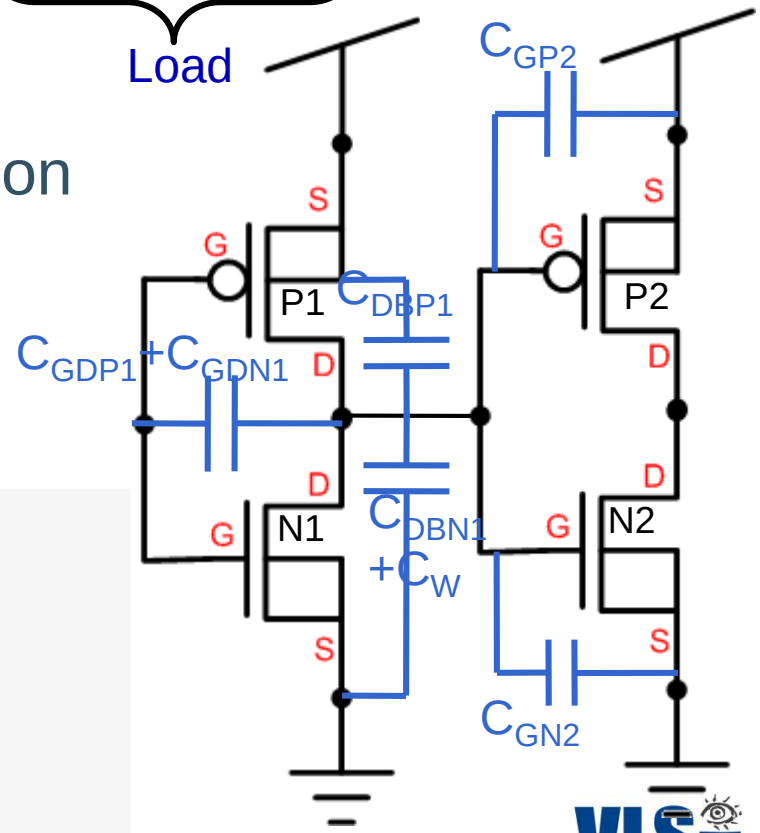
Parasitic Capacitances - Summary

- Altogether, as a very general approximation, we get:

$$C_{load} = \underbrace{2(C_{GDP1} + C_{GDN1})}_{\text{Miller}} + \underbrace{(C_{DBP1} + C_{DBN1})}_{\text{Diffusion}} + \underbrace{(C_{GP2} + C_{GN2})}_{\text{Load}} + C_W$$

- An even more general approximation with ***N fan-out*** gates gives us:

$$C_{load} = C_{out} + C_{wire} + N \cdot C_{in}$$



Last Time…

- CMOS Inverter Capacitance Model for t_{pd}.

Last Time…

- MOS Capacitance Model $C_{load} = C_{out} + C_{wire} + N \cdot C_{in}$
  - » Driver Cap (C_{out} or C_{int}): Diffusion + Miller
  - » Load Cap (C_{in} or C_g): Gate cap, no miller

5.3

5.1 MOSFET Capacitance

5.2 Inverter Delay Capacitance Model

5.3 Driving a Load

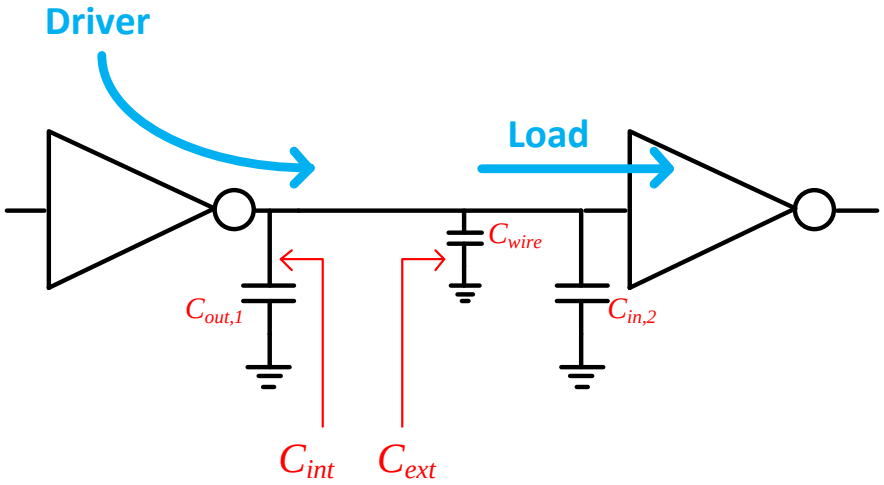
Up till now, we discussed device sizing with an optimal fanout of 1. What happens if we want to cascade more gates to the output?

DRIVING A LOAD

External Capacitance

- Up till now, we assumed our inverter was only driving a copy of itself. This is known as ***intrinsic*** or ***unloaded delay.***
- But we usually will have a larger fanout, and in some cases, we will need to drive large loads.
- Let's remember how we defined our load capacitance:

$$C_{load} = \left(C_{diff} + C_{overlap}\right) + \left(C_{fanout} + C_{wire}\right) \triangleq C_{int} + C_{ext}$$



- We can now write our delay equation according to these components.

$$t_{pd} = 0.69 R_{eq} C_{load} = 0.69 R_{eq} \left(C_{\text{int}} + C_{ext}\right)$$

Sizing Factor (S)

$$t_{pd} = 0.69 R_{eq} \left(C_{\text{int}} + C_{ext} \right)$$

- This means that if we add a larger load, our delay will increase. This is intuitive, as it means we have to supply more current from the same source.
- If we were to *widen* our transistors by a factor *S*, this would *decrease* our resistance and *increase* our *intrinsic* capacitance.

$$C_{\text{int}}^{*} = S \cdot C_{\text{int}} \qquad R_{eq}^{*} = R_{eq} / S$$

Sizing Factor (S)

$$C_{\text{int}}^{*} = S \cdot C_{\text{int}} \qquad R_{eq}^{*} = R_{eq} / S$$

$$t_{pd} = 0.69 R_{eq} \left(C_{\text{int}} + C_{ext} \right)$$

These two factors trade-off, which is why we get an optimal inverter size.

$$t_{pd,unloaded}^{*} = 0.69 R_{eq}^{*} C_{\text{int}}^{*} = 0.69 \left(R_{eq} / S \right) \left(S \cdot C_{\text{int}} \right) = 0.69 R_{eq} C_{\text{int}}$$

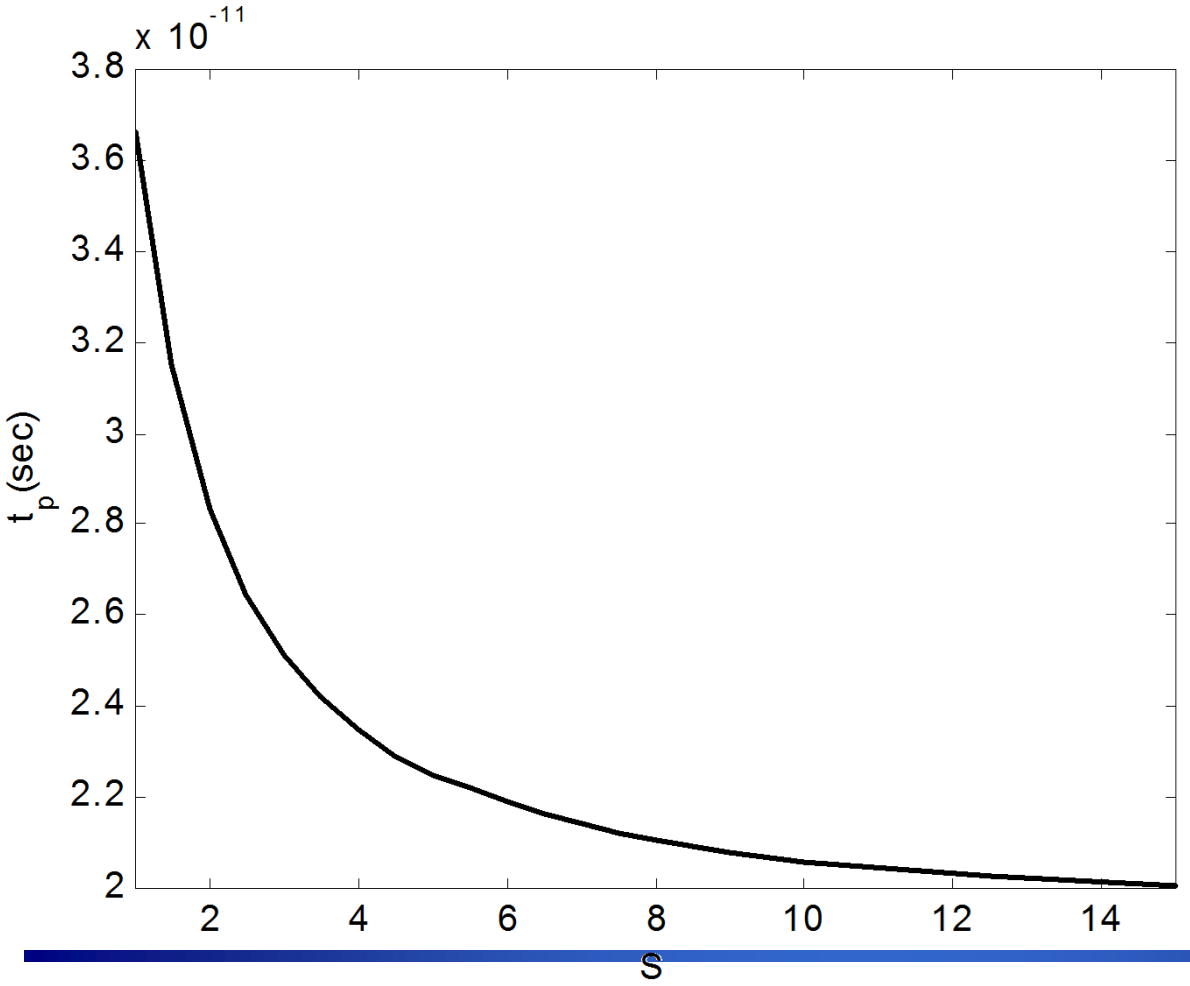
- However, upsizing our gate doesn't affect the external capacitance and therefore decreases the loaded delay.

$$t_{pd}^{*} = 0.69 \frac{R_{eq}}{S} \left(S C_{\text{int}} + C_{ext} \right) = 0.69 R_{eq} C_{\text{int}} \left(1 + \frac{C_{ext}}{S C_{\text{int}}} \right) = t_{p0} \left(1 + \frac{C_{ext}}{S C_{\text{int}}} \right)$$

Sizing Factor (S)

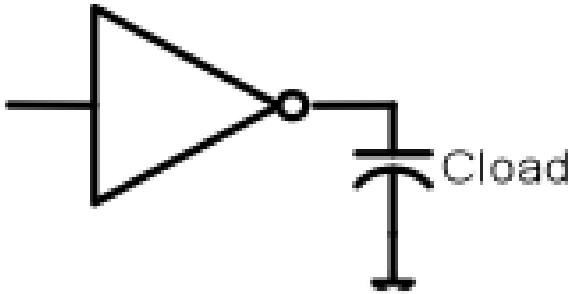
$$t_p = t_{p0}\left(1 + {C_{ext}}/{SC_{\text{int}}}\right)$$

$$t_{p0} \equiv 0.69 R_{eq} C_{\text{int}}$$



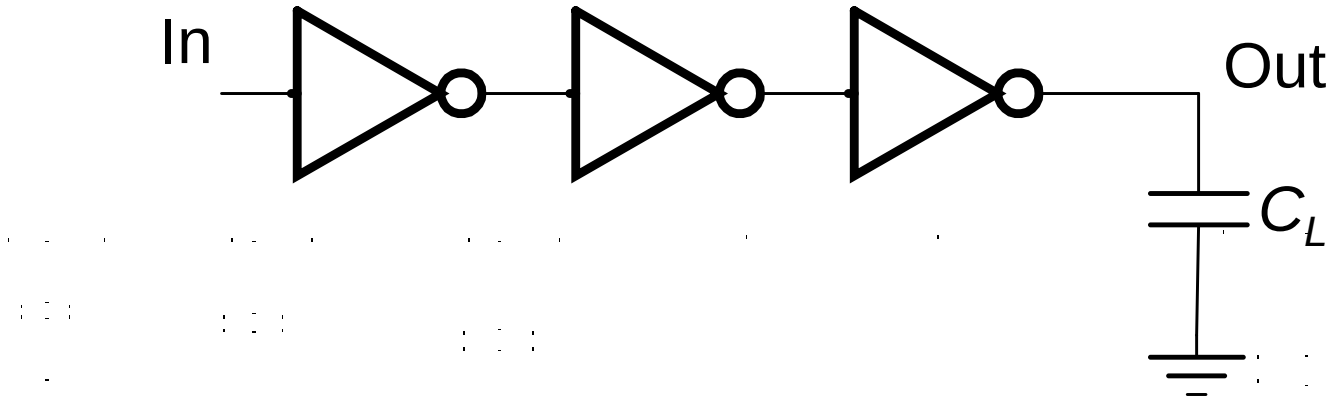
Driving a Large Load

- So now we have a very large load to drive.



- We could just use a *very large* inverter.
  - But then someone would have to drive this large inverter.
- So considering we start with a limited input capacitance, how should we best drive this load?

Inverter Chain

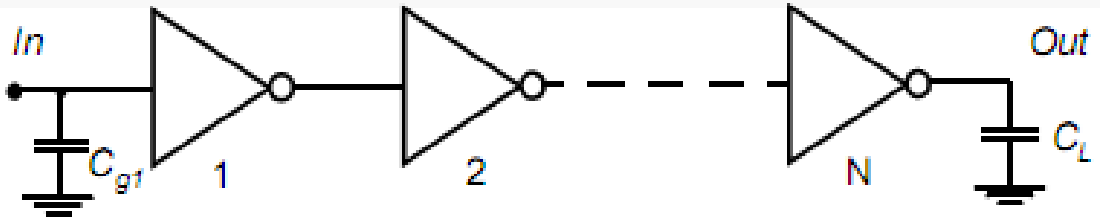


- If C_L is given:
  - » How many stages are needed to minimize the delay?
  - » How to size the inverters?

- Anyone want to guess the solution?

Delay Optimization Problem #1

- To solve an optimization problem, we need a set of constraints:
  - Load Capacitance.
  - Number of Inverters.
  - Size of input capacitance.



Delay Optimization Problem #1

- To explore this problem, we must define a proportionality factor, γ.

- γ is a function of technology*, that describes the relationship between a gate's input gate capacitance (C_g) and its intrinsic output capacitance (C_{int}):

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

* γ is close to *1* for most submicron processes!

Delay Optimization Problem #1

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

- Now, we will write the delay as a function of γ, and the ***effective fanout, f:***

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

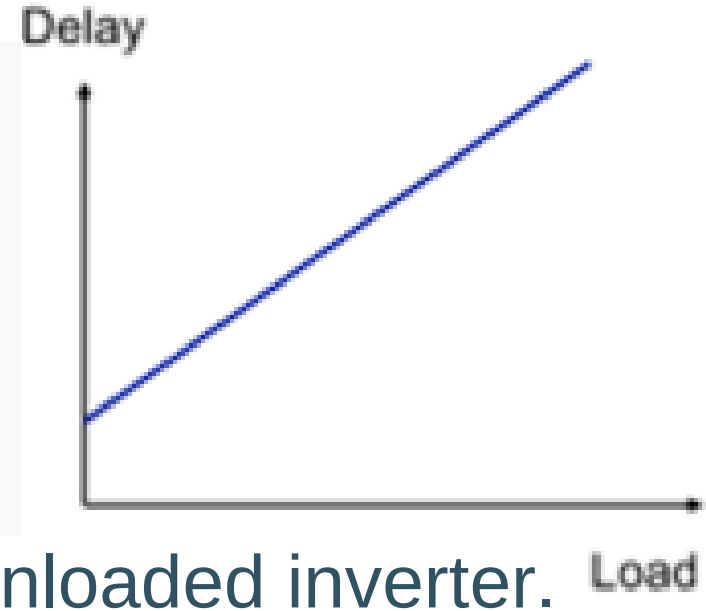
$$f \triangleq \frac{C_{ext}}{C_g}$$

- We can see that the delay for a certain technology is only a function of the effective fanout!

Inverter with Load

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

- So we see that the delay increases with ratio of load to inverter size:

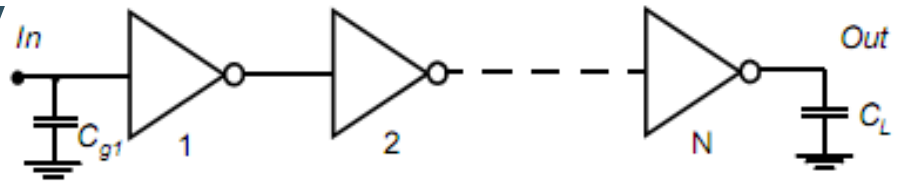


- t_{p0} is the intrinsic delay of an unloaded inverter.
- γ is a technology dependent ratio.
- f is the ***Effective Fanout*** – ratio of load to inverter size

Sizing a chain of inverters

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

- Now we will express the delay of a chain of inverters:



- Assuming a negligible wire capacitance, for the ***j***-th stage, we get:

$$t_{pd,j} = t_{p0}\left(1+\frac{f_j}{\gamma}\right) = t_{p0}\left(1+\frac{C_{g,(j+1)}}{\gamma C_{g,j}}\right)$$

- And we can write the total delay as:

$$t_{pd} = \sum_{j=1}^{N} t_{pd,j} = t_{p0}\sum_{j=1}^{N}\left(1+\frac{C_{g,(j+1)}}{\gamma C_{g,j}}\right)$$

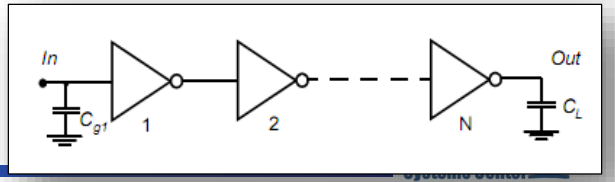
Sizing a chain of inverters

$$t_{pd} = t_{p0}\sum_{j=1}^{N}\left(1+\frac{C_{g,(j+1)}}{\gamma C_{g,j}}\right)$$

- We have ***N-1*** unknowns, so we will derive ***N-1*** partial derivatives: $\frac{\partial t_{pd}}{\partial C_{g,j}} = 0$
- We receive a set of constraints:

$$\frac{C_{g,(j+1)}}{C_{g,j}} = \frac{C_{g,j}}{C_{g,(j-1)}} \Rightarrow C_{g,j} = \sqrt{C_{g,(j+1)}C_{g,(j-1)}}$$

- This means that:
  - Each inverter is sized up by the same factor, f.
  - Each inverter has the same effective fanout, $f_j = f$.
  - Each inverter has the same delay, $t_{p0}(1+f/\gamma)$.



Sizing a chain of inverters

- Now this is interesting…
  what if we multiply the fanout of each stage?:

$$f^N = \prod_{j=1}^{N} f_j = \prod_{j=1}^{N} \frac{C_{g,(j+1)}}{C_{g,j}} = \frac{C_{g,2}}{C_{g,1}} \cdot \frac{C_{g,3}}{C_{g,2}} \cdot \frac{C_{g,4}}{C_{g,3}} \cdots \frac{C_{g,N}}{C_{g,N-1}} \cdot \frac{C_{load}}{C_{g,N}} = \frac{C_{load}}{C_{g,1}}$$

- We found the ratio between input and load capacitance!

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

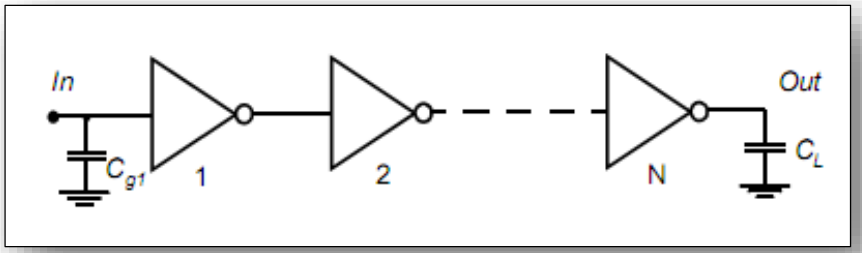
$$F \triangleq \frac{C_{load}}{C_{g,1}}$$

Sizing a chain of inverters

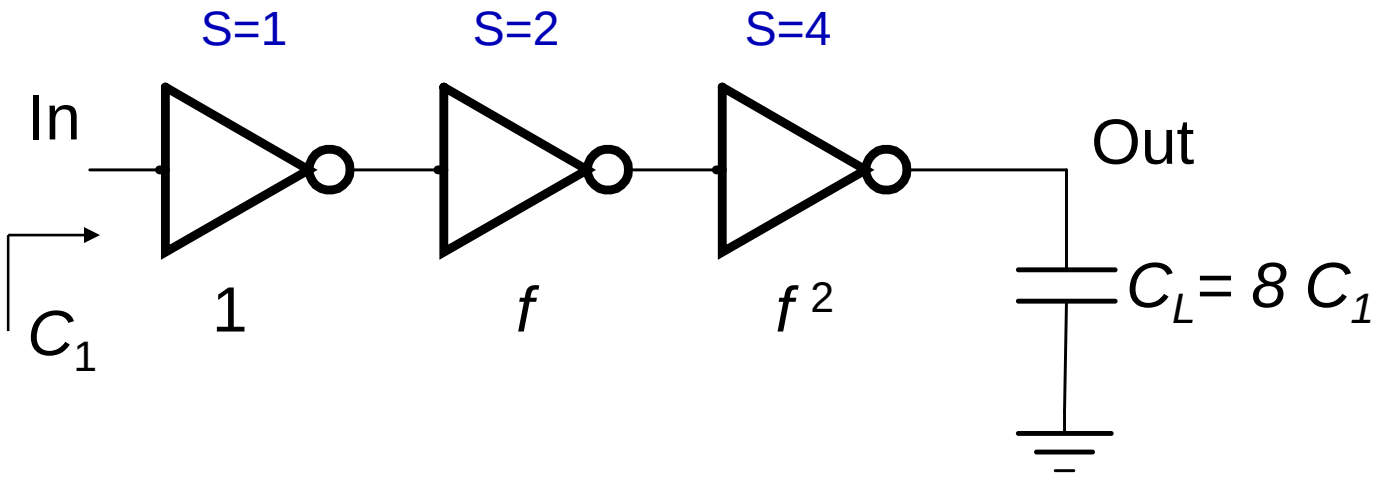
$$f = \sqrt[N]{F} \qquad F \triangleq \frac{C_{load}}{C_{g,1}}$$

- We defined the ***overall effective fanout***, ***F***, between the input and load capacitance of the circuit. Using this parameter, we can express the total delay:

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



Example: 3 Stages



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

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

Delay Optimization Problem #2

- Great – but what is the optimal Number of Stages?
- This is a new Optimization Problem. You are given:
  - The size of the first inverter
  - The size of the load that needs to be driven
- Your goal:
  - Minimize delay by finding optimal number and sizes of gates
- So, need to find N that minimizes: $t_p = N \cdot t_{p0}\left(1+\frac{\sqrt[N]{F}}{\gamma}\right)$

Delay Optimization Problem #2

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

- Starting with a minimum sized inverter with $C_{g,min}$, and driving a given load, C_{load}, we can see that:
  - With a small number of stages, we get a large delay due to the effective fanout (f, F).
  - With a large number of stages, we get a large delay due to the intrinsic delay (Nt_{p0})

- To find the optimal number of stages, we will differentiate and equate to zero.

$$\frac{dt_{pd}}{dN} = 0 \rightarrow \gamma + \sqrt[N]{F} - \frac{\sqrt[N]{F}\ln F}{N} = 0 \rightarrow f_{opt} = \exp\left(1 + \gamma / f_{opt}\right)$$

Delay Optimization Problem #2

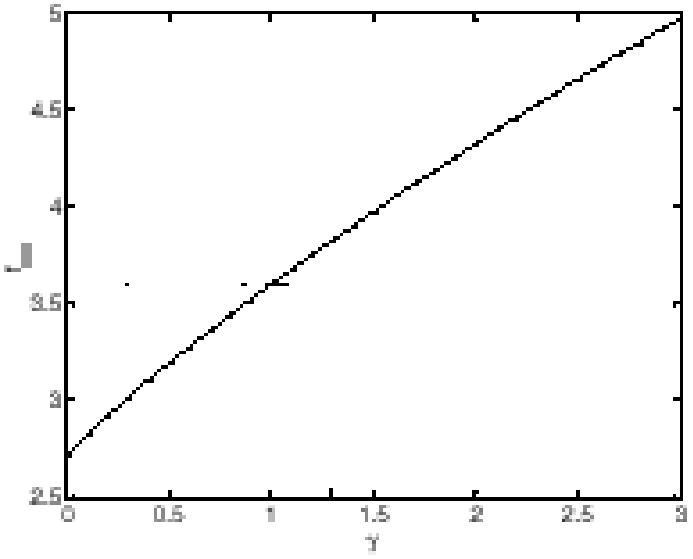
$$f_{opt} = \exp\left(1 + \gamma / f_{opt}\right)$$

- This equation only has an analytical solution for the case of $\gamma=0$. In this esoteric case we get:

$$f_{opt}\Big|_{\gamma=0} = e = 2.718 \qquad N_{opt}\Big|_{\gamma=0} = \ln F$$

- If we take the typical case of $\gamma=1$, we can numerically solve the equation and arrive at:

$$f_{opt}\Big|_{\gamma=1} \approx 3.6 \qquad N_{opt}\Big|_{\gamma=1} \approx \log_{3.6} F$$



Example

- We are given: $C_L = 64C_{\min}$

  $C_{in} = C_{\min}$

- We need to find the optimal number of stages, so $f_{opt}=4$.

$$F = C_L / C_{in} = 64$$

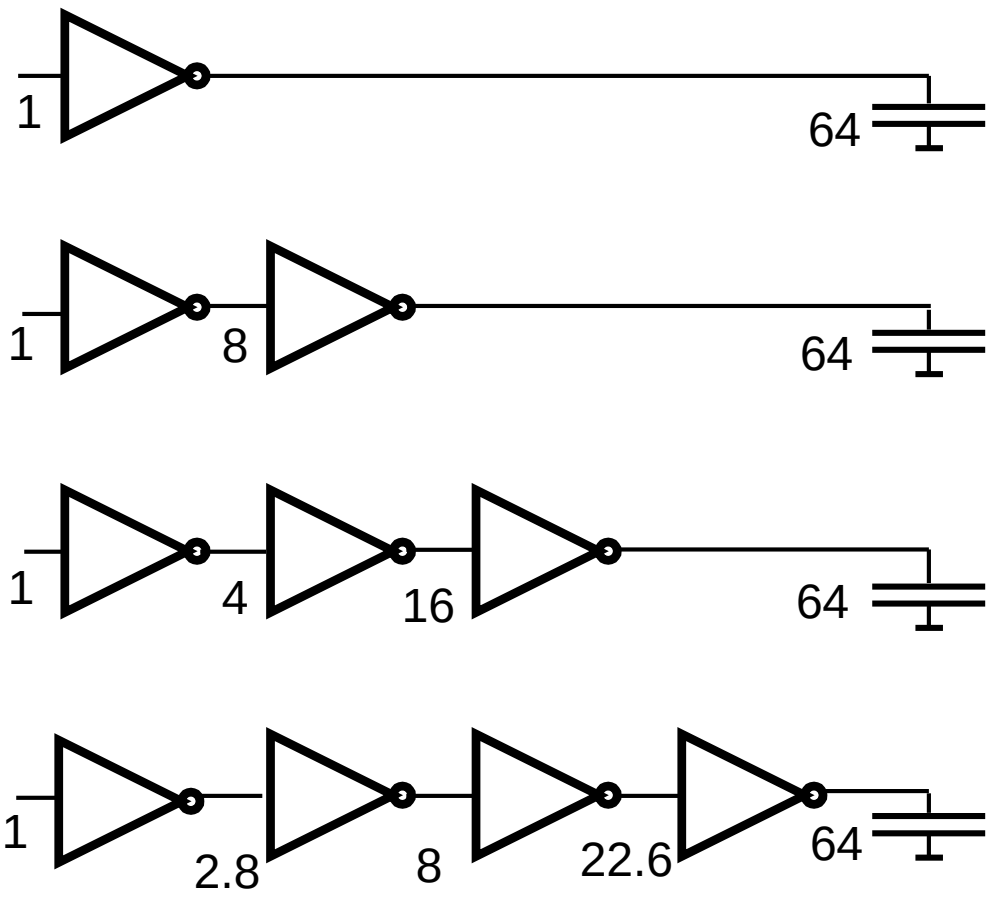
$$N_{opt} = \log_{f_{opt}} F = \log_4 64 = 3$$

- So we need 3 stages that will be sized 1, 4, 16.

- Let's inspect what delay we would have gotten with various number of stages.

Example

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



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

Normalized delay function of F

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

- Let's consider the trade offs of driving loads with various approaches:
  - Unbuffered
  - Two-Stages
  - Optimal FO4 Chain

- For a small load, F=10:
  - Unbuffered delay: $t_{pd}=11t_{p0}$
  - Two Stage delay: $t_{pd}=8.3t_{p0}$
  - Optimal FO4 Chain: $N_{opt}=2 \rightarrow t_{pd}=8.3t_{p0}$

Normalized delay function of F

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

- For a slightly larger load, F=100:
  - Unbuffered delay: $t_{pd}=101t_{p0}$
  - Two Stage delay: $t_{pd}=22t_{p0}$
  - Optimal FO4 Chain: $N_{opt}=4 \rightarrow t_{pd}=16.6t_{p0}$
  - *We see a large benefit for one extra stage, but the inverter chain might not be worth it.*
- How about a really large load, F=10000:
  - Unbuffered delay: $t_{pd}=10001t_{p0}$
  - Two Stage delay: $t_{pd}=202t_{p0}$
  - Optimal FO4 Chain: $N_{opt}=7 \rightarrow t_{pd}=33.1t_{p0}$
  - *But we "pay" for this with a large number of stages and huge final stage inverter (W=1mm)*

What about Energy (and Area)?

- How much additional Energy (and Area) does an inverter chain cost?
  - » Using a single (minimal) inverter, our capacitance would be:

$$C_{1-stage} = C_{\min} + \gamma C_{\min} + C_L$$

  - » But the capacitance of *N* stages

$$C_{N-stages} = (1+\gamma)C_{\min} + (1+\gamma)fC_{\min} + \ldots + (1+\gamma)f^{N-1}C_{\min} + C_L$$

$$= C_{1-stage} + \underbrace{(1+\gamma)fC_{\min} + \ldots + (1+\gamma)f^{N-1}C_{\min}}_{\text{Overhead!}}$$

What about Energy (and Area)?

- So the overhead cap is:

$$C_{overhead} = (1+\gamma)\, fC_{\min}\left(1+ f +\ldots+ f^{N-2}\right)$$

$$= (1+\gamma)\, fC_{\min}\left(\frac{f^{N-1}-1}{f-1}\right)$$

- For example:

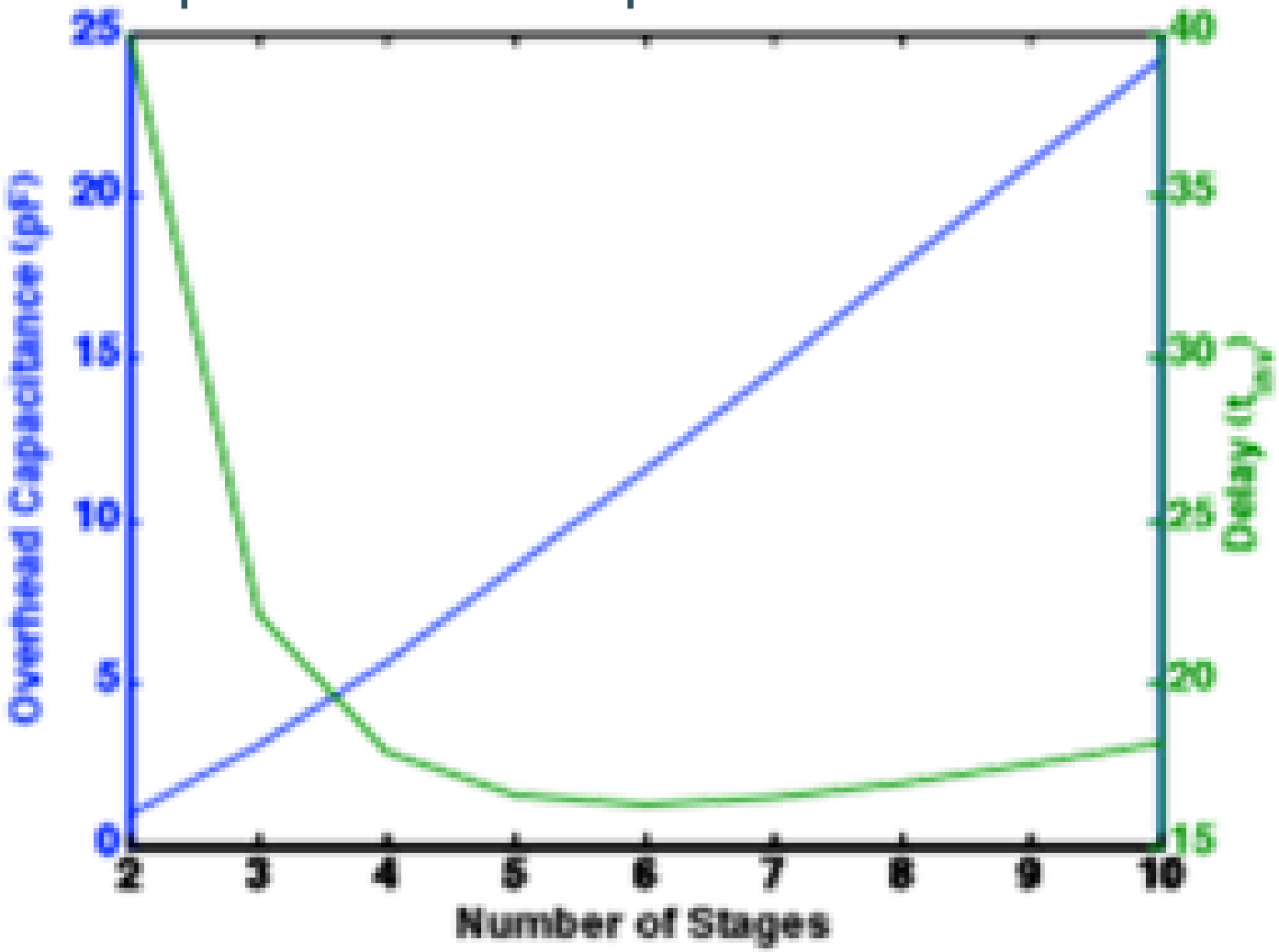
$$C_L = 20\,pF,\ C_{\min} = 50\,fF$$

$$F = \frac{20\,pF}{50\,fF} = 400,\ N_{opt} = 5,\ f = \sqrt[5]{400} = 3.3$$

$$C_{overhead} = 2 \cdot 3.3 \cdot 50 \cdot 10^{-15}\left(\frac{117.6}{4}\right) = 9.7\,pF$$

Example Overhead Numbers

- For the previous example:



Conclusions

- In order to drive a large load, we should:
  - Use a chain of inverters.
  - Each inverter should increase its size by the same amount.
  - To minimize the delay, we should set the effective fanout to about 4.
- Remember!
  - You have to use a whole number of stages (i.e. you can't choose 2.5 stages).
  - Therefore, choose the closest number of stages for close to optimal effective fanout..
  - Choose according to signal polarity or optimal speed.
  - But fewer stages means less power!