ECE 715
System on Chip Design and Test

Lecture 5: SoC Physical Design Issues

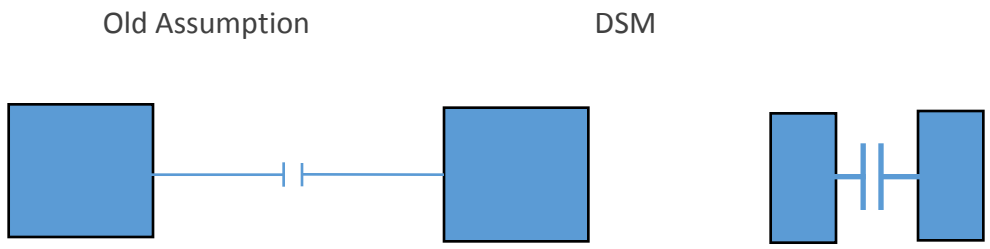
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Design Challenges

1. Non-scalable global wire delay
2. Moving signals across a large die within one clock cycle is not possible.
3. Current interconnection architecture- Buses are inherently non-scalable.
4. Transmission of digital signals along wires is not reliable.

Interconnect Scaling Effects

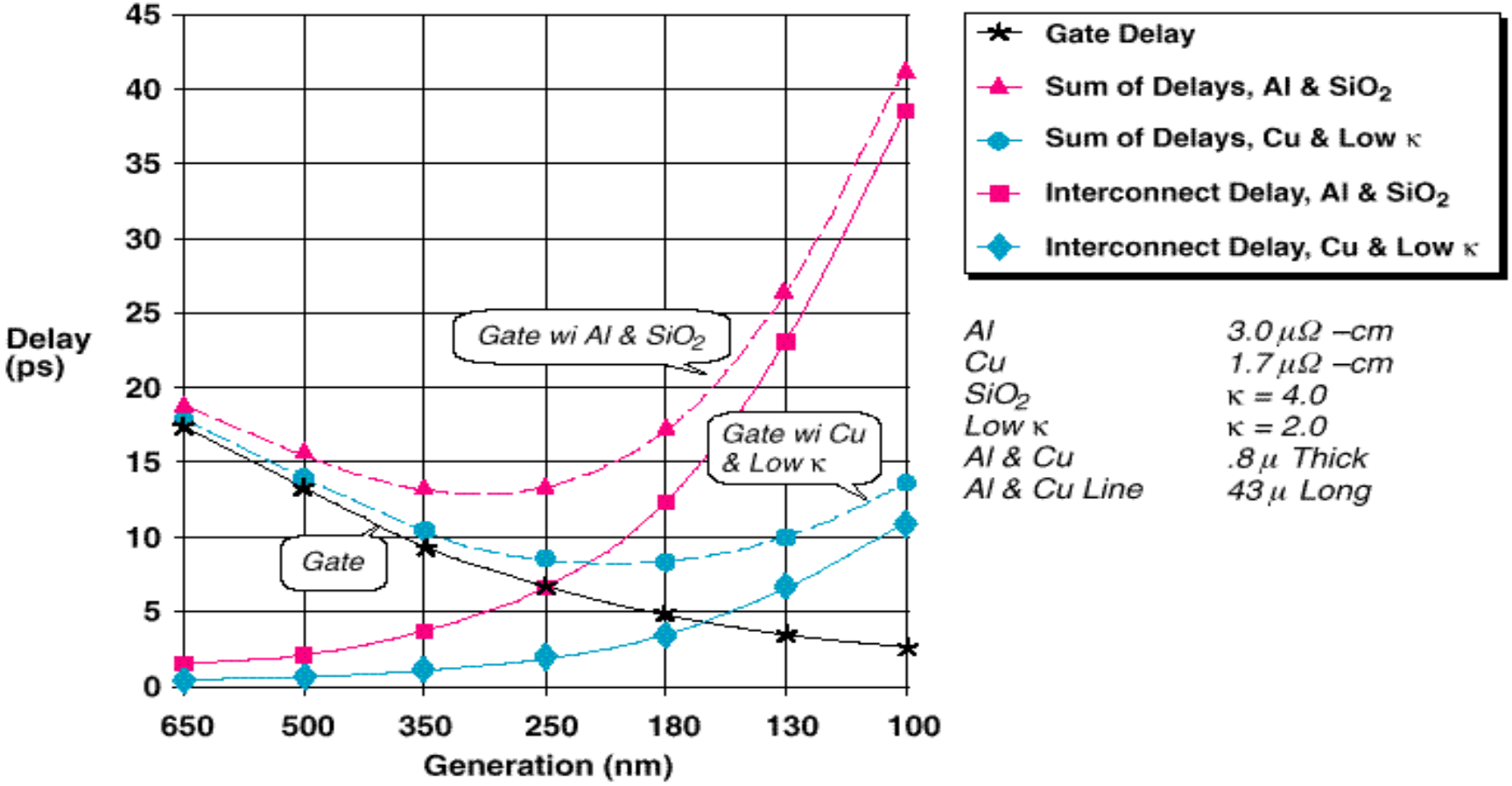
- Dense multilayer metal increases coupling capacitance



- Long/narrow line widths further increases resistance of interconnect

Effect of Advanced Interconnect

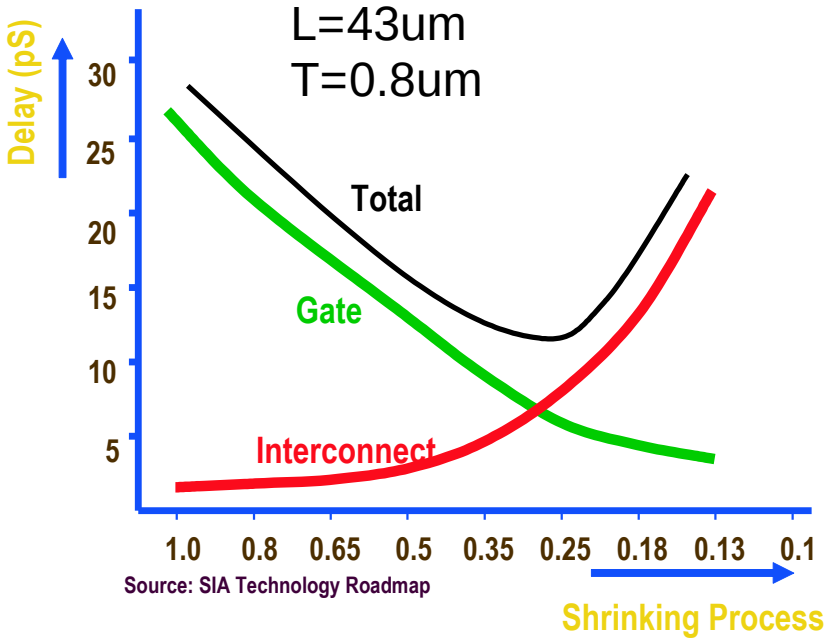
SPEED / PERFORMANCE ISSUE *The Technical Problem*

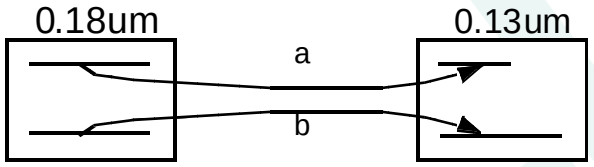


Al	*3.0 μΩ –cm*
Cu	*1.7 μΩ –cm*
SiO_2	*κ = 4.0*
Low κ	*κ = 2.0*
Al & Cu	*.8 μ Thick*
Al & Cu Line	*43 μ Long*

Effect of Wire Scaling on Delay

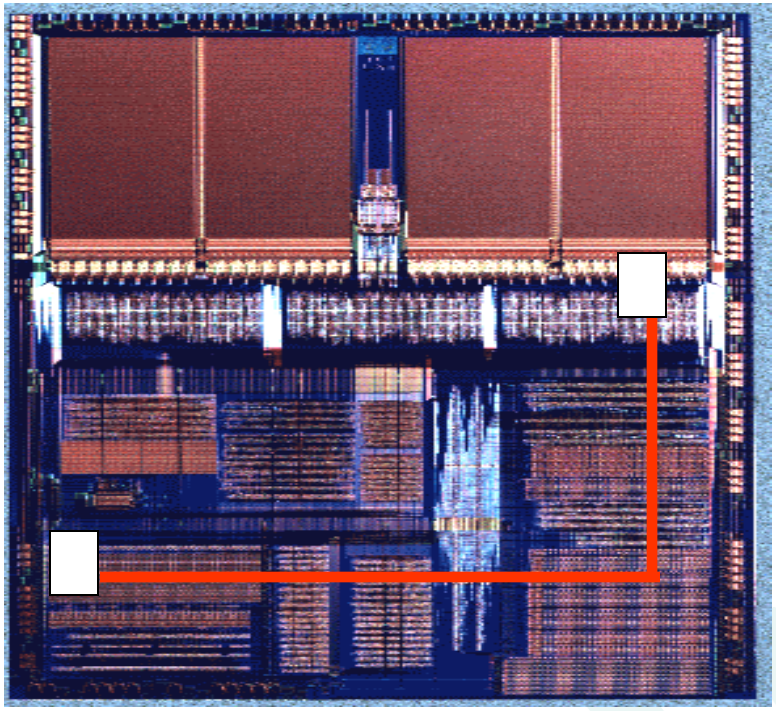
- What happens to wire delay?
- Many people claim that wire delay goes up, as shown in the famous plot from the 1995 SIA roadmap
- But it depends on how you scale the wires and which wires you are talking about.
- In a technology shrink (s< 1)
- There are really two types of wires
- a. Wires that scale L directly by s,
- b. Wires of constant percentage of die size, the global wires of the increasing complex chips
- Delay is different for these two cases as shown here:





Global Wire Delay

- Global wires
  - Non-scalable delay
  - Delay exceeds one clock cycle



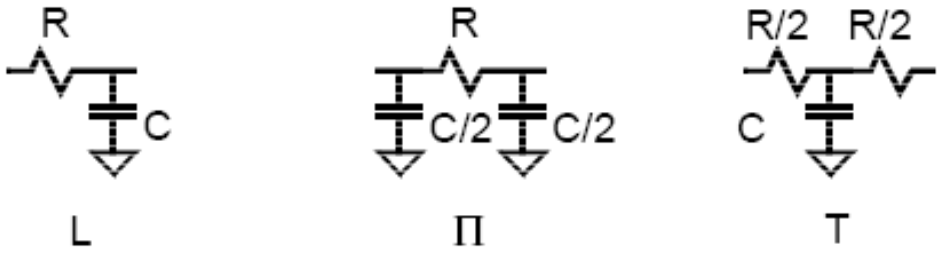
Wire Modeling

Wires are a distributed RC circuit

- Wire has r= resistance/mm and c=capacitance/mm

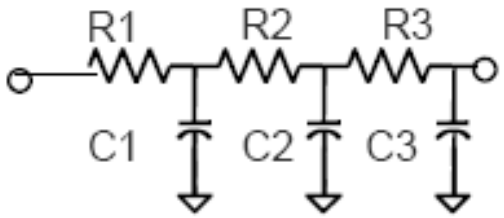
How should you model the interconnect for hand calc?

Use a simple L, Π, or T model assuming C=cL and R = rL



- One of the three models above is not useful for distributed RC modeling
- Even with proper modeling, we still have to deal with RC trees

Elmore Delay



- For a general RC ladder finding the first-order delay is easy!

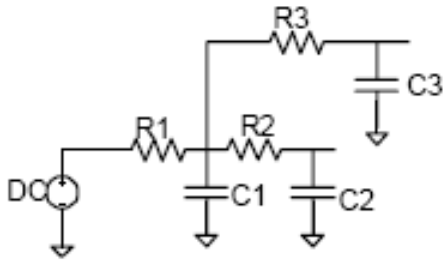
  Delay = R1 C1+ (R1+ R2) C2 + (R1+ R2+R3) C3

  i.e., Delay = $\sum$ Cap_i x (resistance from Cap_i to source)

  = ELMORE DELAY
- Notice this gives the right answer if all C's but one are zero, or all R's but one are zero
- However, in general, it is only a good delay estimate at the end of the tree

Elmore Delay

$$\tau_i = \sum R_{ik} \sum C_k$$



- C_k = every capacitance in the network in sequence
- R_{ik} = common resistance in path between source and node i and source and node k

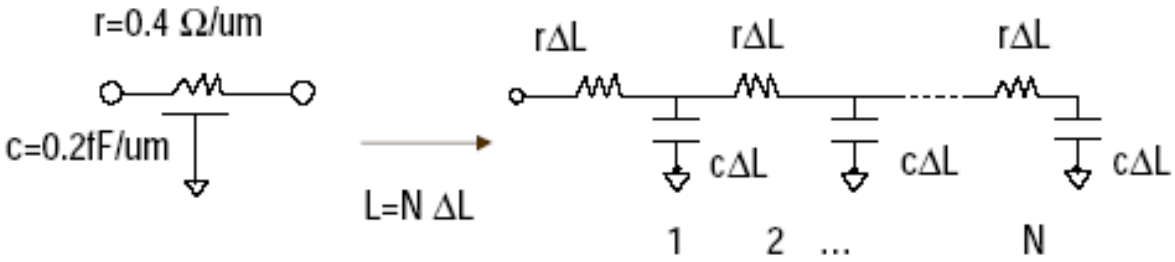
$$\tau_1 = R1C1 + R1C2 + R1C3$$

$$\tau_2 = R1C1 + (R1+R2)C2 + R1C3$$

$$\tau_3 = R1C1 + R1C2 + (R1+R3)C3$$

Delay of a wire

- What is the delay along the distributed line as a function on length L?



Use Elmore delay = $(r\,\Delta L)(c\,\Delta L) + 2(r\,\Delta L)(c\,\Delta L) + \ldots + N(r\,\Delta L)(c\,\Delta L)$

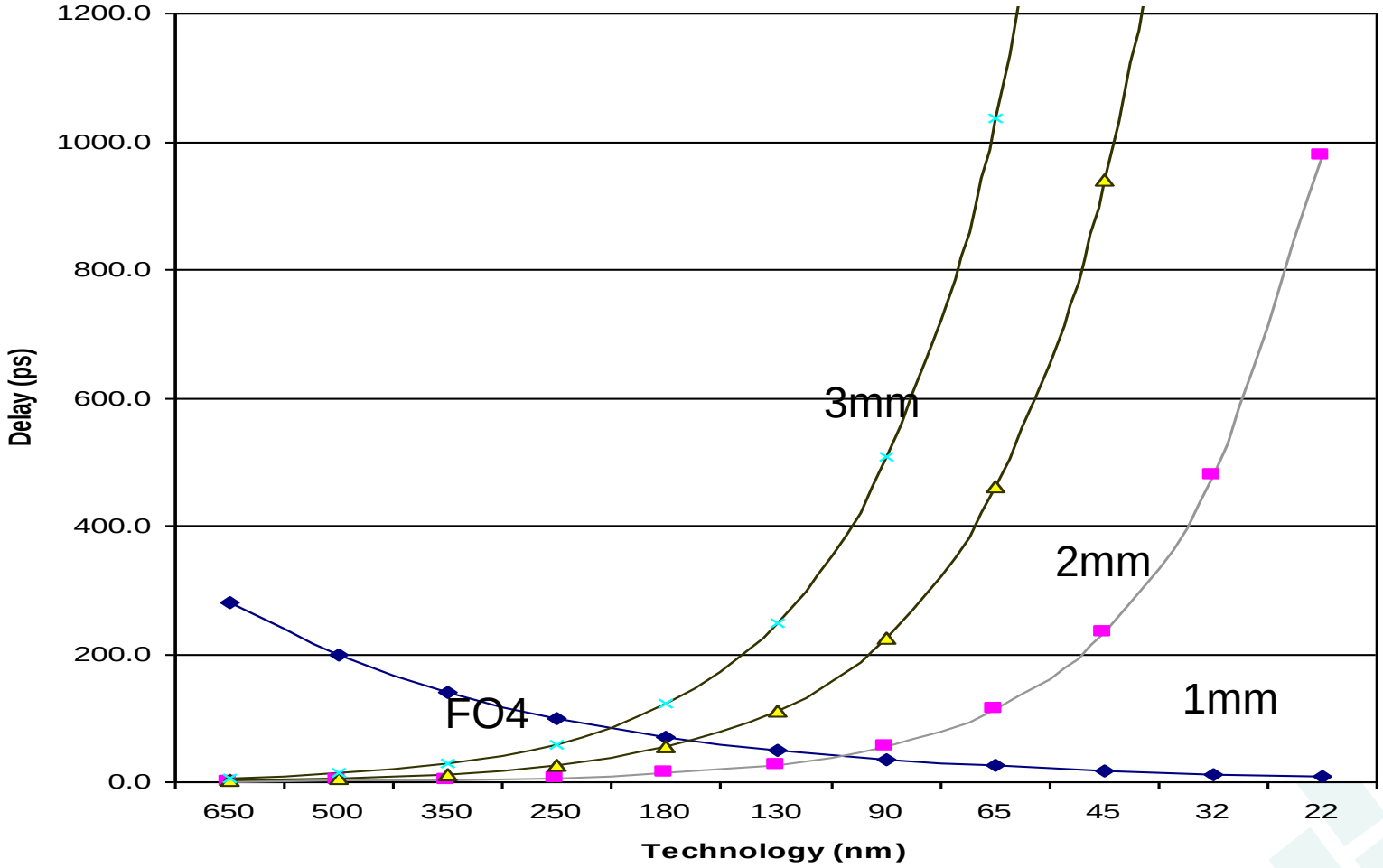
$= (\Delta L)^2 rc(1 + 2 + \ldots + N)$

$= (\Delta L)^2 rc(N)(N+1)/2 \approx (\Delta L)^2 rcN^2$

$= L^2\, rc/2 = RC/2$ (according to Elmore)

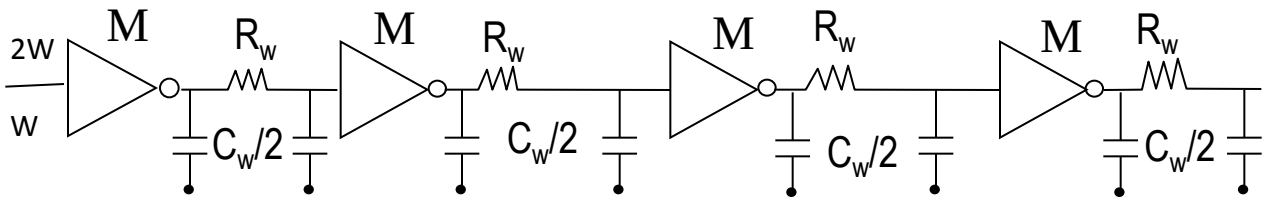
$\approx 0.4rcL^2$ (measured) Note that delay is proportional to length^2

FO4 vs. Wire Delay



Buffer Insertion for Long Wires

- Make Long wires into short wires by inserting buffers periodically. Divide interconnect into N sections as follows:



$R_{eff} = R_{eqn}/M \quad C_{self}=C_j3W*M \quad C_{fanout} = C_g3W*M \quad R_w = R_{int}L/N \quad C_w = C_{int}L/N$

- Then delay through buffers and interconnect is given by:

  $t_p = N *[R_{eff}(C_{self}+ C_W/2) + (R_{eff} + R_W)(C_W/2+C_{fanout})]$

- What is the optimal number of buffers?

  Find N such that $\partial t_p/ \partial N = 0 \Rightarrow N \approx \text{sqrt}(0.4R_{int}C_{int} L^2 /t_{pbuf})$

  where $t_{pbuf} = R_{eff}(C_{self} + C_{fanout})$

- What size should the buffers be?

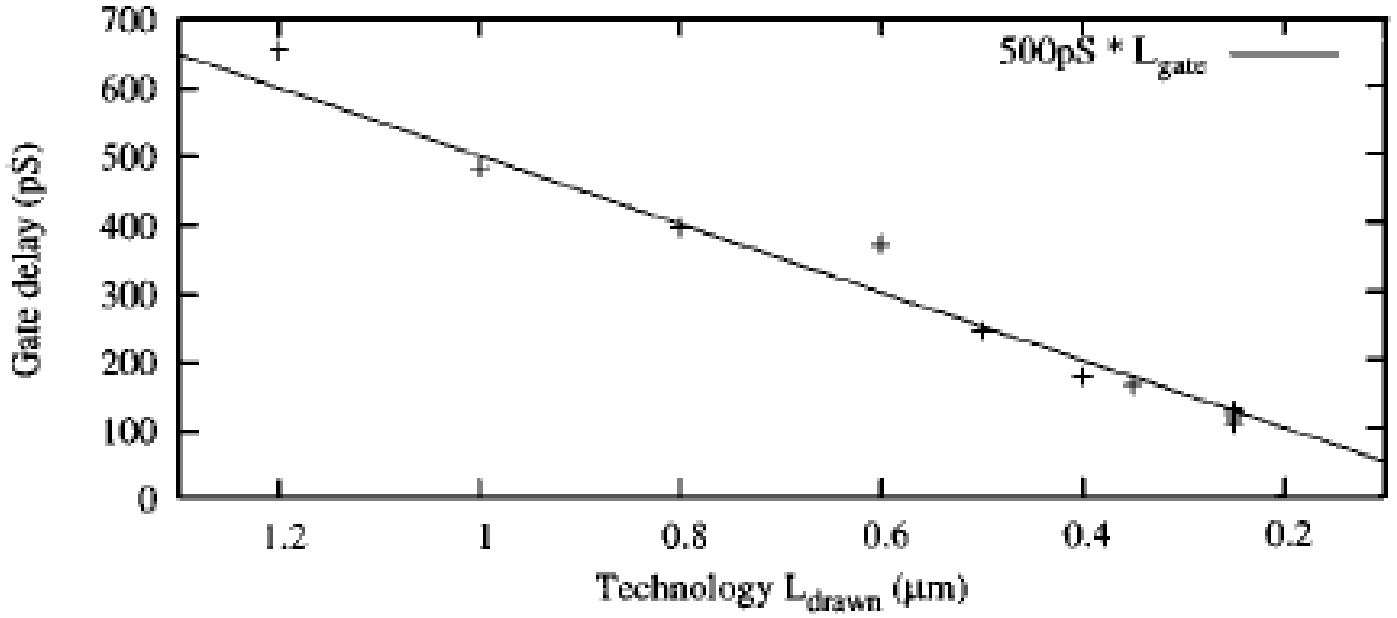
  Find M such that $\partial t_p/ \partial M = 0 \Rightarrow M = \text{sqrt}((R_{eqn}/C_g3W)(C_{int}/R_{int}))$

Issues in Buffer Insertion

- Even number of repeaters needed to avoid logic inversion
- Better strategy to optimize the delay-power product
- Repeaters for global wires require many via cuts from the upper-layer wires all the way down to the substrate
- Floorplanning
- Area and power
- Repeated wires offer increased bandwidth

Gate Delay Scaling

- Gate delay has scaled almost linearly.
- Gate and Diffusion capacitance also scale nicely



Wire Scaling

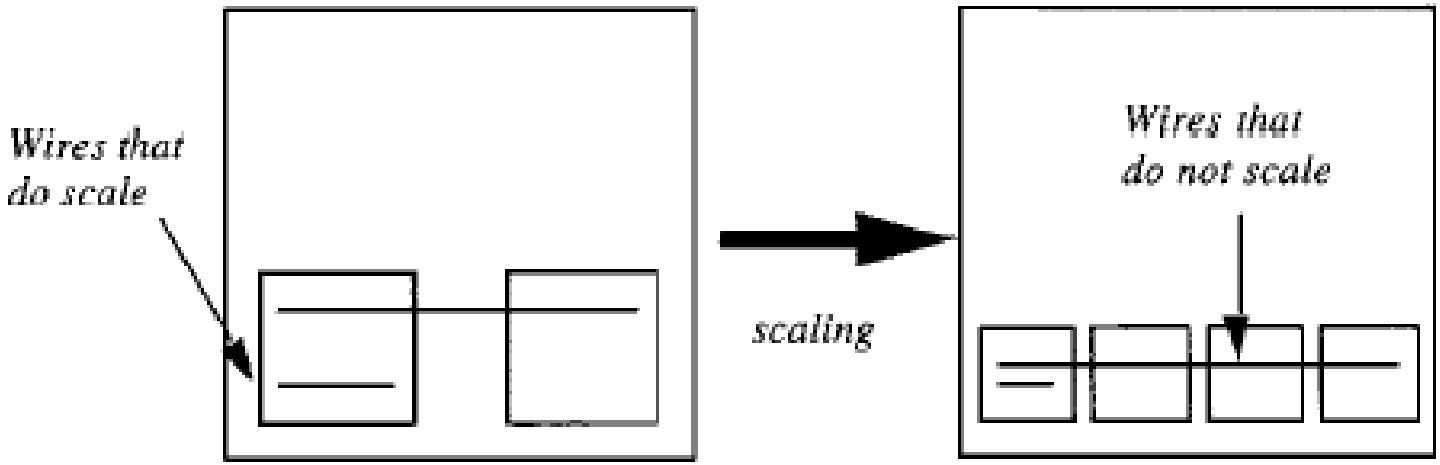
- Resistance: Resistance grows under scaling, since the width and height both scale down

L_drawn	0.18 um	0.13 um	0.10 um	0.07 um	0.05 um	0.035 um
Semi-global pitch, um	**0.36**	**0.26**	**0.20**	**0.14**	**0.10**	**0.07**
Global pitch, um	**0.72**	**0.52**	**0.40**	**0.28**	**0.20**	**0.14**
Chip edge, mm	**19**	**20.7**	**22.8**	**24.9**	**27.4**	**30.1**

Detail analysis of capacitance in later classes

Delay and Bandwidth

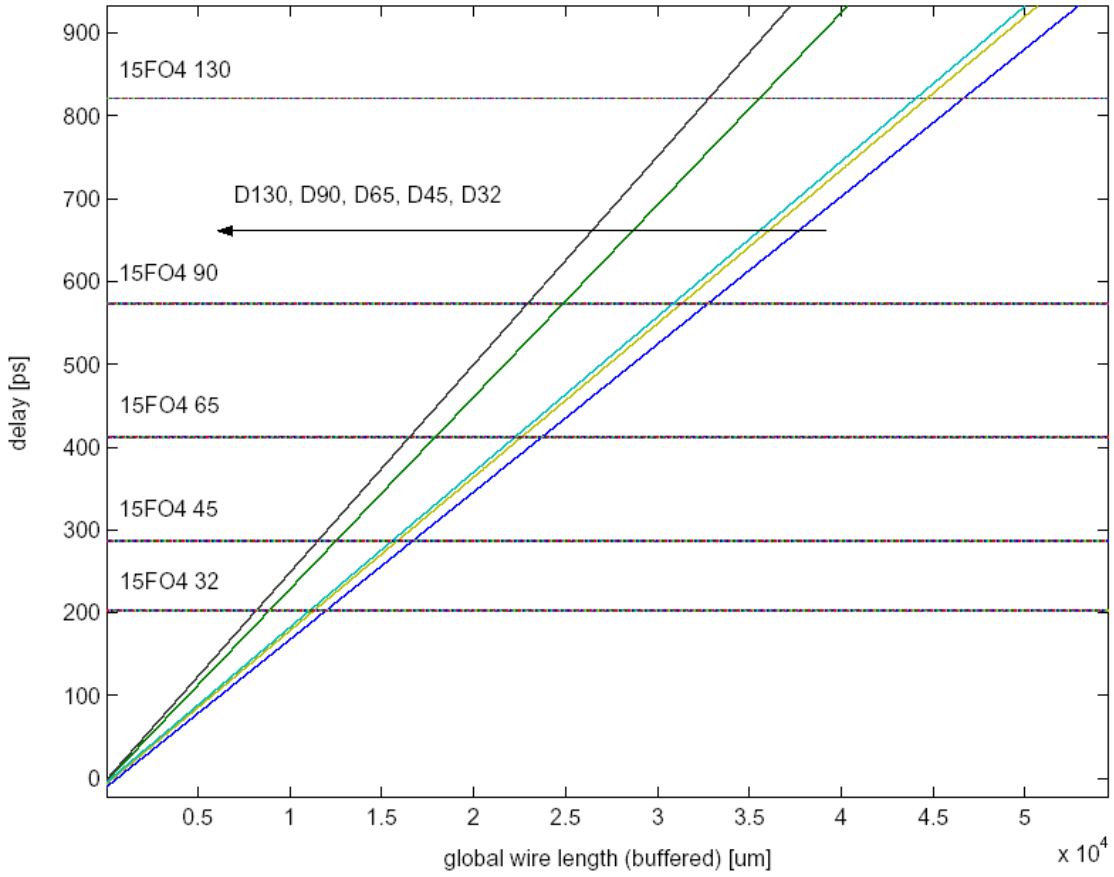
- **Classification of wires**
  - Connects gates locally within blocks, when devices and blocks get smaller, these wires get shorter
  - Connects blocks together, spanning significant portion of the die



Delay and Bandwidth (Cont'd)

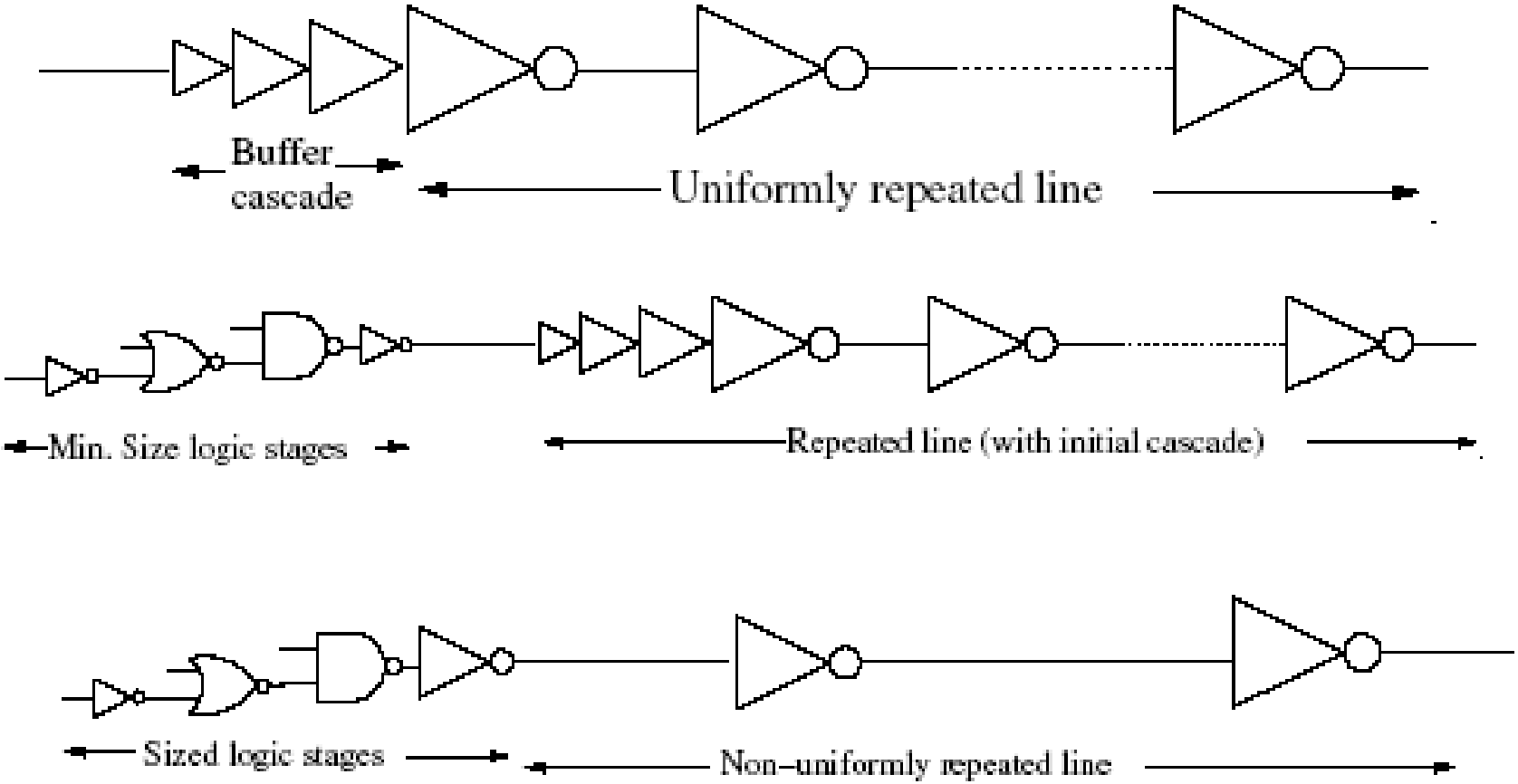
- Wires that scale in length
  - Delay scales with technology
  - Wires span block of 50k gates
- Wires that do not scale in length
  - Increasing delay disparity with gates
  - Relative to gate delay roughly doubles each generation

Global wire delay

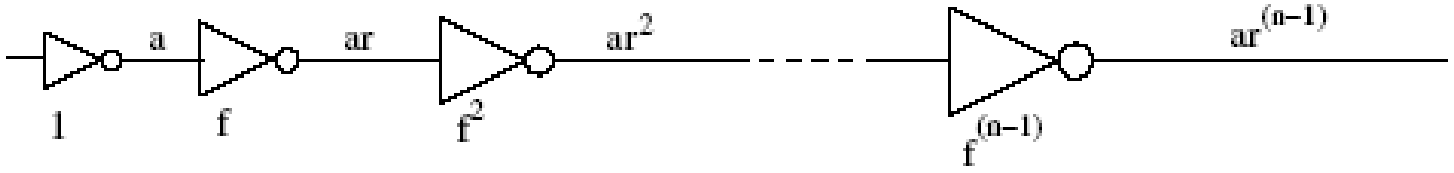


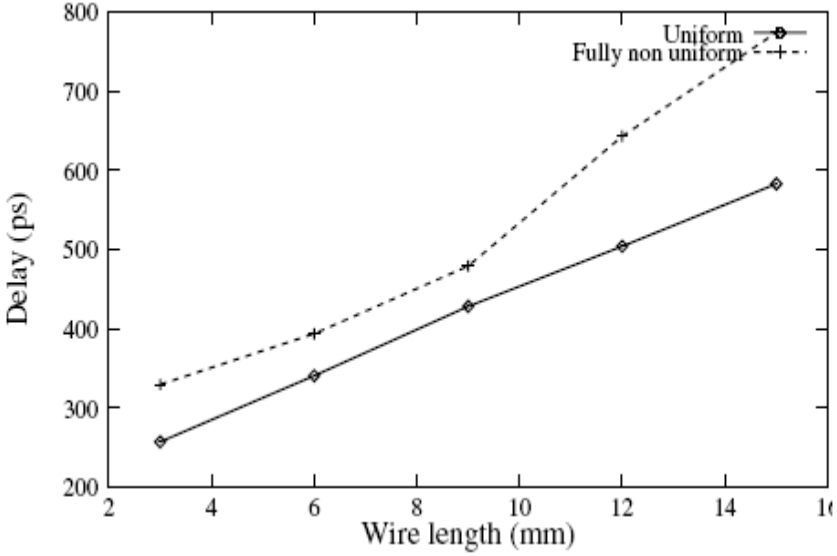
- Global wires limit the system performance

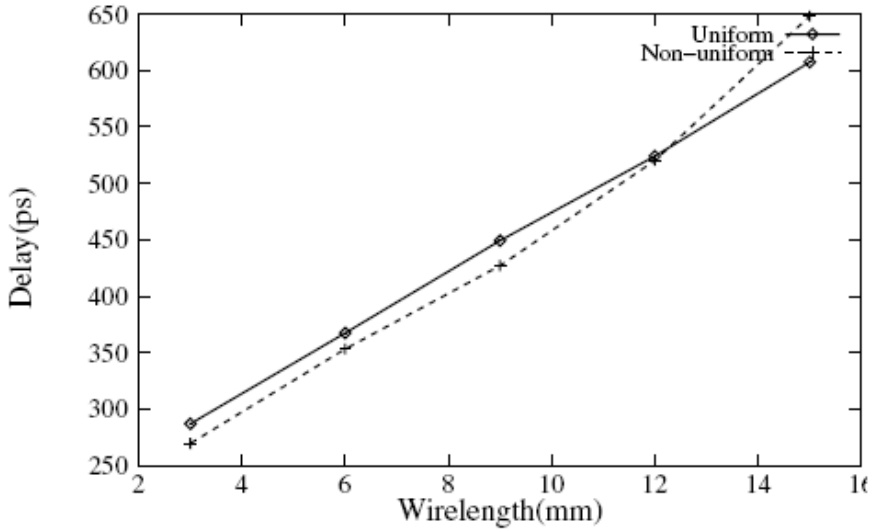
Uniformly Repeated Lines



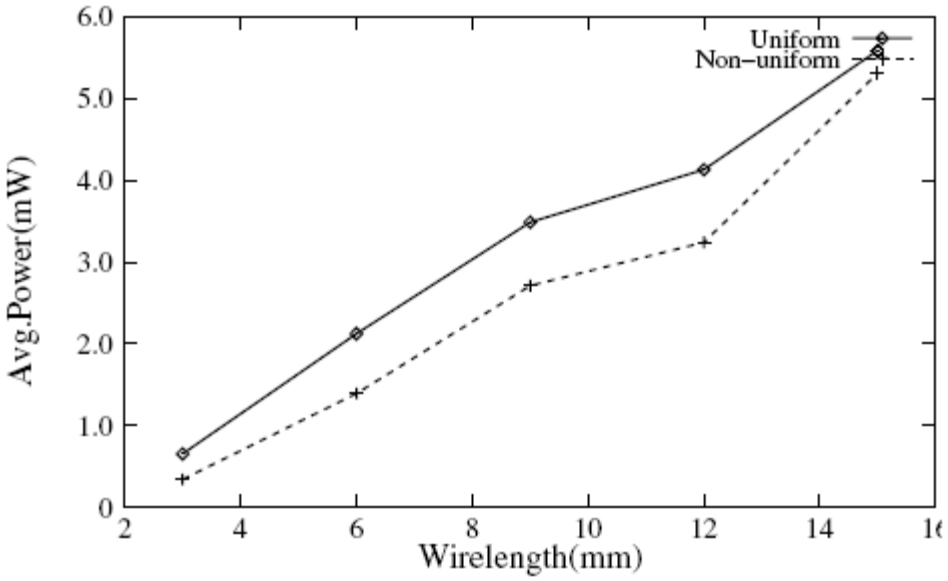
Non-uniform Buffer Insertion







Non-uniform Buffer Insertion (Cont'd)



- Gain in power consumption is due to less number of buffers

Summary

- Single synchronous clock region will span only a small fraction of the chip area
- We should not try to distribute a single low power clock all along the whole chip
- The whole SoC needs to be divided into multiple functional islands with independent frequency
- Synchronization of signals crossing multiple clock boundary is important

Discussion on Project Ideas