Lecture 2:

Wires and Wire Models

Mark Horowitz
Computer Systems Laboratory
Stanford University
horowitz@stanford.edu

Copyright © 2004 Mark Horowitz

Overview

Reading

- Aurora Capacitance extraction from layout
  - Just to see the complexity
- Horowitz Analysis of RC Trees

- Introduction
  - As technology continues to scale to smaller features and larger die, the importance of the wires increases. While smaller features mean faster devices, it does not change the performance of the wires. And if the wire length does not scale (die size is constant or increases) the performance of the wires actually gets worse. Thus a modern circuit designer spends much of her/his time managing the wires in a design.

MOS Scaling

- The velocity saturated model we have been using gives:
  - Idsat = K W $Leff^{-0.5} Tox^{-0.8} (Vgs - Vth)^{1.25}$

- If L,Tox V all scale
  - (note V scaling will be limited by Vth scaling)
  - Current should remains constant per micron
    - 0.6 to 0.8mA/m

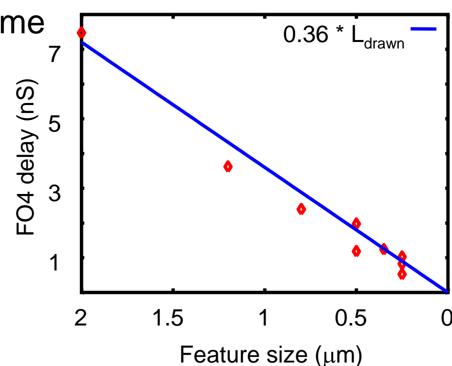
- $\Delta t = CV/i = \alpha \Delta t$ since C, V, i all scale down by α

FO4 Inverter Delay Under Scaling

- Device performance will scale
  - FO4 delay has been linear with tech

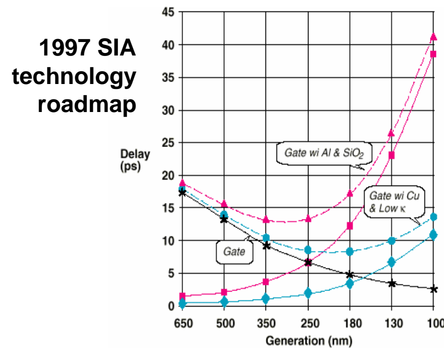
    Approximately 0.36 nS/μm*L_{drawn} at TT

    (0.5nS/μm under worst-case conditions)
- Note: Not all x μ technologies are the same
  - Intel's poly width is about ½ x
- Easy to predict gate performance
  - We can measure them
    - Labs have built 0.04μm devices
  - Key issue is voltage scaling
    - Need to scale Vdd for power
    - Hard since Vth does not scale



On-Chip Wires

- Up until the past decade, on-chip wires were invisible
  – They connected together very slow logic gates
  – Wires were "perfect" and ideal conductors
- In the past decade, gates have sped up, wires slowed down
  – Wires, no longer ideal, can throttle system performance



Source: public.itrs.net

Wire Scaling

- What happens to wire delay?
  – Many people claim that wire delay scales up, like shown in the famous plot from the 1997 SIA roadmap
  – But it depends on how you scale the wires and what wires you are talking about.
- In a new technology shrink ($\alpha < 1$)
  – There are really two types of wires
  – Wires of constant logical span, their length scales by α,
  – Wires of constant percentage of die size, the global wires of the increasing complex dies
  – The delay scaling of these wires is different

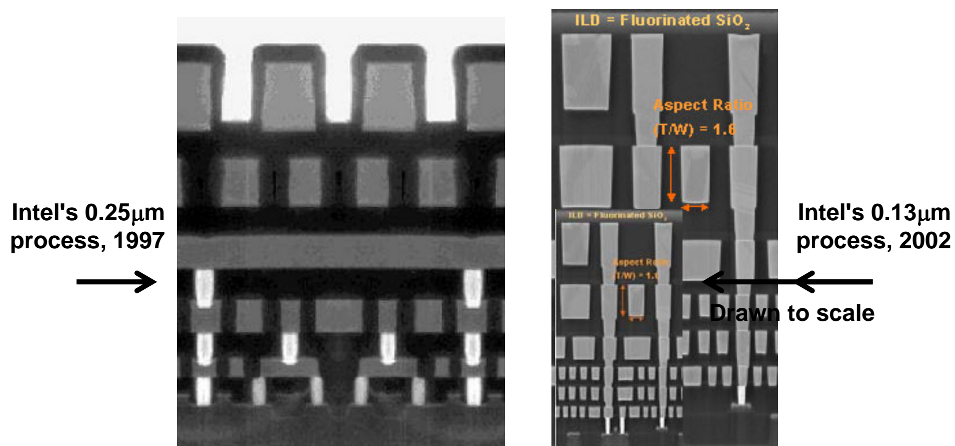
Wire Scaling, cont'd

- As technologies scale:
  - Would like to keep the wires as thick as possible
    - Keeps the resistance from growing rapidly
- Historically wire thickness has hardly scaled down
  - Still 0.8μ in 0.25μ technology
- But there is a limit to the aspect ratio of a wire
  - In 90mn technology, wire pitches are on the order of 0.3μ
  - Wire widths are only .15μ
  - Wires have had to shrink thickness to be made
- Dielectric thickness scales more slowly too
  - Tends to track metal thickness
  - Current technologies around < .5μ
- Helps a little in capacitance, needed with thick metals

Wire Scaling

As chips scale, wires have increased in number and complexity

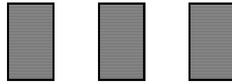
- Cross-sectional areas and spacings have decreased



Source: www.intel.com

Scaling Result

- Result is that wires are taller than they are wide
  - Farther from substrate



- Need to find Resistance and Capacitance for these type of wires
  - Where is the parallel plate?
  - Coupling is a more serious problem in scaled technologies

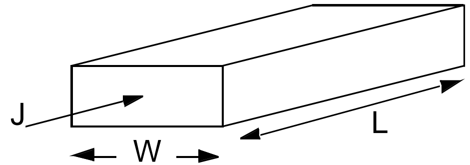
Wire Models

Wires have three important characteristics

- Resistance - relates i to V (carrier flow)
- Capacitance - relates charge (Q) to V (electric energy)
- Inductance - relates flux to i (magnetic energy)

Resistance

- Resistance is the easiest to think about since the current is confined in the material. The instrinsic resistance is pretty small, so we only need to worry about the long wires, which have a very nice shape (rectangular). For more complex shapes, the problem gets more difficult.
- $J = \sigma E$
- $i = J A$; $V = E L$
- $V/i = L/(\sigma A)$
- $R = 1/(\sigma t) * L/W$
- $R = R_{sq} * L/W$; $R_{sq} = 1/(\sigma t)$



Resistance extraction tools need to handle all the resistances even the small ones, and this leads to a number of problems. The tools need to deal with irregular shapes (think about the power and gnd nets) and end up creating a netlist with a very large number of small resistors. Using this netlist can be challenging. In addition you need to distribute the capacitance to the correct segment of the netlist with resistors.

Wire Resistance

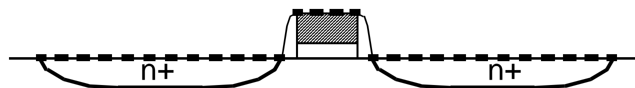
- Since the thickness of the wire is fixed by the fab,
  - Characterize wire by R_{sq}
  - Aluminum
    - $R_{sq} = 0.03\Omega\mu/t_{metal}$
  - Copper
    - $R_{sq} = 0.02\Omega\mu/t_{metal}$,
    - But t_{metal} is less than the whole thickness since you need barrier metals
- Since the thickness of metal must shrink to get finest pitches
  - Have different classes of wires with different properties
  - Metal1,2 — Finest pitch, used only in cells, thin metal
  - Metal 3..N-2 — Semi-global wires
  - Metal N-1,N — Global wires for power and signals (thick)

Wire Resistance For EE371

- Have 7 layers of metal (I think)

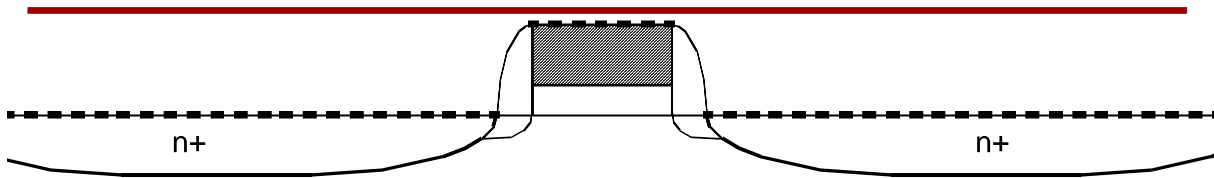
Layer	Min Pitch	Thickness	R_{sq}
M1, M2	0.30	0.30	.073
M3,4,5,6	0.40	0.45	.055
M7	1.0	1.2	.018

Silicide



- Silicide is a metal - silicon alloy
  - Formed to provide lower effective diffusion and poly R
- Thickness is not well controlled,
  - Fabs usually provide an upper bound, not lower bound
  - Self aligned to the spacers, so cover up to diffusion edge
- In many processes, building a resistor is hard
  - All resistive layers have silicide on them
  - Need to add a mask level to block the silicide

Other Resistances



- Series resistance in getting to the channel
  - Caused by the spacer
- Contacts between layers have resistance too
  - Values are very process dependent
  - Depends on the surface layers used
- Assume a couple of ohms for each contact (2Ω)
  - To get low resistance, you need many contacts
  - Need many contact for current density anyhow (electromigration)

Capacitance

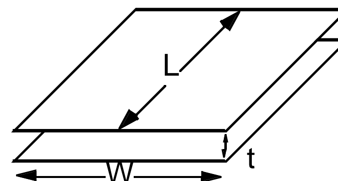
- Two simple models
  - Parallel Plate
  - Cylindrical
- The capacitance of most real objects can be approximated by a combination of these two factors.

- Parallel Plate

$$C = \frac{\varepsilon L * W}{t} \qquad \varepsilon = 0.0345 fF/\mu$$

Fixed by technology

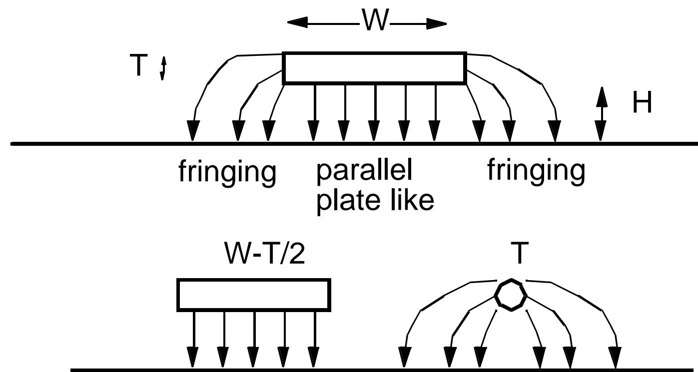
$$C = C_{per_square_micron} * W * L$$



The capacitance can be found by solving Laplace's equation. For an infinite parallel plate capacitor, the E-field does not vary in the vertical direction, and hence the voltage is proportional to the thickness.

Fringe Capacitance

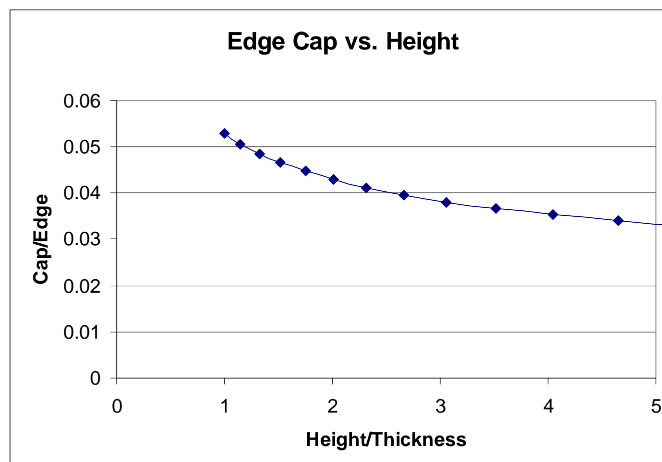
- Yuan and Trick used a simple model for fringe:



$$C_{edge} = \varepsilon L \left\{ \frac{\pi}{\ln\left(1 + \frac{2H}{T}\left(1 + \sqrt{1 + \frac{T}{H}}\right)\right)} - \frac{T}{4H} \right\}$$ (each edge)

Edge Capacitance

- For reasonable range of T/H (.3 to 1) the fringe changes from
- 0.035fF/m edge to 0.05fF/μ edge
- I use 0.05fF/μ for hand calculations (each edge)



Total Capacitance

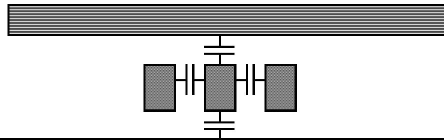
- For calculating the total capacitance
  - Add the fringe (.1fF/μ) to 4 parallel plate caps
- Seems surprising, but it does fit the data
  - And it is easy to remember

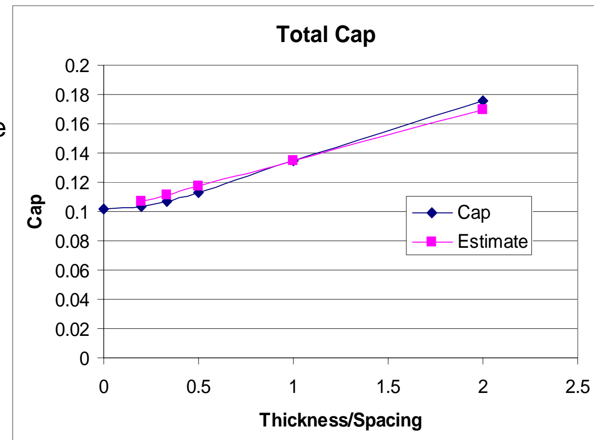
Total capacitance = Parallel Plate + 2*Cedge

The key question is what parallel plate cap?

There are wires on all four sides

Add 4 parallel plate caps!





Sidewall Capacitance

- Caution:
  - Formula is a way to estimate total capacitance
  - Does not give coupling ratio
    - Some of fringe is coupling too.

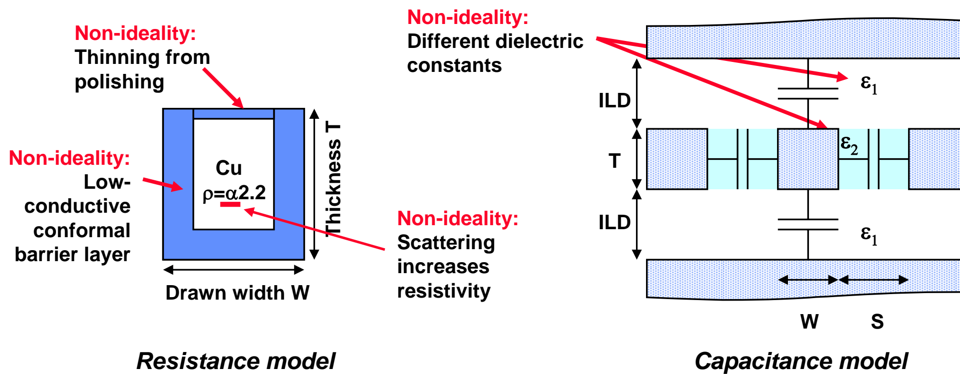
- Parallel plate caps are important
  - Especially sideway caps
  - T/S is now around 2, and can get larger
  - Large source of coupling

- To reduce this coupling, people are starting to use low dielectric material for insulators.
  - SiO_2 ε=3.9

Wire Characteristics

Relevant wire characteristics are resistance and capacitance

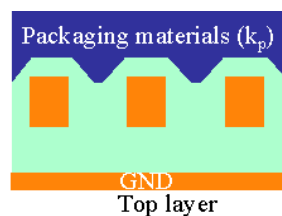
- Inductance (can be) unimportant to delay
- Use simple geometric models for R, C
  - Based on cross-sectional area
  - Account for non-idealities

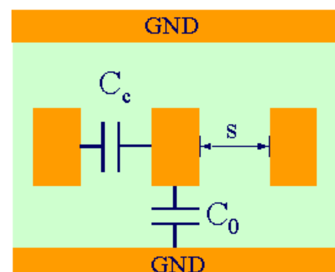


Resistance model *Capacitance model*

Real Wire Capacitance

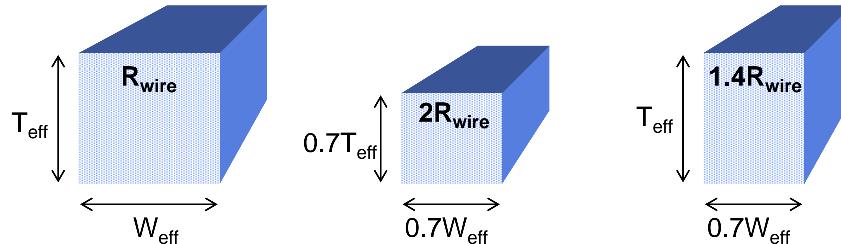
- The situation with real wires is much more complex. Currently people use field solvers to find the capacitance values and coupling for the different metal layers.
- We will use the Berkeley web for calculations you need
  - http://www-device.eecs.berkeley.edu/~ptm/interconnect.html





Wire R and C Scaling

- Resistance grows as cross-sectional area decreases
  - Strict scaling makes resistance increase quickly
  - Wires have been growing taller to compensate



- Capacitance grows slowly
  - Depends on ratio of dimensions
  - As wires stay tall (to help resistance) sideways cap grows
  - Dielectric constants get smaller (slowly) with technologies

Wire Parameters

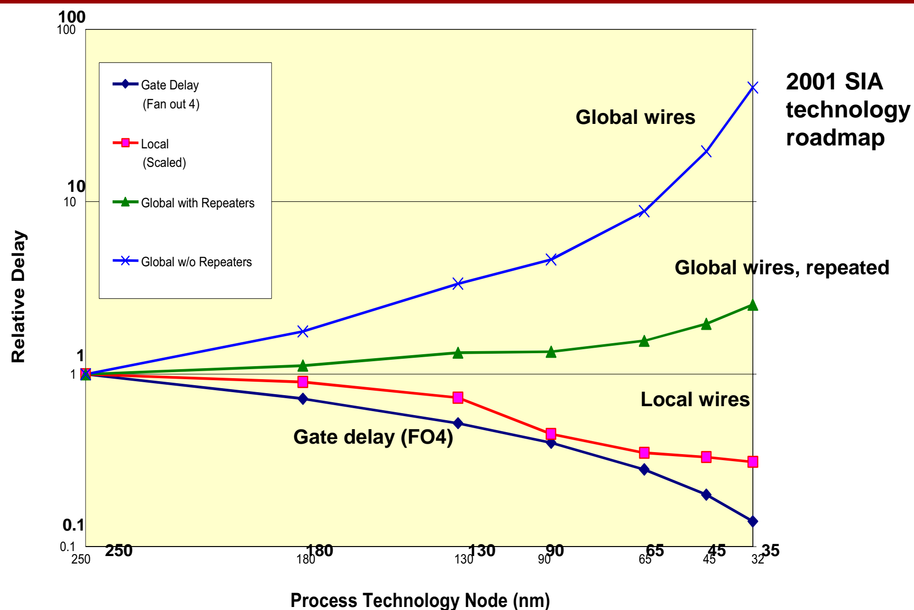
- We will assume that the
  - Intra-level dielectric constant scales = 2.8
  - Inter-level dielectric constant is constant = 3.9
- Total capacitance is now:
  - Cap/μ = 0.1fF + 0.035fF * W/H + 0.025fF * T/S *2 (two edges)
- Resistance:
  - Rsq = 0.03Ωμ/T for Al; 0.02Ωμ/T for Cu
  - Total resistance depends on whether the length scales
- Total RC
  - RC = L^2 * (3/WT + 1/TH + 1.4/WS) * 10-18s

Wire Scaling

Key is to remember to look at the two kinds of wires.

- For wires with a constant logical reach
  - All the wires when you shrink a chip to a new technology
  - All module level wires
  - L scales, so the delay of the wire decreases as T/W grows
  - Ratio of wire to gate delay grows, but slowly
- For global wires, the situation is much worse
  - The length does not scale
  - Delay increases, while the gate delays decrease
  - Should not be surprising
    - Communication has some cost
    - Scaling has finally gotten so many gates on a chip that we are starting to have communication delays on-chip! It is really a complexity issue.

ITRS Wire Delay Predictions



2001 SIA technology roadmap

Wire Limitations

- There are other reasons for wider wires:
- iR drops in power supply line
  - Need to worry about transient currents too
- Electromigration
  - Electron wind can move Al atoms
- iR heating
  - Oxide is a good thermal insulator
  - Wires are resistive and can heat up

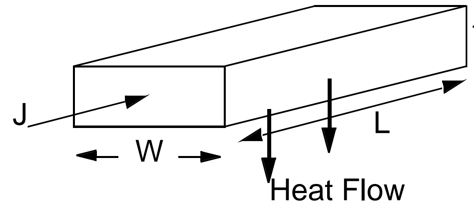
Electromigration

- Electron wind in the material can move the Al atoms
  - This is the result of a unidirectional current
  - Limitation is $1mA/m^2$ (conservative)
  - Issue mostly in supply lines
    - This current is unidirectional
- Most signal wires have AC current
  - Not an issue here
  - Watch wires that connect pMOS to nMOS
    - Wire to pMOS only pulls up
    - Wire to nMOS only pulls down
    - Can have problems in this segment
- Less of an issue in Copper wires,
  - Still a problem in copper vias

Wire Heating

- There is a limit for wires with only AC current
  - Self heating
- Would like wire not to be much hotter than chip

- Heat flow equation is like resistor:
  - Temp = Heat * $R_{thermal}$
  - $R_{thermal}$ =L/ ($\sigma_{thermal}$ A)
  - $R_{thermal}$ =Tox* 1/ (σ_{ox} LW)
- Heat generated is
  - Heat = $i^2R = i^2R_{sq}$ *L/W
  - ΔT = Heat * R = i^2R_{sq} T_{ox}* 1/ (σ_{ox} W^2)



Very simplistic model …

Wire Heating cont'd

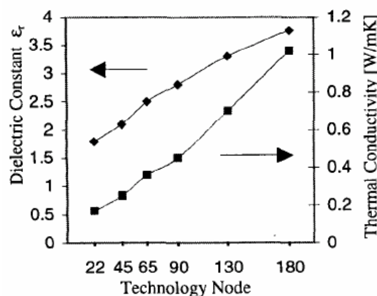


Fig. 1. Both dielectric constant and thermal conductivity of dielectrics decrease for advanced technology nodes.

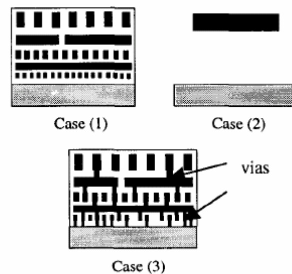


Fig. 2. Configurations of the three thermally worst case scenarios. Current flows through all wires at all metal layers in cases (1) and (3).

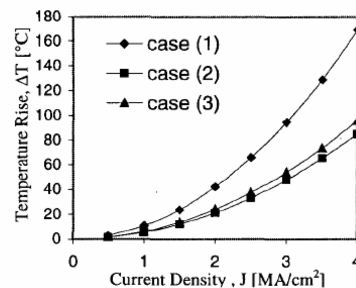
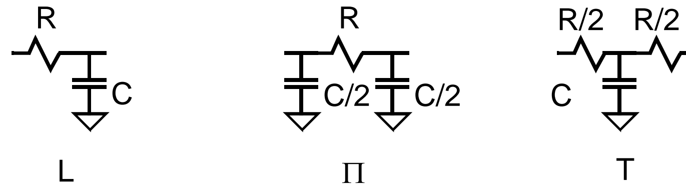


Fig. 3. Temperature rise comparison of the three thermally worst case scenarios. To simulate the more realistic condition, cases (3) is chosen over case (1) for the rest of the thermal analysis in this paper.

- From **Impact of Joule heating on scaling of deep sub-micron Cu/low-k interconnects**
  Ting-Yen Chiang; Shieh, B.; Saraswat, K.C.;
  VLSI Technology, 2002. Digest of Technical Papers. 2002 Symposium on , 11-13 June 2002

Wire Delay

- Wires are a distributed RC circuit
  - Wire has a resistance/mm capacitance/mm
- How should you model these ‘devices’
  - Break into RC sections
  - 3 options, two are ok, one is very bad



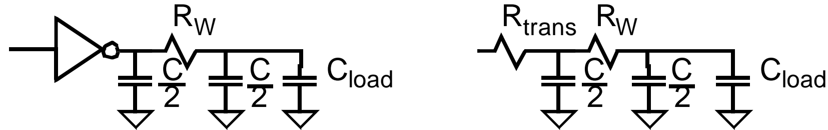
- L section is very bad!
- 64 L sections or 4P sections give 5% accuracy for a RC line

Why use “Π Model”

- A real wire is a distributed RC circuit
  - Can still calculate the Elmore delay (first moment)
  - td = ½ R_T C_T
- Elmore delay of T and Π model is ½ R_T C_T
  - Estimate is unchanged as you add sections
- Elmore delay of L model is R_T C_T for one section
  - Adding sections reduces the delay
  - (N+1)/N * ½ R_T C_T
- Since Elmore delay is a pretty good estimate, the fact that the L model gets it wrong is very bad.

Wire Delay

- Have all the tools you need to need to find the wire delay
  - But be careful about what you mean by wire delay



- Delay without the wire:
  - Delay = $R_{trans}C_{load}$
- Added delay from wire:
  - The intrinsic wire delay (½ R_W C)
  - Added delay from the wire cap (R_{trans} C)
  - Added delay from the wire resistance (RW C_{load})
  - Delay = R_{trans} (C+C_{load}) + RW(C/2+C_{load})

Wire Delay Example

- Assume you have a 5mm M3 wire 0.5μ wide, drives a 0.1pF load.
  - What is the intrinsic wire delay?
  - What size drive should you use?

- Intrinsic delay:
  - Wire capacitance depends on configuration
    - On M2, under M4?
    - Close neighbors? Neighbors switching?
  - Use 0.2fF/μ. C_W= 1pF
  - Wire resistance = 0.055 * 5K/.5 = 550Ω
  - ½ R_WC_W = 270ps

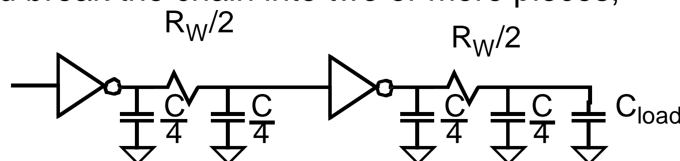
- Driver Size:
  - FO4 => input cap should be approx 1.1pF/4 = .27pF

Coupling Issues

- Wire capacitance is often not to Gnd or Vdd
  - Large fraction of wire capacitance can be coupling
    - This causes noise issues as we saw in EE313
    - Also effects timing
- What is the capacitance of a wire that is 70% coupled to neighbors
  - Depends on the data pattern
  - If neighbor wires transition in the same direction,
    - Then coupling cap is effectively 0, ($\Delta V = 0$ across coupling cap)
    - Cload = 30% of nominal load
  - If neighbor wires transition in the opposite direction,
    - Then coupling capacitance is doubled ($\Delta V = 2V$)
    - Cload = 170% of nominal load
- Change in load is effectively a factor of 5-6x

20mm Wire

- Two options on reducing the delay of the wire
  - Repeaters
  - Wider wire

- Repeaters
  - Basic problem is that the intrinsic wire delay is quadratic
  - Double the length doubles capacitance, and resistance
  - Could break the chain into two or more pieces,
  - 

  - Get linear increase
- But you get 2* TOTAL delay (gate+ wire)

Repeaters

- Can find the optimal number of repeaters
  - Add N repeaters to the line
  - Find the delay in this case
    - It will be N * delay of one section
    - (assume the final load is the same as the repeater)
- Optimize the size of the buffer to minimize the delay
  - Find this optimized delay
- Optimize N to minimize this total delay
  - This N is the optimal number of repeaters

- The optimal distance between repeaters is not very large
  - You will work it out in a homework

Wider Wires

- Resistance decreases with width
- Capacitance grows more slowly
  - Fringe does not change
  - Side wall coupling does not change
  - Only bottom and top plates change
- For our example M3 wire (which is already 2.5x min width)
  - Doubling wire width decreases resistance by 2x
  - Capacitance increases by about 30%
  - Delay drops to .65 previous delay

- Notice that moving to M7 would make the situation much better. Wires are wider and thicker, instead of just wider

Wire Design Issues

- Wire delays are now scaling differently than gates
  - Gate delays will improve with time
  - Wire delays will not
- So you want to create a design that can be shrunk to new tech
  - And you want to get the most performance possible
    - In both technologies
  - What do you do about repeater placement?
- Over design the repeaters
  - Put them in for the more advanced technology
  - Basically assume you have worse wires then you have currently, since they will eventually be that bad