Lecture 2:

Wires and Wire Models

Computer Systems Laboratory
Stanford University
horowitz@stanford.edu

Copyright © 2001 Mark Horowitz
Overview

Reading

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

This last paper is to try to give you a feeling for why the RC delay model that we use works. The model was described in 313, so I won’t hand it out here (but it will be on the web). The paper also contains a section on waveform bounds which is optional reading.

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. To do this, we need to be able to model these wires.

Let’s start by taking a quick look at scaling, and then turn to look at wire properties.
MOS Scaling

For scaling, the velocity saturated model we have been using gives:

\[ I_{dsat} = K W L_{eff}^{-0.5} T_{ox}^{-0.8} (V_{gs} - V_{th})^{1.25} \]

If \( L, T_{ox}, V \) all scale (note V scaling will be limited by \( V_{th} \) scaling)

- Current should remains constant per micron, and will be 0.6 to 0.8mA/µ
- \( \Delta t = CV/i = \alpha \Delta t \) since C, V, i all scale down by \( \alpha \)

![Fanout=4 inverter delay at TT, 90% Vdd, 125C](image-url)
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 $\alpha$.
  - 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

As technologies scale:

- Would like to keep the wires as thick as possible
  - Keeps the resistance down
  - Wire thickness has hardly scaled down; 0.8µ in 0.25µ technology
- Would like to keep dielectric thick too
  - Helps a little in capacitance
  - Needed with thick metals
- Result is that wires are taller than they are wide, and farther from substrate
  - Where is the parallel plate?
  - Coupling is a more serious problem in scaled technologies

Need to find Resistance and Capacitance for these type of wires
Wires have three important characteristics\(^1\)

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

---

1. Inductance’s effect is still small, and we will ignore it mostly. It can be important for noise, so we will discuss it a little later.
Resistance

Resistance is the easiest to think about since the current is confined in the material. The intrinsic 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.\(^1\)

\[
\mathbf{J} = \sigma \mathbf{E}
\]

\[
i = \mathbf{J} \mathbf{A}; \quad \mathbf{V} = \mathbf{E} \mathbf{L}
\]

\[
\mathbf{V}/i = \mathbf{L}/(\sigma \mathbf{A})
\]

\[
\mathbf{R} = 1/(\sigma \mathbf{t}) \times \mathbf{L}/\mathbf{W}
\]

\[
\mathbf{R} = \mathbf{R}_{sq} \times \mathbf{L}/\mathbf{W}; \quad \mathbf{R}_{sq} = 1/(\sigma \mathbf{t})
\]

---

\(^1\) 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
- $R_{sq} = 0.03 \Omega \mu m / t_{metal}$

Copper
- $R_{sq} = 0.02 \Omega \mu m / t_{metal}$
  
  But $t_{metal}$ is less than the whole thickness since you need barrier metals

For our 0.35$\mu$m technology:

<table>
<thead>
<tr>
<th>Layer</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metal4</td>
<td>0.02</td>
</tr>
<tr>
<td>Metal1,2,3</td>
<td>0.04 (in modern technologies M1 is finer and higher R)</td>
</tr>
<tr>
<td>Poly</td>
<td>~2 or 40 (if not silicided)</td>
</tr>
<tr>
<td>ndiff</td>
<td>~2 of 30 (if not silicided)</td>
</tr>
<tr>
<td>pdiff</td>
<td>~2 or 120 (if not silicided)</td>
</tr>
</tbody>
</table>
Silicide

Metal - Silicon compound that is formed to provide a lower resistance path.
- 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
In doing the drain and source engineering there is a shallower inplant under the spacer in a transistor. This improves the high-field effects, but also adds a series resistance to the source and drain. As devices scale this resistance becomes more important -- it scales up in value (gets thinner), while the resistance of the transistor remain constant.

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 (2\(\Omega\))
  - 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\(^1\)

\[
C = \frac{\varepsilon L \times W}{t} \quad \varepsilon = 0.0345\text{fF/μ}
\]

Fixed by technology

\[
C = C_{\text{per\_square\_micron}} \times W \times L
\]

---

1. 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.
Yuan and Trick\textsuperscript{1} used a simple model for fringe:

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

1. Electron Device Letters, EDL3, pps 391
Edge Capacitance

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

Total capacitance = Parallel Plate + 2*C_{edge}

The key question is what parallel plate cap?

- There are wires on all four sides
- Add 4 parallel plate caps!
Sidewall Capacitance

Some of the fringe capacitance goes to the side neighbors too, so we don’t know what the ratio of the ground to coupling capacitance is, but this approximation matches the total capacitance pretty well.

Approximate additional cap

- sideways parallel plate
- \( = 0.035 \times \frac{T}{S} \)
- seems to be a ok match to data

This additional capacitance is important, since T/S ratio might go as high as 3 in future technologies. To reduce coupling, the dielectric constant of the insulator between the wires will be reduced. This is currently SiO\(_2\), (\(\varepsilon=4\)), so the plan is to reduce \(\varepsilon\) at the same rate T/S increases to keep the coupling constant.
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. An example of this type of data is shown in the capacitance data we provide on the web.

Notice that sometimes the ‘ground plane’ is another signal, and that this geometry is not complete, since there can be wires above you too.
Wire Delay

Now we have the data for finding wire delay with scaling:

- **Resistance:**
  - \( R_{sq} = 0.03 \Omega \mu \text{m/T} \) for Al; \( 0.02 \Omega \mu \text{m/T} \) for Cu
  - Total resistance depends on whether the length scales with wire width

- **Capacitance**
  - \( \text{Cap/} \mu = 0.1 \text{fF} + 0.035 \text{fF} \times W/H + 0.035 \text{fF} \times T/S \) (*2 for two edges)
    - fringe down parallel side parallel (could have top parallel too)
    - Assumes a dielectric constant of 4

- **Total RC**
  - \( RC = L^2 \times (3/WT + 1/TH + 2/WS) \times 10^{-18} \text{s} \)
(resistance and capacitance depend only on ratios -- makes scaling cals easier)
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.
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 $1 \text{mA}/\mu^2$ (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, but still a problem in copper vias

- Current limits in copper vias are lower than they were for Tungsten plugs
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:**
  \[
  \text{Temp} = \text{Heat} \times R_{\text{thermal}}
  \]
  \[R_{\text{thermal}} = \frac{L}{\sigma_{\text{thermal}}} \quad A^1\]
  \[R_{\text{thermal}} = T_{\text{ox}} \times \frac{1}{\sigma_{\text{ox}}} \quad LW\]

Heat generated is

\[
\text{Heat} = i^2R = i^2R_{sq} \times \frac{L}{W}
\]

\[
\Delta T = \text{Heat} \times R = i^2R_{sq} \times T_{\text{ox}} \times \frac{1}{\sigma_{\text{ox}}} \times W^2
\]

---

1. This assume a rectangular heat flow. In this problem, the heat will spread out flowing form the wire to the substrate. But this is still Laplace’s equation, so you can use the cap data to find out the correction factor. Take the ratio of the true cap to parallel plate. That is reduction in resistance you will see.
Distributed RC using Elmore Delay - Review

This is an old problem, which has a nice solution for linear systems:

```
Delay = Sum (Cap_i * Resistance from Cap_i to source)
      = R1C1 + (R1 + R2) C2 + (R1 + R2 + R3) C3 +
        (R1 + R2 + R3 + R4) C4
```

Notice this gives the right answer if all C but one are zero, or all R but one are zero. It is the only equation that is correct in all the limits.

Real equation is

```
t_i = \sum_{k=1}^{n} R_{ik}C_k
```

where $t_i$ is the delay for node $i$

$R_{ik}$ is the resistance of path to ground that is common to nodes $i$ and $k$
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

  \[ \begin{align*}
  &\text{L section} \\
  &\text{Π section} \\
  &\text{T section}
  \end{align*} \]

L section is very bad!
64 L sections or 4Π 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)
  \[ t_d = \frac{1}{2} R_T C_T \]

• Elmore delay of T and Π model is \( \frac{1}{2} R_T C_T \) even for one section
  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
  \[ \frac{(N+1)}{N} \times \frac{1}{2} R_T C_T \]

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 find the wire delay

- But be careful about what you mean by wire delay

\[ \text{Delay} = R_{\text{trans}} C_{\text{load}} \]

- Delay without the wire:

- Added delay from wire:
  - The intrinsic wire delay \((1/2 \, R_W \, C)\)
  - Added delay from the wire cap \((R_{\text{trans}} \, C)\)
  - Added delay from the wire resistance \((R_W \, C_{\text{load}})\)

\[ \text{Delay} = R_{\text{trans}} (C + C_{\text{load}}) + R_W (C/2 + C_{\text{load}}) \]
Wire Delay Example

Assume you have a 10mm M2 wire 1.5µ wide that 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 M1?, close neighbors?). Use 0.2fF/µ. CW= 2pF
• Wire resistance = 0.04 * 10K/1.5 = 266Ω
• 1/2 \(R_WC_W = 266\text{ps}\)

Driver Size:
• FO4 => input cap should be approx 2.1pF/4 = .5pF

If the wire is 20mm long, Intrinsic delay is 4X
• Delay is now 1.1ns!
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 effective 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 capacitance is effectively 0, since the voltage across capacitance does not change.
    \[ C_{\text{load}} = 30\% \text{ of nominal load} \]
  - If neighbor wires transition in the opposite direction, then coupling capacitance is doubled since the voltage change across the capacitor is \( 2V_{\text{dd}} \)
    \[ C_{\text{load}} = 170\% \text{ of nominal load} \]
  - 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 pieces, and get linear increase

\[
\frac{R_W}{2} \quad \frac{R_W}{2} \\
\frac{C}{4} \quad \frac{C}{4} \quad \frac{C}{4} \quad \frac{C}{4} \quad C_{load}
\]

But you get \(2\times\) TOTAL delay (gate+ wire)
Receivers

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 number of repeaters is pretty small
• You will work it out in a homework
Wider Wires

Another method to decrease the wire delay is to make the wire wider

- Resistance decreases with width
- Capacitance grows more slowly
  
  Fringe does not change

For our example M2 wire

- Moving from 1.5 to 3µ

  Resistance 0.5

  Capacitance increases by 1.3 to 1.4 (depending on surround)

  Delay = .65 Delay

- If you increase the spacing between the wires from 1 to 3µ

  Decrease cap by 20 to 30%

  Delay = .55 Delay

Notice that moving to M4 would make the situation much better. Wires are wider and thicker