Interconnect

Prof. Krishna Saraswat

Department of Electrical Engineering
Stanford University
Stanford, CA 94305
saraswat@stanford.edu

Outline

• Interconnect scaling issues
• Gate electrode
• Aluminum technology
• Copper technology
Why should we look at interconnects?

Basics and Background
- Interconnect Parameters: resistance, capacitance, inductance
- Interconnect Metrics: Delay and Area Calculations

Scaling Related Issues: Delay
- Problems
- Solutions

Scaling Related Issues: Power

Current Interconnect Technology

Scaling of Minimum Feature Size and Chip Area

Scaling of a chip and interconnections

Scaled wires are:
• Longer (chip area scaling)
• Thinner (minimum dimension scaling)

Scaling of global interconnections
On-chip wires are getting slower

Increase in R, L and C

- Why should we look at interconnects?
- Basics and Background
  - Interconnect Parameters: resistance, capacitance, inductance
  - Interconnect Metrics: Delay and Area Calculations
- Scaling Related Issues: Delay
- Scaling Related Issues: Power
- Current Interconnect Technology
Types of Interconnects

- Dimension based
  - Local
  - Intermediate/semiglobal
  - Global
- Function based
  - Signaling
  - Clocking
  - Power/Gnd distribution

Cross Sectional View:
For Height, Width and Spacing

Top View: For Length

Performance Metrics: Signaling Wires

- Delay
- Power dissipation
- Bandwidth
- Area
- Joule heating
- Data reliability (Noise)
  - Cross talk

- Reliability
  - Electromigration

Depend on R and C and L!
Line Resistance and Capacitance

\[ R = \rho \frac{L}{WH} \]

\[ C_{ILD} = K_{ox} \varepsilon_o \frac{WL}{X_{ox}} \]

\[ C_{IMD} = K_{ox} \varepsilon_o \frac{HL}{L_S} \]

- What metrics does Resistance impact????
  - With scaling of technology L increases, X_{ox}, L_S W and H decrease
  - As a result R, C_{ox} and C_t increase

Capacitance in Multilayer Structures

In general

\[ C_{intot} = C_{ILD} + C_{IMD} = 2l(\frac{e_{ILD}}{AR} + e_{IMD}AR) \]
Capacitance: Impact on Interconnect Metrics

RC-Delay
\[ \tau \propto RC_{\text{inttot}} \]

Power
\[ P = \alpha C_{\text{inttot}} V^2 f \propto C_{\text{inttot}} \]

Crosstalk
\[ X_{\text{talk}} \propto \frac{C_{\text{IMD}}}{C_{\text{inttot}}} \left( \frac{\epsilon_{\text{ILD}}}{\epsilon_{\text{IMD}}} \right) \]

Higher Packing Density
\[ \downarrow \]
Decreased Space Between Interconnects
\[ \downarrow \]
Higher RC-Delay, power and crosstalk

Capacitance Reduction is Important for Performance Enhancement

What Capacitance to use for Delay?

Depends on switching condition on adjacent wires

- Nominal
  \[ C_{\text{inttot}} = C_{\text{IMD}} + C_{\text{ILD}} \]

- Worst Case
  \[ C_{\text{inttot}} = 2C_{\text{IMD}} + C_{\text{ILD}} \]

- Best Case
  \[ C_{\text{inttot}} = C_{\text{ILD}} \]

Not only total capacitance plays a role in delay, IMD plays a very important role.

\[ C_{\text{IMD}} \approx 70\% \text{ of } C_{\text{inttot}} \]
Interconnect Parameters at High Frequencies

- In general complicated: No dedicated return paths

- **Resistance:** $R(f)$
  - Two components: signal and return path
  - Frequency effects
    - Current distribution in signal (Skin effect)
    - Return path choice, thus resistance

- **Inductance:** $L(f)$
  - **Self:** Area enclosed between signal & return path (larger area $\Rightarrow$ larger L): Can effect delay
  - **Mutual:** Effects crosstalk (long range)

- **Capacitance:** $C(f)$
  - Relatively constant

How is current distribution in signal and its return path determined??

Current Distribution and Return Path

- **Impedance** ($Z$) is minimized
  \[ Z = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \]

- At high frequency minimizes $L$
  - Return path closer

- At lower frequency minimizes $R$
  - Return path could be far away if it lowers return resistance

$L \sim 0.3-0.7$ nH/mm

$C \sim 0.15-0.25$ pF/mm

$Z_0 = (L/C)^{1/2} \sim 35-65\Omega$
What About Skin Effect for Resistance?

- Copper

\[ \text{Skin Depth} = \frac{1}{\sqrt{2\pi f \mu_0 \sigma}} = \frac{2.1}{\sqrt{f}} \]  
  \( f \) in GHz, S.D in \( \mu \text{m} \)

- Be careful when comparing skin depth to dimensions
  - Which dimension?: tall vs. wide
  - Depends on the return path
  - Frequency of interest is not clock freq. But rise time associated freq.

- Example: \( f = 3\text{GHz} \implies \text{skin depth} = 1.2 \mu \text{m} \)

\[ \begin{array}{c}
1.2 \mu \text{m} \\
\hline
1 \mu \text{m} \\
\hline
3 \mu \text{m}
\end{array} \]

Case 1: Important (from sides)  
Case 2: Not important (from vertical direction)

Summary: R, L and C Parameters

- Can we not model anything because of complexity of R and L?

- No, there is a lot we can predict without frequency dependencies
  Using \( R_{dc} \) and C

  - Most interconnects are still not skin effect limited: \( R_{dc} \) ok
  - L is important for inductive cross talk
    - Not so important for delay
  - Delay and power important for most interconnects

- \( R_{dc} \) we need to model more accurately
The Truth about Interconnect Delay: Don’t neglect the driver!

- **Driver Delay**
  \[ Z_{dr} \left( C_{diff} + C_wL + C_L \right) \]

- **Interconnect Delay**
  - **LC regime**
    \[ \frac{l}{\nu} = \frac{L_v}{C_v} \left( \frac{L_w}{C_w} = Z_0(C/L) \right) \]
  - **RC regime**
    \[ (0.5R_w)(C_wL + C_L) \]

- Total Delay combination of three delays
- Dominant component depends on
  - wire length & cross-sectional dimensions
  - driver size: larger \( Z_{dr} \) & higher \( C_{diff} \)
- Delay RC only under certain conditions !!!


---

Delay: Local wires

- Driver small, wire length short, cross section small \( Z_{dr} > Z_0 \)
- Inductance (LC) almost never important
- Mostly driver delay dominates \( \Rightarrow \) want low \( C_w \)

\[ Delay \sim Z_{dr} \left( C_{diff} + C_wL + C_L \right) \]
Delay: Semiglobal and Global Wires

Case 1 (Small driver)
• $Z_{dr} > Z_0$ (same as local wire case)
• Inductance not important

$$Delay = Z_0 \left(\frac{1}{R} + L + C_{LC}\right) + 0.5 R C_{LC} l^2$$

Case 2a (Large driver)
• $Z_{dr} < Z_0$ and $\frac{L}{R} \cdot \frac{l}{Z_0} < 1$
• Inductance important
• Entire RLC model must be considered

Case 2b (Large driver)
• $Z_{dr} < Z_0$ and $\frac{L}{R} \cdot \frac{l}{Z_0} > 1$
• Inductance not important
• Slow diffused RC signal

• Delay has another component
  – number of round trips needed to ring up the line ($Z_d > Z_{dr}$ desirable)
• Inductance only important for case 2a
• Very long wires => RC delay conservative (diffuse RC transmission line model better)

Delay: When is inductance important? (II)

- Global wires are becoming more RC in terms of delay in future
- However $L$ is still very important in Cross talk (long range) and noise
- $L$ also important in delay for wide global wires

$$l_{crit} = \frac{2.77 L_{w}}{R_w} \sqrt{\frac{1}{C_w}} = \frac{2.77 Z_0}{R_w}$$


J. A. Davis et. al. Proc. IEEE, March 2001
Chip Size

Memory: SRAM, DRAM

• Device Size Limited
• Regular compact structure
• Needs fewer interconnect levels

Logic, e.g., μ-Processors

• Wire Pitch Limited
• Irregular structure
• Needs more interconnect levels
• Performance impacted more by interconnects

Wire-length distribution

• Wire-length distribution (in terms of gate pitches) for a futuristic logic circuit with 180 million gates.
• Metal tiers determined by $L_{\text{Loc}}$, $L_{\text{Semi-global}}$, and $L_{\text{Global}}$ boundaries defined by design constraints, such as maximum allowable delay, current density, etc.
• More wires can be accommodated in the lower levels.
• By placing wires in higher levels design constraints can be met but will need more metal levels.
Rent’s Rule

\[ T = k N^P \]

- \( T \) = # of I/O terminals
- \( N \) = # of gates
- \( k \) = avg. I/O's per gate
- \( P \) = Rent’s exponent

![Graph showing Rent's Rule](image)

Number of Gates, \( N \)
Number of I/O pins, \( T \)
Rent’s Rule fit
Intel Data
\( T = 2.09 N^{0.36} \)

Determination of Wire-length Distribution

- Conservation of I/O’s
  \[ T_A + T_B + T_C = T_{A-to-B} + T_{A-to-C} + T_{B-to-C} + T_{ABC} \]
  \[ T_{A-to-B} = T_A + T_B - T_{AB} \]
  \[ T_{B-to-C} = T_B + T_C - T_{BC} \]

- Values of \( T \) within a block or collection of blocks are calculated using Rent’s rule, e.g.,
  \[ T_A = k (N_A)^p \]
  \[ T_{ABC} = k (N_A + N_B + N_C)^p \]

- Recursive use of Rent’s rule gives wire-length distribution for the whole chip

Ref: Davis & Meindl, IEEE TED, March 1998
Why should we look at interconnects?
Basics and Background
Scaling Related Issues: Delay
- Problems
- Solutions
Scaling Related Issues: Power
Current Interconnect Technology

Future Problems (Delay)

Simple Scaling Scenarios
- **Local**: Wires whose length shrinks
  - \( S_1 \): AR maintained (3D shrink)
  - \( R \) up by \( \alpha \) (worse) where \( \alpha = \text{scaling factor} \)
  - \( C \) down by \( \alpha \) (geometrical effect)
  - \( C \) down by low-k material
  - RC delay down as low-k
  - Delay going up compare to gate delay
- **Semiglobal/Global**: Length does not shrink
  - Much worse than local

All types of signal wires delays are deteriorating wrt gate delay with scaling even with new low-k materials!
Scaling of Interconnect Cross Section Dimensions

ITRS ‘99 dimensions: local, semi-global, global wires

Solutions to Mitigate the Interconnect Problems

- **Technological Solutions**
  - Material Solutions: Lower resistivity materials and lower-dielectric constant (Existing Paradigm)
  - Future Solutions: 3-D integration and Optical Interconnects

- **Circuit Solutions**
  - Repeaters (Existing Paradigm)
  - Future Solutions: Low-swing signaling and near speed of light electrical interconnects

- **Architectural/Combination Solutions**
Impact of Interconnect Resistivity

- Will superconductors really improve the circuit speed?
- Is Cooling conventional conductors to 77°K sufficient?

Interconnect and gate delay vs chip area and minimum feature sizes for various interconnect materials

Delay calculated for the longest interconnect on a chip

\[ L = \sqrt{\frac{\text{Chip area}}{2}} \]
Limits of 4 Commonly Used Materials for Interconnections

Maximum length limited by $\tau_G = \tau_I$


Why Cu and Low-k Dielectrics?

Reduced resistivity and dielectric constant results in reduction in number of metal layers as more wires can be placed in lower levels of metal layers.
Low Dielectric Constant (Low-k) Materials

**Oxide Derivatives**
- F-doped oxides (CVD) \( k = 3.3-3.9 \)
- C-doped oxides (SOG, CVD) \( k = 2.8-3.5 \)
- H-doped oxides (SOG) \( k = 2.5-3.3 \)

**Organics**
- Polyimides (spin-on) \( k = 3.0-4.0 \)
- Aromatic polymers (spin-on) \( k = 2.6-3.2 \)
- Vapor-deposited parylene; parylene-F \( k \approx 2.7; k \approx 2.3 \)
- F-doped amorphous carbon \( k = 2.3-2.8 \)
- Teflon/PTFE (spin-on) \( k = 1.9-2.1 \)

**Highly Porous Oxides**
- Xerogels \( k = 1.8-2.5 \)

**Air** \( k = 1 \)

---

Repeater As a SOLUTION

![Repeater Diagram]

Propagation delay of a long interconnect line is

\[
\tau_L = \frac{3.56 \cdot K_{\text{ox}} \varepsilon_0 \rho}{k^2} L^2
\]

By breaking the long line into \( n \) smaller lines the delay of each line is reduced quadratically

\[
\frac{\tau_L}{n} = \frac{3.56 \cdot K_{\text{ox}} \varepsilon_0 \rho}{k^2} \left( \frac{L}{n} \right)^2
\]

The total wire delay is thus reduced significantly as \( \tau_G \) reduces with scaling

\[
\left( \frac{\tau_L}{n} + \tau_G \right) n = \frac{3.56 \cdot K_{\text{ox}} \varepsilon_0 \rho}{k^2} \left( \frac{L}{n} \right)^2 + m \tau_G
\]

However, repeaters have Power and Area penalties.
Repeaters: How Good Is It?

- Even with repeaters, 7.5X Clock at 35nm node 8X increase compared to 180nm node
- By increasing the distance between the repeaters power can be reduced at the expense of delay

Number of Repeaters Required

- ITRS wire dimensions: justified based on barely enough metal levels to fit the wires
- Separation of memory and logic area because different wire length distributions
- Rent’s rule based distribution for logic area
  - A fraction of the chip area would be occupied by repeaters
  - Additional power will be consumed by repeaters
Repeater Area Penalty

- Significant area occupied by repeaters in future
- Via blockage non-negligible for wire-limited chips

Why should we look at interconnects?
- Basics and Background
- Scaling Related Issues: Delay

Scaling Related Issues: Power
- Power Dissipation
- Power Removal (thermal Problem)
- Power Distribution
- Current Interconnect Technology
Chip Power: Breakdown

- Dynamic Power: $CV^2f$
- Leakage power: devices
- Short circuit power during switching
- Analog components (sense amps etc.): static power

Dynamic Power

- Clocking
  - Latches
  - Clocking Interconnects
- Signaling
  - Devices
  - Signaling Interconnects
- I/O
  - Buffers
  - Off-chip load

- Interconnect power
  - Due to $C_{int}$: dissipated in devices
  - Due to $R_{int}$: Joule heating (makes things worse)

Power breakdown at the 180nm node

- Clock (36%)
- Signaling Interconnects (46%)
- Local lines (29%)
- Repeaters (4%)
- Global lines (4%)
- Semi-global lines (9%)
- Latches (13%)
- Logic (Dynamic power) (15%)
- Memory (Dynamic power) (23%)
- Memory (leakage power) (4%)
Power breakdown at the 50nm node

- Logic (27%)
- Signaling Interconnects (27%)
- Memory (17%)
- Clock (28%)
- Local lines
- Repeater
- Global lines
- Distribution (Interconnects)
- Semi-global lines
- Logic (leakage power)
- Memory (leakage power)
- Latches
- Logic (Dynamic power)
- Memory (Dynamic power)

μprocessor Power Projections

ITRS projections for total power dissipation on chip
Total power dissipation on chip

- Power dissipation rising to exorbitant proportions !!!
- Need to come up with novel schemes to reduce power in each department

Thermal problem

- Higher T ⇨
  - higher R
  - lower reliability
- Better circuit design techniques needed to reduce power
- Better cooling techniques needed
The problems Caused by Increased Power

>100A will flow on these wires

**RELIABILITY**
Electromigration induced hillocks and voids

**PERFORMANCE**
As $T \uparrow$, $R \uparrow$, $RC$ delay $\uparrow$
10°C $\uparrow$, Speed $\downarrow$ 5%

Mean time to failure

$$MTF = \frac{A}{r^m J^n} \exp\left(\frac{E_a}{kT}\right)$$

10°C $\uparrow$, MTF $\downarrow$ 50%

Impact of Vias on the Thermal Characteristics of low-k Interconnects

- Vias have much higher thermal conductivity than the dielectric materials (ILD)
- Can be efficient thermal dissipation paths
Simulation of Wire Temperature: Role of Vias

- Commonly used case 1 model overestimates interconnect temperature
- Case 3 represents the most realistic worst case condition

3D Thermal Analysis of Interconnects

THERMAL-ELECTRICAL ANALOGY

<table>
<thead>
<tr>
<th>Thermal</th>
<th>Electrical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature T [K]</td>
<td>Voltage V [V]</td>
</tr>
<tr>
<td>Heat flux q [W]</td>
<td>Charge Q [C]</td>
</tr>
<tr>
<td>Thermal resistance R [K/W]</td>
<td>Electrical resistance R [V/A]</td>
</tr>
<tr>
<td>Thermal capacitance C [J/K]</td>
<td>Electrical capacitance C [C/V]</td>
</tr>
<tr>
<td>Heat diffusion</td>
<td>RC transmission line</td>
</tr>
<tr>
<td>∇²T = RC_1 \frac{dT}{dt}</td>
<td>∇²V = RC \frac{dV}{dt}</td>
</tr>
</tbody>
</table>

3-D THERMAL CIRCUIT

EFFECT OF VIAS

Polymer, ANSYS
Polymer, HSPICE
SiO₂, ANSYS
SiO₂, HSPICE

ILD: SiO₂
polymer
air

Chiang and Saraswat, VLSI Symp., June 2001
Wire Temperature vs. Low-k ILD

- Temperature of global interconnects rise sharply for low-k ILD materials.
- Embedded low-k approach, e.g., air-gap shows excellent results

Chiang, Shieh and Saraswat, VLSI Symp, June 2002

Embedded Low-k Dielectric Approach

Why embed Low-k dielectric only between metal lines?
- Capacitance is dominated by $C_{IMD}$
- Use a low-k dielectric as IMD
- Heat flows vertically
- Use a high thermal conductivity material $\Rightarrow$ SiO$_2$

Source: Y. Nishi
**Impact of Joule Heating on RC Delay**

- RC delay is strong function of current density because of Joule Heating
- Greater RC degradation for lower-k materials

**Temperature in Multilevel Metal Layers**

- With the help of vias as efficient thermal paths, the wire temperature can be significantly lower than that predicted from overly simplified 1-D thermal model.
- Therefore, the thermal problem associated with low-k insulators is not as bad as it appears.
- Beyond 45nm node closer packing of vias will alleviate the temperature rise problem.
Global Signaling Wire: Repeater Power Minimization With Delay Tradeoff

- Tolerable delay penalty depends on architecture
- Still 20W of power dissipation due to repeaters at 50nm node
- With about 20% more delay power dissipation by global wires with repeaters on them is now \( \sim 60+20=80 \text{W} \) at 50nm node

Current Interconnect Technologies

- Current Al technology (Courtesy of Motorola)
- Current Cu technology (Courtesy of IBM)
DC Resistance Modeling with Scaling: Technology Impact (I)

Diffusion barrier
- Consumes progressively larger fractional area
  - Barrier thickness (BT) doesn't scale
  - Higher AR => larger barrier area
- Technology dictates
  - Minimum thickness: reliability constraints
  - Profile: deposition technology

Electron surface scattering
- Reduced electron mobility with scaling
- Depends on
  - Ratio of $\lambda_{mfp}$ to thickness
  - Interface quality: Roughness
- Technology dictates
  - Temperature
  - Copper/barrier interface quality (P)

Cu effective $\rho$ increases in future

Problems in Scaling of Interconnections

AS $\lambda$ DECREASES
- Resistivity increases as grain size decreases
- Resistivity increases as main conductor size decreases but not the surrounding film size
• Aspect ratio increase tradeoffs:
  ▪ Better delay and electromigration
  ▪ Worse power and cross talk

• In future increasing aspect ratio may not help

• Explains why AR dropped when Al to Cu switch

• Pay attention to different metrics simultaneously rather than just delay
• Design window quite complex

Requirements of the interconnection materials

Electrical
  • Low resistivity of conductors
  • Low capacitance => low dielectric constant
    – Low RC delay
    – Low power dissipation (CV^2f loss)
    – Low cross talk
  • Low contact resistance

Processing
  • Ability to contact shallow junctions
  • Ease of deposition of thin films of the material
  • Ability to withstand the chemicals and high temperatures required in the fabrication process
  • Ability to be thermally oxidized
  • Ability to be defined into fine patterns - dry etching

Reliability
  • Resistance to electromigration
  • Good adhesion to other layers - low physical stress
  • Stability of electrical contacts to Si and other layers
  • Good MOS properties
Outline

• Interconnect scaling issues
• Silicide Gate technology
• Aluminum technology
• Copper technology