INTERCONNECTS IN VLSI DESIGN

Prof. D. Bhattacharya

Figure. Interconnect crisis in which interconnect RC becomes dominant over intrinsic gate delay in determining the overall device performance.

<table>
<thead>
<tr>
<th>TABLE 1 Interconnect Technology Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Local Line Width (μm)</td>
</tr>
<tr>
<td>Local Line Spacing (μm)</td>
</tr>
<tr>
<td>Local Line Thickness (μm)</td>
</tr>
<tr>
<td>Dielectric Thickness (μm)</td>
</tr>
<tr>
<td>Wire Width (μm)</td>
</tr>
<tr>
<td>Power Supply (V)</td>
</tr>
<tr>
<td>Frequency (MHz)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 2 Device Technology Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Local Line Width (μm)</td>
</tr>
<tr>
<td>Local Line Spacing (μm)</td>
</tr>
<tr>
<td>Local Line Thickness (μm)</td>
</tr>
<tr>
<td>Dielectric Thickness (μm)</td>
</tr>
<tr>
<td>Wire Width (μm)</td>
</tr>
<tr>
<td>Power Supply (V)</td>
</tr>
</tbody>
</table>

Integrated Circuits
Basic Components

Active Devices
(Transistors)

Until the late 1980s, the greatest challenges to processing lay in forming the active devices. As the design rule shrunk beyond 0.25 μm, the interconnects have become too important to ignore.

Local Interconnects (short)

Global Interconnects (long)

Table. Overall Technology Roadmap from National Technology Roadmap for Semiconductors (NTRS'97).

<table>
<thead>
<tr>
<th>Technology (nm)</th>
<th>250</th>
<th>180</th>
<th>150</th>
<th>130</th>
<th>100</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device intrinsic delay (ps)</td>
<td>70.5</td>
<td>51.1</td>
<td>48.7</td>
<td>45.8</td>
<td>39.2</td>
<td>21.9</td>
</tr>
<tr>
<td>1 mm (ps)</td>
<td>59</td>
<td>49</td>
<td>51</td>
<td>44</td>
<td>52</td>
<td>42</td>
</tr>
<tr>
<td>2cm un-optimized (ps)</td>
<td>2080</td>
<td>1970</td>
<td>2060</td>
<td>2070</td>
<td>2890</td>
<td>3520</td>
</tr>
<tr>
<td>2cm optimized (ps)</td>
<td>890</td>
<td>790</td>
<td>770</td>
<td>700</td>
<td>770</td>
<td>670</td>
</tr>
<tr>
<td>Projected clock period (ps)</td>
<td>1333</td>
<td>1333</td>
<td>714</td>
<td>625</td>
<td>590</td>
<td>480</td>
</tr>
</tbody>
</table>
As device dimensions are miniaturised, interconnection dimensions must also be reduced to take full advantage of the scaling process.

### Table. Scaling of MOS Transistors

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensions ( (W, L, t_{ox}, X) )</td>
<td>( 1/S )</td>
</tr>
<tr>
<td>Substrate doping ( (N_{sub}) )</td>
<td>( S )</td>
</tr>
<tr>
<td>Voltages ( (V_{DD}, V_{TN}, V_{TP}) )</td>
<td>( 1/S )</td>
</tr>
<tr>
<td>Current per device ( (I_{DS} = \frac{W}{L} \cdot t_{ox} \cdot (V_{DD} - V_{TP})^2) )</td>
<td>( 1/S )</td>
</tr>
<tr>
<td>Gate capacitance ( (C_g = \frac{W}{L} \cdot \frac{t_{ox}}{L}) )</td>
<td>( 1/S )</td>
</tr>
<tr>
<td>Transistor on-resistance ( (R_{on} = \frac{L}{t_{ox}}) )</td>
<td>1</td>
</tr>
<tr>
<td>Intrinsic gate delay ( (\tau = \frac{C_{ox} \cdot \Delta V}{I_{DS}} = R_{on} \cdot C_g) )</td>
<td>( 1/S )</td>
</tr>
<tr>
<td>Power-dissipation per gate ( (P = IV) )</td>
<td>( 1/S^3 )</td>
</tr>
<tr>
<td>Power-delay product per gate ( (P \cdot \tau) )</td>
<td>( 1/S^3 )</td>
</tr>
<tr>
<td>Area per device ( (A = W \cdot L) )</td>
<td>( 1/S^3 )</td>
</tr>
<tr>
<td>Power-dissipation density ( (P/A) )</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table. Scaling of Local and Global Interconnections

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-sectional dimensions ( (W_{oc}, H_{oc}, W_{ap}, l_{oc}) )</td>
<td>( 1/S )</td>
</tr>
<tr>
<td>Resistance per unit length ( (R_{oc} = \frac{W_{oc}}{H_{oc} \cdot t_{oc}}) )</td>
<td>( S^2 )</td>
</tr>
<tr>
<td>Capacitance per unit length ( (C_{oc} = \frac{W_{oc}}{H_{oc} \cdot t_{oc}}) )</td>
<td>1</td>
</tr>
<tr>
<td>RC constant per unit length ( (R_{oc} \cdot C_{oc}) )</td>
<td>( S^2 )</td>
</tr>
<tr>
<td>Local interconnection length ( (l_{oc}) )</td>
<td>( 1/S )</td>
</tr>
<tr>
<td>Local interconnection RC delay ( (R_{oc} \cdot C_{oc} \cdot l_{oc}) )</td>
<td>1</td>
</tr>
<tr>
<td>Die size ( (D_e) )</td>
<td>( S_e )</td>
</tr>
<tr>
<td>Global interconnection length ( (l_{ext}) )</td>
<td>( S_e )</td>
</tr>
<tr>
<td>Global interconnection RC delay ( (R_{ext} \cdot C_{ext} \cdot l_{ext}) )</td>
<td>( S^2 e^5 )</td>
</tr>
<tr>
<td>Transmission line time of flight ( (l_{oc}/v_f) )</td>
<td>( S_e )</td>
</tr>
</tbody>
</table>

\( S \): Scaling factor for device dimensions.
\( S_e \): Chip scaling factor.

The chip size increases from 5 x 5 mm\(^2\) to 10 x 10 mm\(^2\) \( \Rightarrow S_e \) is 2.

### Interconnect Scaling
- Global interconnects get longer due to larger die size
- Wire scaling increases \( R, L \) and \( C \)
- Example: local vs. global interconnect delay

### Why Al?
- Low cost, easily purified.
- Low resistivity.
- Good adherence to Si and SiO\(_2\).
- Good bondability.
- Good patternability.
- Ease of deposition.

### Problems with Al!
Device Dimensions Decreased
- Current Density Increases
- Decreased Reliability
  - (Electromigration, shorting between levels of Al)
  - Solution
    - Alternative Metal / Metal Composite
      - Tungsten, Al / Cu / Al / Mg, Al / Ti / Co
      - Al / Ta / Al, Al / Ti / Si, Al / Ni, Al / Cr, Al / Cu / Co.
**Major criteria for Interconnect design**

- **Delay**
- **Cross Talk**
- **Noise**
- **Electromigration**

---

**Interconnect Delay Problem**

- Local interconnect has sped up (shorter wires)
- Global interconnect has slowed down (RC doesn't scale)

---

The scaling properties of the delay components for a critical path that extends across the chip, to a first order, are as follows:

- \( R_{drain} = \frac{W}{L_i} \mu C_{ox} \left( V_{dd} - V_T \right) \propto \frac{1}{l} \)
- \( C_{par} = \frac{W_i L_i + W_f L_f}{l_{par}} \propto \frac{1}{l} \)
- \( C_e = \frac{W_i L_i}{l_{eox}} \propto \frac{1}{l_{eox}} \)

where \( l_{par} \) and \( l_{eox} \) are the gate and field oxide thicknesses.

---

**Use of Lumped Model**

- Line length \( l \ll l_e \), or equivalently \( l \ll l_e \)
- Rule of thumb \( l \approx \frac{l_e}{4} \)
- \( Z_{REV} > Z_{par} \)

**Gate delay**

\[ T_{gate} = \frac{Z_{REV} \left( C_e + C_{par} \right)}{Z_{REV}} \]

where

- \( Z_{REV} \) is the effective source resistance of the driving gate
- \( C_e \) is interconnection capacitance
- \( C_{par} \) is the input capacitance of the receiving gate.

---

**Significant Delay results when**

- \( C_i = C_{par} \)
- \( C_{g} = \frac{W_i L_i + W_f L_f}{l_{par}} \)

Assume a large size inverter in 2 μm CMOS technology:

- \( W_i = 10 \mu m \), \( W_f = 20 \mu m \)
- \( l_{par} = l_{eox} = 2 \mu m \)

For large length \( l_{par} \), Interconnection Capacitance: \( C_i \propto l_{par} \)

- \( C_{g} = 2 \mu F/cm \)

Gate Capacitance:

- \( C_{gg} = \frac{W_i L_i + W_f L_f}{l_{par}} \)

For \( C_{gg} \propto C_{par} \propto 0.262 \mu m \). For a fan-out of 1.

The critical interconnection length would be even smaller for a small size inverter.

With a fan-out of 10, critical interconnection length = 2.6μm.

---

Global and local interconnect delays versus gate delays.
Interconnection Capacitance Components:
1. Area component (also referred to as parallel plate capacitance component)
2. Fringing field component
3. Wire-to-wire capacitance component

To improve delay
1. Increase Dielectric Thickness
2. Reduce Wire Width
3. Reduce Spacing

Cross Talk Noise
- As wires are brought closer with scaling, capacitive coupling becomes significant
- Adjacent wires on same layer have stronger coupling

Cross Talk and Delay
- Capacitive cross talk can affect delay
- If aggressor(s) switch in opposite direction, effective coupling capacitance is doubled
- On the other hand, if aggressor(s) switch in the same direction, Cc is eliminated
- Significant difference in RC delay depending on adjacent switching activity

The interconnect capacitance $C_i$ at any node is given by

\[ C_i = C_{ij} + C_{ik} + C_{ik'} \]

- the overlap capacitance $C_{ij}$ due to the overlap between two conductors in different planes,
- the lateral capacitance $C_{ik}$ between two conductors in the same plane,
- the fringe capacitance $C_{ik'}$ that represents coupling between two conductors in different planes.
Table. Cu and AlCu wiring with \( t = 1.0 \) mm, \( Z_{RXY} = 5000 \Omega

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>( \Delta )</th>
<th>( \Delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay</td>
<td>+17</td>
<td>+17</td>
</tr>
<tr>
<td>(ns)</td>
<td>123</td>
<td>123</td>
</tr>
<tr>
<td>Fulltime</td>
<td>220</td>
<td>220</td>
</tr>
<tr>
<td>(ns)</td>
<td>191</td>
<td>191</td>
</tr>
<tr>
<td>Crosstalk</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>(mV)</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>(GHz)</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Wire Resistance

- For older technologies, we could ignore wire resistance since the number of squares of wire was small (since the wires were all short anyway).
- These days, the wire are long and thin, and the number of squares has gone up considerably.
- As the wire gets longer and thinner, \( R_{wire} \) gets bigger.

- Now interconnects have to be modeled as an RC tree or ladder.

Example of a global wire (Al line)

<table>
<thead>
<tr>
<th>Length</th>
<th>1 cm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width</td>
<td>2 ( \mu m )</td>
</tr>
<tr>
<td>Thickness</td>
<td>5 ( \mu m )</td>
</tr>
<tr>
<td>( I_{eq} )</td>
<td>1.0 ( \mu m )</td>
</tr>
<tr>
<td>Resistance</td>
<td>300 Ohm</td>
</tr>
<tr>
<td>Capacitance</td>
<td>2.0 pF</td>
</tr>
</tbody>
</table>

Distributed RC constant = 60 nsec.
50% delay = 240 nsec.

Design Requirement: \( R_{eq} = \rho / l_{eq} \),

Typical gate delay = 1 to 2 nsec.

Interconnection Resistance

Interconnection resistances is \( R_{int} = \rho / l_{eq} \),

where
\( \rho \) is the resistivity

\( W_1, H_1, \) and \( l_1 \) are the width, thickness, and the length of the interconnection.

Wire resistance increases as \( S_1 \), when ideal scaling is applied.

Interconnection Capacitance

The capacitance of the center one of three adjacent lines above a ground plane is expressed as

\[
C_{int} = \frac{W_1}{\epsilon_{\lambda}} \left( 2.8 \frac{H_1}{l_1} \right)^{3/2} \left( 0.04 \frac{W_1}{l_1} + 1.65 \frac{H_1}{l_1} - 0.14 \frac{H_1}{l_1} \right) \left( \frac{W_1}{l_1} \right)^{3/2} \frac{W_1}{l_1}
\]
Methods for improving interconnection RC Delay
(a) Multilayer Interconnection
(b) Repeaters

Additional Considerations (second order effects)
(a) Area and Yield Penalty
(b) Source/Drain and Contact Resistances
(c) Step Coverage Resistance

When to Use Transmission Line Analysis
Transmission line behavior becomes significant when the rise time $t_r$ of a signal is less than or comparable to the transmission line line time-of-flight delay $t_f$.

As a rule of thumb, transmission line phenomena become significant when

$$t_r < 2.5 t_f$$

and the line acts as a lumped capacitor when

$$t_r < 5 t_f$$

where $t_f$ = Time of flight $= \frac{L}{v}$

In between is the gray area where either transmission line analysis or lumped approximations can be used, depending on the application and required accuracy:

$$2.5 t_f < t_r < 5 t_f$$

A signal with a 0.5 nsec rise time and 300 μm as the insulator ($\varepsilon_r = 4$) yields $t_f = 3\text{ cm}$ as the interconnection length when $t_r = 2.5 t_f$. On-chip interconnections (1cm) can be treated as lumped elements until rise times reach 150 psec.

Figure. Calculated step response waveforms. The dashed curve is the waveform using the lumped-capacitance approximation.

Figure. Crosstalk amplitude versus interconnection length for two systems of free interconnections with different terminal conditions.

Figure. Layout of the bievel crossing interconnections, including the line endings for the first-level interconnections.

Figure. Dependences of the delay time and the rise time on the interlevel separation.

Figure. Dependences of the delay time and the rise time on the number of crossing interconnection lines.
Alternative Conductor Materials: Au, Ag, Cu

Problems associated with the choices
- Less compatible with the IC processing than Al.
- Poor adherence to insulator or substrate (Au, Cu).
- Incompatibility with high temperature processing (Au, Ag).
- Electromigration problem (Ag).
- High Cost (Au, Ag).
- Copper residues are toxic, need careful handling.

Likely Choice: Copper
- Electromigration resistance much higher than that of Al.
- Lifetimes are orders of magnitude higher than Al alloy.

Problems with Copper
- It diffuses very fast in both oxide and Si. Unblocked Cu atoms can reach silicon area and destroy devices by causing severe Vt drift and junction leakage.
- It cannot be readily patterned using conventional subtractive plasma etching, because of the lack of volatile halide by-products.
- It is readily oxidized in ambient air at low temperatures (<200°C) and forms no self protective layer to stop further oxidation and erosion.

Process Issues:
- Cu barrier design.
- Reliable lift off technique to obtain high aspect ratio.
- Cu Patterning.

A wire’s resistance as technologies migrated from Al to Cu:
A thin barrier layer on three sides needed to prevent copper from diffusing into surrounding oxide.
Alternative Dielectric Materials

NTRS calls for $\varepsilon_r$ values moving down to 1.5 by 2005(?) with 70 nm technology. Materials described as good candidates:

- Low k carbon doped oxide: $\varepsilon_r = 2.9$ (Interlevel dielectric)
- SiCN for interconnect layer
- Aromatic Polyethers (high thermal stability, low moisture absorption)
- Accuspin® T-24, siloxane based polymer: $\varepsilon_r = 2.5$

Multilevel System

Effected by:

- Increase of Metal Pitch (Metal cross-section and spacing)

Results in:

- Decrease of length of metal available/unit area/level

Needs:

- More metal levels to maintain same packing density
- More process steps, reduction of yield

10 levels may not be manufacturable even at relatively wide design rules >.35µ.

Interconnect Metal Layers

- Local wires have high density to accommodate the increasing number of devices
- Global wires have low RC (tall, wide, thick, sparse wires)

For a 7-layer metal (7L) process might include:

1. Local interconnect layers (e.g., M1-M3) remain at near-minimum dimensions and pitch in order to achieve routing density. For short lines (e.g., several hundred microns or less), thinner metal offers less lateral coupling capacitance and driven loads. Maximum wire width is limited by the aspect ratio upper bound.
2. Layers M2-M3 (and maybe M4) will support a mix of local and “near-global” wiring. Long wires within a single block.
3. Power distribution layers (e.g., M6-M7; maybe M5) also support the top-level clock distribution (mesh or balanced tree). Should be as thick as possible for reliability. Thick wire conserves area, but suffers from increased lateral capacitive coupling.
4. Global interconnect layers (e.g., M4-M6) support inter-block signal runs with length on the order of 3000µm - 15000µm. To satisfy delay and signal integrity constraints, at least three degrees of freedom are available: line width and spacing, repeater insertion and shield wiring.
5. All layers are subject to mutual pitch matching, via sizing, etc. considerations. Hence, available widths and spacings on one layer are not independent of the widths and spacings on a second layer.

Copper process issues:

- It diffuses very fast in both oxide and Si. If unblocked, Cu atoms can reach the silicon area and destroy devices by causing severe V, drift and junction leakage.
- It cannot be readily patterned using conventional subtractive plasma etching, because of the lack of volatile halide by-products.
- It is readily oxidized in air at low temperatures ($\sim$ 200°C) and forms no self-protective layer to stop further oxidation and erosion.

The copper integration issues:

Cross contamination,
Cu barrier design,
Trench / via etches,
Cu resistivity.
Dielectric configuration used for low-k interconnect systems.

The materials as good candidates for the ILD:

Aromatic polyethers are one of the promising candidates because they have a low dielectric constant, a high thermal stability, and low moisture absorption, and they show good adhesion to various inorganic substrates. Such as Si, SiN, Al, and TiN.

A new spin-on polymer with excellent physical and mechanical properties is described. Acetone polymer T-24 is a siloxane based polymer which has a dielectric constant of 2.5 after cure, the lowest reported this class of materials. In addition, thick, crack free films can be prepared which exhibit excellent planarity and gap-fill and are suitable for both subtractive aluminum and damascene processes.

The physical processes which cause wear-out include:

- oxide shorts;
- metallization failure due to electromigration or corrosion;
- threshold-voltage shifting effects in MOS devices;
- electrostatic discharge and electrical overstress;
- alpha particle induced soft errors.

Electromigration wear-out is one of the major interconnection failure mechanisms in VLSI integrated circuits.

Electromigration:

![Electromigration](image)

Figure. Electromigration-related failure modes. (a) A broken line caused by removal of metal molecules (b) Wrinkled structure formed by accumulation of metal molecules (c) A “kinked” line going through the insulator between the two adjoining interconnection levels causing a short circuit.

### TABLE. Dependence of the Median Time to Failure on the Wire Interconnection Line with a Length of 25 μm and Median Grain 0.75 μm.

<table>
<thead>
<tr>
<th>Width (μm)</th>
<th>Median Time to Failure (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>165</td>
</tr>
<tr>
<td>1.0</td>
<td>220</td>
</tr>
<tr>
<td>1.5</td>
<td>270</td>
</tr>
<tr>
<td>2.0</td>
<td>305</td>
</tr>
<tr>
<td>2.5</td>
<td>335</td>
</tr>
</tbody>
</table>

Reduction of Electromigration:

1. Substrate Overcoating
2. Alloying of Metallization
3. Encapsulated Multilayer Interconnections
4. Gold Metallization
5. Deposition Techniques
**Power Distribution Noise:**

IR Voltage drop along the power lines:

- Resistance of the line increases with scaling down of the minimum feature size.
- Total current supported by the lines rises with increased speed and higher integration density.
- Magnitude of the voltage drops to be tolerated is smaller, since the supply voltage is scaled down.

**Problems caused by IR voltage drops**

- The noise margin of a static circuit such as an inverter is reduced.
- Positive power supply level is reduced at some points by an amount IR in the supply path.
- A glitch in the clock signal causes latches to switch inadvertently or lose their data.
- Rise in current densities favors higher electromigration rate.

**Suggested Remedies**

- Thicker conductor
- Dedicated power planes with comb like structures
- Increased number of power planes
- More power leads between the dice and the package.

**Inductive voltage fluctuation along the power lines:**

Points to note.

- Inductive noise increases with larger amount of current switching as many circuits switch at the same time.
- These current transients generate large potential drops due to inductance of the power distribution network.
- This inductive voltage drop causes the power supply level to go down.
- The resulting voltage glitch is proportional to the switching speed, the number of drivers simultaneously switched and the effective inductance of the power lines.
- This noise is referred to as power supply level fluctuation, simultaneous switching noise, delta-I, dv/dt, and L di/dt noise.

**Simultaneous switching noise can affect the circuits in the following manners:**

- Chip to chip delays are increased, because when the power supply level goes down, the output current of all active drivers is reduced.
- The drivers on the sending chip, which are otherwise quiet, may generate noise pulses due to switching of active drivers. If the sum of this disturbance and the crosstalk is sufficient large, it can cause:
  1. False switching in the asynchronous clock lines
  2. Introduce additional delay in the synchronous lines
- It may affect the gates on the sending chip as well as active drivers. When Off-chip drivers turn on, ON-chip gate delays increase and the power supply voltage goes down.

**Reduction of noise margin with decreasing supply voltage**

[Diagram of noise margin reduction with decreasing supply voltage]
Power Supply Level Fluctuations: Methods to Minimize

- **Decoupling Capacitor**
  A decoupling capacitor reduces power supply level fluctuation by
  (i) charging up during steady state, and
  (ii) assuming the role of power supply during switching.

- **Multiple Power and Ground Pins**
  With multiple power and ground pins, each connection can supply a fraction of the total current. This reduces the effective inductance and can support a larger current spike with little power supply level fluctuation.

- **Tailoring Driver’s Turn-on Characteristics**
  If off-chip drivers are designed such that they do not turn on sharply, current transients can be reduced.

Thank You