10. Interconnects in CMOS Technology

Jacob Abraham

Department of Electrical and Computer Engineering
The University of Texas at Austin

VLSI Design
Fall 2020

September 29, 2020

Introduction to Wires on a Chip

Most of chip is wires (interconnect)

- Most of the chip is covered by wires, many layers of wires
- Transistors: little things under wires
- Wires as important as transistors
  - Affect
    - Speed
    - Power
    - Noise
- Alternating layers usually run orthogonally

Intel Damascene copper

IBM air gap between Cu
Wire Geometry

- Pitch = \( w + s \)
- Aspect Ratio, \( AR = \frac{t}{w} \)
  - Old processes had \( AR \ll 1 \)
  - Modern processes have \( AR \approx 2 \) to pack in many skinny wires

Layer Stack

- Number of metal layers has been increasing
  - AMI 0.6 mm process has 3 metal layers
  - Modern processes use 6-10+ metal layers

- Example: Intel 180 nm process
  - M1: thin, narrow (\(< 3\lambda\))
    - High density cells
  - M2-M4: thicker
    - For longer wires
  - M5-M6: thickest
    - For \( V_{DD}, \) GND, CLK

<table>
<thead>
<tr>
<th>Layer</th>
<th>( T ) [( \mu m )]</th>
<th>( W ) [( \mu m )]</th>
<th>( S ) [( \mu m )]</th>
<th>( AR )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1720</td>
<td>800</td>
<td>800</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1600</td>
<td>800</td>
<td>800</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1080</td>
<td>540</td>
<td>540</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>700</td>
<td>530</td>
<td>530</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>700</td>
<td>320</td>
<td>320</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>460</td>
<td>250</td>
<td>250</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Wire Resistance

\[ \rho = \text{resistivity} \ (\Omega \ast m) \]

\[ R = \frac{\rho \ l \ w}{l \ w} = R_\square \]

- \( R_\square \) = sheet resistance \((\Omega / \square)\)
- \( \square \) is a dimensionless unit
- Count number of squares
  - \( R = R_\square \ast (\# \ of \ squares) \)

Choice of Metals

- Until the 180 nm generation, most wires were aluminum
- Modern processes often use copper
  - Cu atoms diffuse into silicon and damage FETs
  - Must be surrounded by a diffusion barrier

<table>
<thead>
<tr>
<th>Metal</th>
<th>Bulk Resistivity ((\mu\Omega \ast cm))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silver (Ag)</td>
<td>1.6</td>
</tr>
<tr>
<td>Copper (Cu)</td>
<td>1.7</td>
</tr>
<tr>
<td>Gold (Au)</td>
<td>2.2</td>
</tr>
<tr>
<td>Aluminum (Al)</td>
<td>2.8</td>
</tr>
<tr>
<td>Tungsten (W)</td>
<td>5.3</td>
</tr>
<tr>
<td>Molybdenum (Mo)</td>
<td>5.3</td>
</tr>
</tbody>
</table>
Sheet Resistance

Typical sheet resistances in 180 nm process

<table>
<thead>
<tr>
<th>Layer</th>
<th>Sheet Resistance (Ω/□)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diffusion (silicided)</td>
<td>3–10</td>
</tr>
<tr>
<td>Diffusion (no silicide)</td>
<td>50–200</td>
</tr>
<tr>
<td>Polysilicon (silicided)</td>
<td>3–10</td>
</tr>
<tr>
<td>Polysilicon (no silicide)</td>
<td>50–400</td>
</tr>
<tr>
<td>Metal1</td>
<td>0.08</td>
</tr>
<tr>
<td>Metal2</td>
<td>0.05</td>
</tr>
<tr>
<td>Metal3</td>
<td>0.05</td>
</tr>
<tr>
<td>Metal4</td>
<td>0.03</td>
</tr>
<tr>
<td>Metal5</td>
<td>0.02</td>
</tr>
<tr>
<td>Metal6</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Contact Resistance

- Contacts and vias also have 2-20 Ω resistance
- Use many contacts for lower R
  - Many small contacts for current crowding around periphery
- Multiple contacts also help improve the yield (failure or high resistance of a contact will have only a small effect on the overall resistivity)
Wire Capacitance

- Wire has capacitance per unit length
  - To neighbors
  - To layers above and below
- \( C_{\text{total}} = C_{\text{top}} + C_{\text{bot}} + 2C_{\text{adj}} \)

Capacitance Trends

- Parallel plate equation: \( C = \epsilon A/d \)
  - Wires are not parallel plates, but obey trends
  - Increasing area (W, t) increases capacitance
  - Increasing distance (s, h) decreases capacitance
- Dielectric Constant
  - \( \epsilon = k \epsilon_0 \)
  - \( \epsilon_0 = 8.85 \times 10^{14} \text{ F/cm} \)
  - \( k = 3.9 \) for \( SiO_2 \)
  - Processes are starting to use low-k dielectrics
    - \( k \approx 3 \) (or less) as dielectrics use air pockets

Department of Electrical and Computer Engineering, The University of Texas at Austin
J. A. Abraham, September 29, 2020
For Deep Sub-Micron (DSM) (or nanoscale) processes, fringing model applies

For DSM processes, parallel plate model applies
M2 Capacitance Data

- Typical wires have $\approx 0.2 \text{ fF/} \mu \text{m}$
- Compare to $2 \text{ fF/} \mu \text{m}$ for gate capacitance

Diffusion and Polysilicon

- Diffusion capacitance is very high (about $2 \text{ fF/} \mu \text{m}$)
  - Comparable to gate capacitance
  - Diffusion also has high resistance
  - Avoid using diffusion *runners* for wires!
- Polysilicon has lower $C$ but high $R$
  - Use for transistor gates
  - Occasionally for very short wires between gates
Lumped Element Models

- Wires are a distributed system
  - Approximate with lumped element models

![Lumped Element Models Diagram]

- 3-segment \( \pi \)-model accurate to 3% in simulation
- L-model needs 100 segments for same accuracy!
- Use single segment \( \pi \)-model for Elmore delay

When to use Lumped versus Distributed Models

- First find the total R and total C for the wire.
  - If \( RC \gg t_r \) (or \( t_f \)) of driver then use distributed (II or T) model
  - If \( RC \leq t_r \) (or \( t_f \)) of driver then use lumped (L) model
- It is safe to use distributed model always, but this results in more circuit elements and larger simulation times.
- To find number of distributed elements to use
  - Increase the number of elements, and stop when the error between \( k \) and \( k + 1 \) elements is acceptably small.
- Distributed RC delay is about half that of lumped RC
- This can be validated by using the Elmore model for the distributed wire (see previous slide)
- Rule of Thumb: for a distributed wire, propagation delay can be estimated as \( \sim RC/2 \).
Example

- Metal2 wire in 180 nm process
  - 5 mm long
  - 0.32 µm wide
- Construct a 3-segment π-model
  - \( R_{\square} = 0.05 \Omega/\square \implies R = 781 \Omega \)
  - \( C_{\text{permicron}} = 0.2 fF/\mu m \implies C = 1 pF \)

Wire RC Delay

- Estimate the delay of a 10x inverter driving a 2x inverter at the end of the 5mm wire from the previous example
  - \( R = 2.5 k\Omega \ast \mu m \) for gates
  - Unit inverter: 0.36 µm nMOS, 0.72 µm pMOS

\( t_{pd} = 1.1 \text{ ns} \)
Crosstalk

- A capacitor does not like to change its voltage instantaneously
- A wire has high capacitance to its neighbor
  - When the neighbor switches from $1 \rightarrow 0$ or $0 \rightarrow 1$, the wire tends to switch too
  - Called capacitive coupling or crosstalk
- Crosstalk effects
  - **Noise** on nonswitching wires
  - Increased **delay** on switching wires

Crosstalk Delay

- Assume layers above and below on average are quiet
  - Second terminal of capacitor can be ignored
  - Model as $C_{gnd} = C_{top} + C_{bot}$
- Effective $C_{adj}$ depends on behavior of neighbors
  - **Miller Effect**

<table>
<thead>
<tr>
<th></th>
<th>$\Delta V$</th>
<th>$C_{eff}(A)$</th>
<th>MCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$V_{DD}$</td>
<td>$C_{gnd} + C_{adj}$</td>
<td>1</td>
</tr>
<tr>
<td>Switching with A</td>
<td>0</td>
<td>$C_{gnd}$</td>
<td>0</td>
</tr>
<tr>
<td>Switching opposite A</td>
<td>$2V_{DD}$</td>
<td>$C_{gnd} + 2C_{adj}$</td>
<td>2</td>
</tr>
</tbody>
</table>
Crosstalk Noise

- Crosstalk causes noise on nonswitching wires
- If victim is floating:
  - model as capacitive voltage divider

\[ \Delta V_{\text{victim}} = \frac{C_{\text{adj}}}{C_{\text{gnd-v}} + C_{\text{adj}}} \Delta V_{\text{aggressor}} \]

![Diagram of crosstalk noise](image)

Driven Victims

- Usually victim is driven by a gate that fights noise
- Noise depends on relative resistances
- Victim driver is in linear region, aggressor in saturation
- If sizes are same, \( R_{\text{aggressor}} = 2 - 4 \times R_{\text{victim}} \)

\[ \Delta V_{\text{victim}} = \frac{C_{\text{adj}}}{C_{\text{gnd-v}} + C_{\text{adj}}} \frac{1}{1 + k} \Delta V_{\text{aggressor}} \]

\[ k = \frac{\tau_{\text{aggressor}}}{\tau_{\text{victim}}} = \frac{R_{\text{aggressor}}(C_{\text{gnd-a}} + C_{\text{adj}})}{R_{\text{victim}}(C_{\text{gnd-v}} + C_{\text{adj}})} \]
Noise Implications

- So what if we have noise?
- If the noise is less than the noise margin, nothing happens
- Static CMOS logic will eventually settle to correct output even if disturbed by large noise spikes
  - But glitches cause extra delay
  - Also cause extra power from false transitions
- Dynamic logic never recovers from glitches
- Memories and other sensitive circuits also can produce the wrong answer
**Wire Engineering**

Goal: achieve delay, area, power goals with acceptable noise

- Degrees of freedom
  - Width
  - Spacing
  - Layer
  - Shielding

**Repeaters**

- R and C are proportional to $l$
- RC delay is proportional to $l^2$
  - Unacceptably great for long wires
- Break long wires into N shorter segments
  - Drive each one with an inverter or buffer
Repeater Design

- How many repeaters should we use?
- How large should each one be?
- Equivalent Circuit
  - Wire length \( l \)
  - Wire Capacitance \( C_w \cdot l \), Resistance \( R_w \cdot l \)
  - Inverter width \( W \) (nMOS = \( W \), pMOS = \( 2W \))
  - Gate Capacitance \( C' \cdot W \), Resistance \( R/W \)

Repeater Results

- Write equation for Elmore Delay
  - Differentiate with respect to \( W \) and \( N \)
  - Set equal to 0, solve

\[
\frac{l}{N} = \sqrt{\frac{2RC'}{RwC_w}}
\]

\[
\frac{td_{pd}}{l} = \left( 2 + \sqrt{2} \right) \sqrt{RC'R_wC_w}
\]

\sim 60–80 \text{ ps/mm in 0.18\( \mu \) process}

\[
W = \sqrt{\frac{RC_w}{RwC'}}
\]
Clock Distribution

High peak currents to drive typical clock loads ($\approx 1000$ pF)

\[ I_{\text{peak}} = CV \frac{dV}{dt} \]
\[ P_d = CV_{DD}^2 f \]

H-Trees

Department of Electrical and Computer Engineering, The University of Texas at Austin
J. A. Abraham, September 29, 2020
Matching Delays in Clock Distribution

- Balance delays of paths
- Match buffer and wire delays to minimize skew
- Issues
  - Load of latch (driven by clock) is data-dependent (capacitance depends on source voltage)
  - Process variations
  - IR drops and temperature variations
- Tools to support clock tree design

Clocking in the Itanium Processor

- 0.18\(\mu\) technology
- 1GHz core clock
- 200 MHz system clk
- Core clocking
  - 260 \(mm^2\)
  - 1 primary driver
  - 5 repeaters
  - 157,000 clocked latches

Source for the slides on Itanium: Intel/HP
Clock Generation

- Phase Shifted Clocks
- Ratio N
- Core clock of frequency N times system frequency
- Clock qualifiers of frequency core/N

Second Level Clock Buffer (SLCB)

Core Clock Distribution

- Repeaters (5)
- SLCBs (33)

Each SLCB ~70 tap points of ~8 gaters each
Second Level Clock Buffer (SLCB)

SLCB Schematic

- Coarse control
- Fine control
- Delay
- Clock distribution
- Debug mode
- 256ps delay range for debug purposes

First Level Route Geometry

- Lateral shielding
- 6 Layer metal process. Clocks are routed on the top two layers (m5 & m6)
- N-2 layer parallel shielding maintains constant impedance
Measured Skew

- Initial layout error
- 50ps final skew