Advanced Topics

- Packaging
- Power Distribution
- I/O

Packages

- Package functions
  - Electrical connection of signals and power from chip to board
  - Little delay or distortion
  - Mechanical connection of chip to board
  - Removes heat produced on chip
  - Protects chip from mechanical damage
  - Compatible with thermal expansion
  - Inexpensive to manufacture and test
Package Types

• Through-hole vs. surface mount

Multichip Modules

• Pentium Pro MCM
  – Fast connection of CPU to cache
  – Expensive, requires known good dice

DIP: Dual-inline package, PGA: Pin grid array, PLCC: Plastic leadless chip carrier
BGA: Ball grid array, QFP: Quad flat pack, TSOP: Thin small outline package

IBM z900 mainframe:
20 CPUs, 8 cache chips, 1km of Interconnect, 127mm on a side, 1.3 kW power
Chip-to-Package Bonding

- Traditionally, chip is surrounded by *pad frame*
  - Metal pads on 100 – 200 μm pitch
  - Gold *bond wires* attach pads to package
  - *Lead frame* distributes signals in package
  - Metal *heat spreader* helps with cooling

Advanced Packages

- Bond wires contribute parasitic inductance
- Fancy packages have many signal, power layers
  - Like tiny printed circuit boards
- *Flip-chip* places connections across surface of die rather than around periphery
  - Top level metal pads covered with solder balls
  - Chip flips upside down
  - Carefully aligned to package (done blind!)
  - Heated to melt balls
  - Also called *C4* (Controlled Collapse Chip Connection)
Heat Dissipation

- 60 W light bulb has surface area of 120 cm²
- Itanium 2 die dissipates 130 W over 4 cm²
  - Chips have enormous power densities
  - Cooling is a serious challenge
- Package spreads heat to larger surface area
  - Heat sinks may increase surface area further
  - Fans increase airflow rate over surface area
  - Liquid cooling used in extreme cases ($$$)

Example

- Your chip has a heat sink with a thermal resistance to the package of 4.0° C/W.
- The resistance from chip to package is 1° C/W.
- The system box ambient temperature may reach 55° C.
- The chip temperature must not exceed 100° C.
- What is the maximum chip power dissipation?

- \((100-55 \, \text{C}) / (4 + 1 \, \text{C/W}) = 9 \, \text{W}\)
### Power Distribution

- Power Distribution Network functions
  - Carry current from pads to transistors on chip
  - Maintain stable voltage with low noise
  - Provide average and peak power demands
  - Provide current return paths for signals
  - Avoid electromigration & self-heating wearout
  - Consume little chip area and wire
  - Easy to lay out

### Power Requirements

- \( V_{DD} = V_{DDnominal} - V_{droop} \)
- Want \( V_{droop} < +/- 10\% \) of \( V_{DD} \)
- Sources of \( V_{droop} \)
  - IR drops
  - L di/dt noise
- \( I_{DD} \) changes on many time scales
Input / Output

- Input/Output System functions
  - Communicate between chip and external world
  - Drive large capacitance off chip
  - Operate at compatible voltage levels
  - Provide adequate bandwidth
  - Limit slew rates to control di/dt noise
  - Protect chip against electrostatic discharge
  - Use small number of pins (low cost)

I/O Pad Design

- Pad types
  - $V_{DD}$ / GND
  - Output
  - Input
  - Bidirectional
  - Analog
Output Pads

- Drive large off-chip loads (2 – 50 pF)
  - With suitable rise/fall times
  - Requires chain of successively larger buffers
- Guard rings to protect against latchup
  - Noise below GND injects charge into substrate
  - Large nMOS output transistor
  - p+ inner guard ring
  - n+ outer guard ring
  - In n-well
Interconnects

• Wire Resistance
• Wire Capacitance
• Wire RC Delay
• Crosstalk
• Wire Engineering
• Repeaters

Introduction

• Chips are mostly made of wires called *interconnect*
  – In stick diagram, wires set size
  – Transistors are little things under the wires
  – Many layers of wires
• Wires are as important as transistors
  – Speed
  – Power
  – Noise
• Alternating layers run orthogonally
Wire Geometry

- Pitch = w + s
- Aspect ratio: AR = t/w
  - Old processes had AR << 1
  - Modern processes have AR ≈ 2
    - Pack in many skinny wires

Layer Stack

- AMI 0.6 μm process has 3 metal layers
- Modern processes use 6-10+ metal layers
- Example:
  - Intel 180 nm process
  - M1: thin, narrow (< 3λ)
    - 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 (nm)</th>
<th>W (nm)</th>
<th>S (nm)</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel 180 nm process</td>
<td>6</td>
<td>1720</td>
<td>800</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1600</td>
<td>800</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1880</td>
<td>540</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>700</td>
<td>330</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>700</td>
<td>330</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>490</td>
<td>250</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Wire Resistance

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

\[ R = \frac{\rho}{l} \frac{1}{w} = R_\square \frac{1}{w} \]

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

Choice of Metals

- Until 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 \text{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>

Contacts Resistance

- Contacts and vias also have 2-20 Ω
- Use many contacts for lower R
  - Many small contacts for current crowding around periphery

![Diagram showing current crowding around periphery]
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: $C = \varepsilon A/d$
  - Wires are not parallel plates, but obey trends
  - Increasing area ($W, t$) increases capacitance
  - Increasing distance ($s, h$) decreases capacitance
- Dielectric constant
  - $\varepsilon = k\varepsilon_0$
- $\varepsilon_0 = 8.85 \times 10^{-14} \text{ F/cm}$
- $k = 3.9$ for SiO$_2$
- Processes are starting to use low-k dielectrics
  - $k \approx 3$ (or less) as dielectrics use air pockets
- Typical (M2) wires have $\sim 0.2 \text{ fF/\mu m}$
  - Compare to $2 \text{ fF/\mu m}$ for gate capacitance
Diffusion & Polysilicon

- Diffusion capacitance is very high (about 2 fF/μm)
  - 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
  - 3-segment π-model is accurate to 3% in simulation
  - L-model needs 100 segments for same accuracy!
  - Use single segment π-model for Elmore delay
Example

- Metal2 wire in 180 nm process
  - 5 mm long
  - 0.32 μm wide
  - Number of squares = 5000/0.32 = 15625
- Construct a 3-segment π-model
  - $R_\square = 0.05 \Omega/\square$  \[\Rightarrow R = 15625 \times 0.05 = 781 \Omega\]
  - $C_{\text{permicron}} = 0.2 \text{ fF/μm}$  \[\Rightarrow C = 0.2 \text{ fF/μm} \times 5000 \text{ μm} = 1 \text{ pF}\]

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.
  - $R = 2.5 \text{ kΩ} \times \mu\text{m}$ for gates
  - Unit inverter: 0.36 μm nMOS, 0.72 μm pMOS
  - Unit inverter has 4λ = 0.36μm wide nMOS, 8λ = 0.72μm wide pMOS
  - Unit inverter: effective resistance of $(2.5 \text{ kΩ} \times \mu\text{m})/(0.36\mu\text{m}) = 6.9 \text{ kΩ}$
  - Capacitance: $(0.36\mu\text{m} + 0.72 \mu\text{m}) \times (2\text{ fF/μm}) = 2\text{fF}$

  \[
  \begin{array}{c}
    \text{Driver} \\
    690 \Omega
  \end{array} \quad \begin{array}{c}
    781 \Omega \\
    500 \text{ fF}
  \end{array} \quad \begin{array}{c}
    \text{Wire} \\
    500 \text{ fF}
  \end{array} \quad \begin{array}{c}
    \text{Load} \\
    4 \text{ fF}
  \end{array}
  \]

- $t_{pd} = 1.1 \text{ ns}$
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->0 or 0->1, the wire tends to switch too.
  – Called capacitive *coupling* or *crosstalk*.
• Crosstalk effects
  – Noise on non-switching 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 $C_{\text{gnd}} = C_{\text{top}} + C_{\text{bot}}$
• Effective $C_{\text{adj}}$ depends on behavior of neighbors
  – *Miller effect*

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

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

\[ \Delta V_{\text{victim}} = \frac{C_{\text{adj}}}{C_{\text{sym}} + C_{\text{adj}}} \Delta V_{\text{aggressor}} \]

Coupling Waveforms

- Simulated coupling for \( C_{\text{adj}} = C_{\text{victim}} \)
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

Diagram:

1. Driver
2. Wire Length $l$
3. N Segments
4. Repeater
5. ... (repeated N times)
6. Receiver
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

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

\(~60-80\) ps/mm

in \( 180 \) nm process