Lecture 15: Scaling & Economics
Outline

- Scaling
  - Transistors
  - Interconnect
  - Future Challenges
- Economics
Recall that Moore’s Law has been driving CMOS.
Why?

- Why more transistors per IC?
  - Smaller transistors
  - Larger dice

- Why faster computers?
  - Smaller, faster transistors
  - Better microarchitecture (more IPC)
  - Fewer gate delays per cycle
Scaling

- The only constant in VLSI is constant change
- Feature size shrinks by 30% every 2-3 years
  - Transistors become cheaper
  - Transistors become faster and lower power
  - Wires do not improve (and may get worse)
- Scale factor $S$
  - Typically $S = \sqrt{2}$
  - Technology nodes

![Feature Size vs Year Graph]
Dennard Scaling

- Proposed by Dennard in 1974
- Also known as constant field scaling
  - Electric fields remain the same as features scale
- Scaling assumptions
  - All dimensions ($x, y, z => W, L, t_{ox}$)
  - Voltage ($V_{DD}$)
  - Doping levels
## Device Scaling

### Parameter | Sensitivity | Dennard Scaling
--- | --- | ---
L: Length | 1/S | 
W: Width | 1/S | 
t\_\text{ox}: gate oxide thickness | 1/S | 
V\_\text{DD}: supply voltage | 1/S | 
V\_t: threshold voltage | 1/S | 
NA: substrate doping | S | 
\(\beta\): | \(W/(Lt_{\text{ox}})\) | S
I\_\text{on}: ON current | \(\beta(V_{\text{DD}}-V_t)^2\) | 1/S
R: effective resistance | \(V_{\text{DD}}/I_{\text{on}}\) | 1
C: gate capacitance | \(WL/t_{\text{ox}}\) | 1/S
\(\tau\): gate delay | RC | 1/S
f: clock frequency | \(1/\tau\) | S
E: switching energy / gate | \(CV_{\text{DD}}^2\) | 1/S^3
P: switching power / gate | Ef | 1/S^2
A: area per gate | WL | 1/S^2
Switching power density | P/A | 1
Switching current density | \(I_{\text{on}}/A\) | S
Observations

- Gate capacitance per micron is nearly independent of process
- But ON resistance * micron improves with process
- Gates get faster with scaling (good)
- Dynamic power goes down with scaling (good)
- Current density goes up with scaling (bad)
Example

- Gate capacitance is typically about 1 fF/μm
- The typical FO4 inverter delay for a process of feature size $f$ (in nm) is about 0.5$f$ ps
- Estimate the ON resistance of a unit $(4/2 \lambda)$ transistor.
Real Scaling

- $t_{ox}$ scaling has slowed since 65 nm
  - Limited by gate tunneling current
  - Gates are only about 4 atomic layers thick!
  - High-k dielectrics have helped continued scaling of effective oxide thickness
- $V_{DD}$ scaling has slowed since 65 nm
  - SRAM cell stability at low voltage is challenging
- Dennard scaling predicts cost, speed, power all improve
  - Below 65 nm, some designers find they must choose just two of the three
Wire Scaling

- Wire cross-section
  - w, s, t all scale
- Wire length
  - Local / scaled interconnect
  - Global interconnect
    - Die size scaled by $D_c \approx 1.1$
## Interconnect Scaling

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity</th>
<th>Scale Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>w: width</td>
<td>1/(S)</td>
<td></td>
</tr>
<tr>
<td>s: spacing</td>
<td>1/(S)</td>
<td></td>
</tr>
<tr>
<td>t: thickness</td>
<td>1/(S)</td>
<td></td>
</tr>
<tr>
<td>h: height</td>
<td>1/(S)</td>
<td></td>
</tr>
<tr>
<td>(D_c): die size</td>
<td>1/(S)</td>
<td>(D_c)</td>
</tr>
<tr>
<td>(R_w): wire resistance/unit length</td>
<td>(1/wt)</td>
<td>(S^2)</td>
</tr>
<tr>
<td>(C_{wf}): fringing capacitance / unit length</td>
<td>(t/s)</td>
<td>1</td>
</tr>
<tr>
<td>(C_{wp}): parallel plate capacitance / unit length</td>
<td>(w/h)</td>
<td>1</td>
</tr>
<tr>
<td>(C_w): total wire capacitance / unit length</td>
<td>(C_{wf} + C_{wp})</td>
<td>1</td>
</tr>
<tr>
<td>(t_{wu}): unpeated RC delay / unit length</td>
<td>(R_w C_w)</td>
<td>(S^2)</td>
</tr>
<tr>
<td>(t_w): repeated RC delay / unit length</td>
<td>(\sqrt{RCR_w C_w})</td>
<td>(\sqrt{S})</td>
</tr>
<tr>
<td>Crosstalk noise</td>
<td>(w/h)</td>
<td>1</td>
</tr>
<tr>
<td>(E_w): energy per bit / unit length</td>
<td>(C_w V_{DD}^2)</td>
<td>(1/S^2)</td>
</tr>
</tbody>
</table>
# Interconnect Delay

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity</th>
<th>Local / Semiglobal</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>l: length</td>
<td>1/S</td>
<td></td>
<td>D_c</td>
</tr>
<tr>
<td>Unrepeated wire RC delay</td>
<td>1</td>
<td></td>
<td>S^2 D_c^2</td>
</tr>
<tr>
<td>Repeated wire delay</td>
<td>sqrt(1/S)</td>
<td></td>
<td>D_c sqrt(S)</td>
</tr>
<tr>
<td>Energy per bit</td>
<td>1/S^3</td>
<td></td>
<td>D_c^c/S^2</td>
</tr>
</tbody>
</table>
Observations

- Capacitance per micron is remaining constant
  - About 0.2 fF/μm
  - Roughly 1/5 of gate capacitance
- Local wires are getting faster
  - Not quite tracking transistor improvement
  - But not a major problem
- Global wires are getting slower
  - No longer possible to cross chip in one cycle
Semiconductor Industry Association forecast
  - Intl. Technology Roadmap for Semiconductors

<table>
<thead>
<tr>
<th>Year</th>
<th>2009</th>
<th>2012</th>
<th>2015</th>
<th>2018</th>
<th>2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature size (nm)</td>
<td>34</td>
<td>24</td>
<td>17</td>
<td>12</td>
<td>8.4</td>
</tr>
<tr>
<td>$L_{gate}$ (nm)</td>
<td>20</td>
<td>14</td>
<td>10</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>$V_{DD}$ (V)</td>
<td>1.0</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.65</td>
</tr>
<tr>
<td>Billions of transistors/die</td>
<td>1.5</td>
<td>3.1</td>
<td>6.2</td>
<td>12.4</td>
<td>24.7</td>
</tr>
<tr>
<td>Wiring levels</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Maximum power (W)</td>
<td>198</td>
<td>198</td>
<td>198</td>
<td>198</td>
<td>198</td>
</tr>
<tr>
<td>DRAM capacity (Gb)</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Flash capacity (Gb)</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
</tr>
</tbody>
</table>
Scaling Implications

- Improved Performance
- Improved Cost
- Interconnect Woes
- Power Woes
- Productivity Challenges
- Physical Limits
Cost Improvement

- In 2003, $0.01 bought you 100,000 transistors
  - Moore’s Law is still going strong

Source: Dataquest/Intel

[Moore03]
Interconnect Woes

- SIA made a gloomy forecast in 1997
  - Delay would reach minimum at 250 – 180 nm, then get worse because of wires
- But…
  - Misleading scale
  - Global wires
- 100 kgate blocks ok

[SIA97]
Reachable Radius

- We can’t send a signal across a large fast chip in one cycle anymore
- But the microarchitect can plan around this
  - Just as off-chip memory latencies were tolerated
Dynamic Power

- Intel VP Patrick Gelsinger (ISSCC 2001)
  - If scaling continues at present pace, by 2005, high speed processors would have power density of nuclear reactor, by 2010, a rocket nozzle, and by 2015, surface of sun.
  - “Business as usual will not work in the future.”

- Attention to power is increasing

![Power Density Graph](image)
Static Power

- $V_{DD}$ decreases
  - Save dynamic power
  - Protect thin gate oxides and short channels
  - No point in high value because of velocity sat.
- $V_t$ must decrease to maintain device performance
- But this causes exponential increase in OFF leakage
- Major future challenge
Productivity

- Transistor count is increasing faster than designer productivity (gates / week)
  - Bigger design teams
    - Up to 500 for a high-end microprocessor
  - More expensive design cost
  - Pressure to raise productivity
    - Rely on synthesis, IP blocks
  - Need for good engineering managers
Physical Limits

- Will Moore’s Law run out of steam?
  - Can’t build transistors smaller than an atom…
- Many reasons have been predicted for end of scaling
  - Dynamic power
  - Subthreshold leakage, tunneling
  - Short channel effects
  - Fabrication costs
  - Electromigration
  - Interconnect delay
- Rumors of demise have been exaggerated
VLSI Economics

- Selling price $S_{\text{total}}$
  
  $$S_{\text{total}} = \frac{C_{\text{total}}}{1-m}$$

- $m$ = profit margin

- $C_{\text{total}}$ = total cost
  
  - Nonrecurring engineering cost (NRE)
  - Recurring cost
  - Fixed cost
NRE

- Engineering cost
  - Depends on size of design team
  - Include benefits, training, computers
  - CAD tools:
    - Digital front end: $10K
    - Analog front end: $100K
    - Digital back end: $1M

- Prototype manufacturing
  - Mask costs: $5M in 45 nm process
  - Test fixture and package tooling
Recurring Costs

- Fabrication
  - Wafer cost / (Dice per wafer * Yield)
  - Wafer cost: $500 - $3000
  - Dice per wafer: $N = \pi \left[ \frac{r^2}{A} - \frac{2r}{\sqrt{2A}} \right]$
  - Yield: $Y = e^{-AD}$
    - For small $A$, $Y \approx 1$, cost proportional to area
    - For large $A$, $Y \rightarrow 0$, cost increases exponentially

- Packaging
- Test
Fixed Costs

- Data sheets and application notes
- Marketing and advertising
- Yield analysis
Example

- You want to start a company to build a wireless communications chip. How much venture capital must you raise?

- Because you are smarter than everyone else, you can get away with a small team in just two years:
  - Seven digital designers
  - Three analog designers
  - Five support personnel
Solution

- Digital designers:
  - $70k salary
  - $30k overhead
  - $10k computer
  - $10k CAD tools
  - Total: $120k * 7 = $840k

- Analog designers
  - $100k salary
  - $30k overhead
  - $10k computer
  - $100k CAD tools
  - Total: $240k * 3 = $720k

- Support staff
  - $45k salary
  - $20k overhead
  - $5k computer
  - Total: $70k * 5 = $350k

- Fabrication
  - Back-end tools: $1M
  - Masks: $5M
  - Total: $6M / year

- Summary
  - 2 years @ $7.91M / year
  - $16M design & prototype

Digital designers:

- $70k salary
- $30k overhead
- $10k computer
- $10k CAD tools
- Total: $120k * 7 = $840k

Analog designers

- $100k salary
- $30k overhead
- $10k computer
- $100k CAD tools
- Total: $240k * 3 = $720k

Support staff

- $45k salary
- $20k overhead
- $5k computer
- Total: $70k * 5 = $350k

Fabrication

- Back-end tools: $1M
- Masks: $5M
- Total: $6M / year

Summary

- 2 years @ $7.91M / year
- $16M design & prototype