Introduction

Eniac, 1946
(first stored-program computer)
Occupied 50x30 feet room,
weighted 30 tonnes,
contained 18000 electronic valves,
consumed 25 KW of electrical power;
capable to perform 100K calc. per second

PC, 2003

PDA, 2003

Bionic, 2003

Introduction (cont’d)

Continuous growth in performance due to advances in technology and innovations in computer design
- 25-30% yearly growth in performance during 1970s
  • Mainframes and minicomputers dominated the industry
- Microprocessors enabled 35% yearly growth in performance (late 1970s)
- RISCs (Reduced Instruction Set Computers) enabled 50% yearly growth in performance (early 1980s)
  • Performance improvements through pipelining and ILP (Instruction Level Parallelism)
Effect of this Dramatic Growth

- Significant enhancement of the capability available to computer user
  - Example: your today’s PC of less than $1000 has more performance, main memory and disk storage than $1 million computer in 1980s
- Microprocessor-based computers dominate
  - Workstations and PCs have emerged as major products
  - Minicomputers - replaced by servers
  - Mainframes - replaced by multiprocessors
  - Supercomputers - replaced by large arrays of microprocessors

Changing Face of Computing

- In the 1960s mainframes roamed the planet
  - Very expensive, operators oversaw operations
  - Applications: business data processing, large scale scientific computing
- In the 1970s, minicomputers emerged
  - Less expensive, time sharing
- In the 1990s, Internet and WWW, handheld devices (PDA), high-performance consumer electronics for video games set-top boxes have emerged
- Dramatic changes have led to 3 different computing markets
  - Desktop computing, Servers, Embedded Computers
Desktop Computing

- Spans low-end ( <$1K) to high-end (≈$10K) systems
- Optimize price-performance
  - Performance measured in the number of calculations and graphic operations
  - Price is what matters to customers
- Arena where the newest highest-performance processors appear
- Market force: clock rate appears as the direct measure of performance

Embedded Computers

- Computers as parts of other devices where their presence is not obviously visible
  - E.g., home appliances, printers, smart cards, cell phones, palmtops
- Wide range of processing power and cost
  - ≈$1 (8-bit, 16-bit processors), $10 (32-bit capable to execute 50M instructions per second), ≈$100-200 (high-end video games and network switches)
- Requirements
  - Real-time performance requirement (e.g., time to process a video frame is limited)
  - Minimize memory requirements, power

Servers

- Provide more reliable file and computing services (Web servers)
- Key requirements
  - Availability – effectively provide service 24/7/365 (Yahoo!, Google, eBay)
  - Reliability – never fails
  - Scalability – server systems grow over time, so the ability to scale up the computing capacity is crucial
  - Performance – transactions per minute

Computing Classes: A Summary

<table>
<thead>
<tr>
<th>Feature</th>
<th>Desktop</th>
<th>Server</th>
<th>Embedded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price of the system</td>
<td>$1K-$10K</td>
<td>$10K-$10M</td>
<td>$100-$100K</td>
</tr>
<tr>
<td>Price of the processor</td>
<td>$100-$1K</td>
<td>$200-$2K</td>
<td>$2-$200</td>
</tr>
<tr>
<td>Sold per year (from 2000)</td>
<td>300M</td>
<td>4M</td>
<td>300M (only 32-bit and 64-bit)</td>
</tr>
<tr>
<td>Critical system design issues</td>
<td>Price, performance, availability, graphics performance</td>
<td>Throughput, performance, availability, scalability performance</td>
<td>Price, power consumption, application-specific performance</td>
</tr>
</tbody>
</table>
Task of Computer Designer

- "Determine what attributes are important for a new machine; then design a machine to maximize performance while staying within cost constraints.”

- Aspects of this task
  - instruction set design
  - functional organization
  - logic design and implementation (IC design, packaging, power, cooling...)

Technology Trends

- Integrated circuit technology – 55% /year
  - Transistor density – 35% per year
  - Die size – 10-20% per year

- Semiconductor DRAM
  - Density – 40-60% per year (4x in 3-4 years)
  - Cycle time – 33% in 10 years
  - Bandwidth – 66% in 10 years

- Magnetic disk technology
  - Density – 100% per year
  - Access time – 33% in 10 years

- Network technology (depends on switches and transmission technology)
  - 10Mb/100Mb (10 years), 100Mb/1Gb (5 years)
  - Bandwidth – doubles every year (for USA)

What is Computer Architecture?

Computer Architecture covers all three aspects of computer design

- Instruction Set Architecture
  - the computer visible to the assembler language programmer or compiler writer (registers, data types, instruction set, instruction formats, addressing modes)

- Organization
  - high level aspects of computer’s design such as the memory system, the bus structure, and the internal CPU (datapath + control) design

- Hardware
  - detailed logic design, interconnection and packing technology, external connections

Processor and Memory Capacity

MOORE’s Law

<table>
<thead>
<tr>
<th>Year</th>
<th>Size</th>
<th>Cycle time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>64 kb</td>
<td>250 ns</td>
</tr>
<tr>
<td>1983</td>
<td>256 Kb</td>
<td>220 ns</td>
</tr>
<tr>
<td>1986</td>
<td>1 Mb</td>
<td>190 ns</td>
</tr>
<tr>
<td>1992</td>
<td>6 Mb</td>
<td>145 ns</td>
</tr>
<tr>
<td>1996</td>
<td>64 Mb</td>
<td>130 ns</td>
</tr>
<tr>
<td>1999</td>
<td>256 Mb</td>
<td>100 ns</td>
</tr>
<tr>
<td>2000</td>
<td>256 Mb</td>
<td>70 ns</td>
</tr>
<tr>
<td>2002</td>
<td>1 Tb</td>
<td>50 ns</td>
</tr>
</tbody>
</table>

Intel 4004, 2300 tr

Intel P4 – 55M tr

Intel McKinley – 221M tr

Reuters, Monday 11 June 2001: Intel engineers have designed and manufactured the world’s smallest and fastest transistor in size of 0.02 microns in size. This will open the way for microprocessors of 1 billion transistors, running at 20 GHz by 2007.

37 transistors per chip, every 1.5 years

Intel engineers have designed and manufactured the world’s smallest and fastest transistor in size of 0.02 microns in size. This will open the way for microprocessors of 1 billion transistors, running at 20 GHz by 2007.
Technology Directions: SIA Roadmap

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature size (nm)</td>
<td>180</td>
<td>130</td>
<td>100</td>
<td>70</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Logic trans/cm²</td>
<td>8.2M</td>
<td>18M</td>
<td>30M</td>
<td>48M</td>
<td>180M</td>
<td>300M</td>
</tr>
<tr>
<td>Gold Trans (np)</td>
<td>1.735</td>
<td>2.85</td>
<td>3.55</td>
<td>4.15</td>
<td>4.80</td>
<td>5.05</td>
</tr>
<tr>
<td>Chip size (in²)</td>
<td>1.867</td>
<td>2.553</td>
<td>3.492</td>
<td>4.776</td>
<td>6.532</td>
<td>8.935</td>
</tr>
<tr>
<td>Clock (MHz)</td>
<td>1250</td>
<td>2100</td>
<td>3500</td>
<td>6000</td>
<td>10000</td>
<td>16000</td>
</tr>
<tr>
<td>Chip size (in²)</td>
<td>240</td>
<td>430</td>
<td>520</td>
<td>620</td>
<td>750</td>
<td>900</td>
</tr>
<tr>
<td>Wiring levels</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Power supply (V)</td>
<td>1.8</td>
<td>1.5</td>
<td>1.2</td>
<td>0.9</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>High-perf pow (W)</td>
<td>60</td>
<td>130</td>
<td>160</td>
<td>200</td>
<td>250</td>
<td>300</td>
</tr>
</tbody>
</table>

Cost, Price, and Their Trends

- Price – what you sell a good for
- Cost – what you spent to produce it

Understanding cost

- Learning curve principle – manufacturing costs decrease over time (even without major improvements in implementation technology)
  - Best measured by change in yield – the percentage of manufactured devices that survives the testing procedure
- Volume (number of products manufactured)
  - Decreases the time needed to get down the learning curve
  - Decreases cost since it increases purchasing and manufacturing efficiency
- Commodities – products sold by multiple vendors in large volumes which are essentially identical
  - Competition among suppliers lowers cost

Prices of DRAM and Intel Pentium III

Integrated Circuits Variable Costs

\[
\text{Cost} = \frac{\text{Die cost} + \text{Testing cost} + \text{Packaging cost}}{\text{Final test yield}}
\]

Cost of die = \( \frac{\text{Cost of wafer}}{\text{Dies per wafer} \times \text{Die yield}} \)

Dies per wafer = \( \frac{\pi \times \text{Wafer diameter}^2}{\text{Die area}} \)

Example: Find the number of dies per 20 cm wafer for a die that is 1.5 cm on a side.

- Die area = 1.5 \times 1.5 = 2.25 cm²
- Die per wafer = \( \frac{3.14 \times 20 \times 20 \times 2.25}{3.14 \times 20 \times 2.5^2} = 110 \)
Integrated Circuits Cost (cont’d)

What is the fraction of good dies on a wafer – die yield

Empirical model
- defects are randomly distributed over the wafer
- yield is inversely proportional to the complexity of the fabrication process

\[ \text{Die yield} = \text{Wafer yield} \times \left(1 + \frac{\text{Defects per unit area} \times \text{Die area}}{\alpha} \right) \]

- Wafer yield accounts for wafers that are completely bad (no need to test them); We assume the wafer yield is 100%
- Defects per unit area: typically 0.4 – 0.8 per cm²
- \( \alpha \) corresponds to the number of masking levels; for today’s CMOS, a good estimate is \( \alpha = 4.0 \)

Example: Find die yield for dies with 1 cm and 0.7 cm on a side; defect density is 0.6 per square centimeter

- For larger die: \( (1 + 0.6 \times 1/4)^4 = 0.57 \)
- For smaller die: \( (1 + 0.6 \times 0.49/4)^4 = 0.75 \)

Die costs are proportional to the fourth power of the die area

\[ \text{Die cost} = f(\text{Die area}^4) \]

In practice \( \text{Die cost} = f(\text{Die area}^4) \)

Real World Examples

<table>
<thead>
<tr>
<th>Chip</th>
<th>RL</th>
<th>Die cost</th>
<th>Wafer cost</th>
<th>Defect [cm²]</th>
<th>Area [cm²]</th>
<th>Dies/ wafer</th>
<th>Yield</th>
<th>Die cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDIX</td>
<td>2</td>
<td>1.05</td>
<td>3000</td>
<td>1.0</td>
<td>43</td>
<td>350</td>
<td>74%</td>
<td>38</td>
</tr>
<tr>
<td>Greycity</td>
<td>4</td>
<td>2.65</td>
<td>6100</td>
<td>5.6</td>
<td>115</td>
<td>172</td>
<td>56%</td>
<td>812</td>
</tr>
<tr>
<td>PowerPC 601</td>
<td>4</td>
<td>1.60</td>
<td>51700</td>
<td>1.3</td>
<td>121</td>
<td>115</td>
<td>38%</td>
<td>553</td>
</tr>
<tr>
<td>HI PA 7100</td>
<td>1</td>
<td>1.60</td>
<td>3100</td>
<td>1.6</td>
<td>100</td>
<td>66</td>
<td>50%</td>
<td>313</td>
</tr>
<tr>
<td>Dec Alpha</td>
<td>1</td>
<td>1.70</td>
<td>61500</td>
<td>1.7</td>
<td>234</td>
<td>33</td>
<td>5%</td>
<td>419</td>
</tr>
<tr>
<td>Sun SPARC</td>
<td>5</td>
<td>0.70</td>
<td>81700</td>
<td>1.6</td>
<td>256</td>
<td>48</td>
<td>12%</td>
<td>372</td>
</tr>
<tr>
<td>Pentium</td>
<td>6</td>
<td>0.70</td>
<td>91500</td>
<td>1.5</td>
<td>296</td>
<td>40</td>
<td>9%</td>
<td>217</td>
</tr>
</tbody>
</table>

Typical in 2002:
- 8 cm diameter wafer, 4-6 metal layers, wafer cost $5K-$6K


Things to Remember

- Computing classes: desktop, server, embedd.
- Technology trends:
  - Learning curve: manufacturing costs decrease over time
  - Volume: the number of chips manufactured
  - Commodity

Cost
- Capacity
- Speed

<table>
<thead>
<tr>
<th>Logic</th>
<th>4x in 3 years</th>
<th>2x in 3 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM</td>
<td>4x in 3 years</td>
<td>2x in 3 years</td>
</tr>
<tr>
<td>Disk</td>
<td>4x in 3 years</td>
<td>2x in 3 years</td>
</tr>
</tbody>
</table>
### Things to Remember (cont’d)

- **Cost of an integrated circuit**

\[
\text{IC cost} = \frac{\text{Die cost} + \text{Testing cost} + \text{Packaging cost}}{\text{Final test yield}}
\]

- **Cost of die**

\[
\text{Cost of die} = \frac{\text{Cost of wafer}}{\text{Dies per wafer} \times \text{Die yield}}
\]

- **Dies per wafer**

\[
\text{Dies per wafer} = \frac{\pi \times (\text{Wafer diameter} / 2)^2}{\text{Die area}} = \frac{\pi \times \text{Wafer diameter}}{\sqrt{3} \times \text{Die area}}
\]

- **Die yield**

\[
\text{Die yield} = \text{Wafer yield} \times \left(1 - \frac{\text{Defects per unit area} \times \text{Die area}}{\alpha}\right)^{1/2}
\]