The Big Picture: Where are We Now?

° The Five Classic Components of a Computer

° Today’s Topic: Memory System

Who Cares About the Memory Hierarchy?

Processor-DRAM Memory Gap (latency)

μProc

60%/yr. (2X/1.5yr)

“Moore’s Law”

Processor-Memory Performance Gap:
(grows 50% / year)

DRAM 9%/yr. (2X/10 yrs)

Memory Hierarchy: Principles of Operation

° At any given time, data is copied between only 2 adjacent levels:
  - Upper Level: the one closer to the processor
    - Smaller, faster, and uses more expensive technology
  - Lower Level: the one further away from the processor
    - Bigger, slower, and uses less expensive technology

° Block:
  - The minimum unit of information that can either be present or not present in the two level hierarchy

Memory Hierarchy: Terminology

° Hit: data appears in some block in the upper level (example: Block X)
  - Hit Rate: the fraction of memory access found in the upper level
  - Hit Time: Time to access the upper level which consists of RAM access time + Time to determine hit/miss

° Miss: data needs to be retrieve from a block in the lower level (Block Y)
  - Miss Rate = 1 - (Hit Rate)
  - Miss Penalty: Time to replace a block in the upper level + Time to deliver the block to the processor

° Hit Time << Miss Penalty
Memory Hierarchy: How Does it Work?

- **Temporal Locality** (Locality in Time): If an item is referenced, it will tend to be referenced again soon.
  - Keep more recently accessed data items closer to the processor
- **Spatial Locality** (Locality in Space): If an item is referenced, items whose addresses are close by tend to be referenced soon.
  - Move blocks consists of contiguous words to the upper levels

Memory Hierarchy of a Modern Computer System

- By taking advantage of the principle of locality:
  - Present the user with as much memory as is available in the cheapest technology.
  - Provide access at the speed offered by the fastest technology.

Memory Hierarchy Technology

- Random Access:
  - “Random” is good: access time is the same for all locations
  - **DRAM**: Dynamic Random Access Memory
    - High density, low power, cheap, slow
    - Dynamic: need to be “refreshed” regularly
  - **SRAM**: Static Random Access Memory
    - Low density, high power, expensive, fast
    - Static: content will last “forever”
- “Non-so-random” Access Technology:
  - Access time varies from location to location and from time to time
  - Examples: Disk, tape drive, CDROM
  - The next two lectures will concentrate on random access technology
    - The Main Memory: DRAMs
    - Caches: SRAMs

Technology Trends

<table>
<thead>
<tr>
<th>Year</th>
<th>DRAM Size (MB)</th>
<th>Cycle Time (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>64</td>
<td>250</td>
</tr>
<tr>
<td>1983</td>
<td>256</td>
<td>220</td>
</tr>
<tr>
<td>1986</td>
<td>1</td>
<td>190</td>
</tr>
<tr>
<td>1989</td>
<td>4</td>
<td>165</td>
</tr>
<tr>
<td>1992</td>
<td>16</td>
<td>145</td>
</tr>
<tr>
<td>1995</td>
<td>64</td>
<td>120</td>
</tr>
</tbody>
</table>

SPARCstation 20’s Memory Module

- Supports a wide range of sizes:
  - Smallest: 4 MB = 16 2Mb DRAM chips, 8 KB of Page Mode SRAM
  - Biggest: 64 MB = 32 16Mb chips, 16 KB of Page Mode SRAM
Summary:

Two Different Types of Locality:
- Temporal Locality (Locality in Time): If an item is referenced, it will tend to be referenced again soon.
- Spatial Locality (Locality in Space): If an item is referenced, items whose addresses are close by tend to be referenced soon.

By taking advantage of the principle of locality:
- Present the user with as much memory as is available in the cheapest technology.
- Provide access at the speed offered by the fastest technology.

DRAM is slow but cheap and dense:
- Good choice for presenting the user with a BIG memory system

SRAM is fast but expensive and not very dense:
- Good choice for providing the user FAST access time.