Big Picture

• Memory
  – Just an “ocean of bits”
  – Many technologies are available

• Key issues
  – Technology (how bits are stored)
  – Placement (where bits are stored)
  – Identification (finding the right bits)
  – Replacement (finding space for new bits)
  – Write policy (propagating changes to bits)

• Must answer these regardless of memory type
Memory Performance Gap

![Graph showing the performance gap between processor and memory over the years from 1980 to 2010. The graph illustrates a significant increase in processor performance compared to memory performance.]
# Types of Memory

<table>
<thead>
<tr>
<th>Type</th>
<th>Size</th>
<th>Speed</th>
<th>Cost/bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register</td>
<td>&lt; 1KB</td>
<td>&lt; 1ns</td>
<td>$$$$$</td>
</tr>
<tr>
<td>On-chip SRAM</td>
<td>8KB-6MB</td>
<td>&lt; 10ns</td>
<td>$$</td>
</tr>
<tr>
<td>Off-chip SRAM</td>
<td>1Mb – 16Mb</td>
<td>&lt; 20ns</td>
<td>$$</td>
</tr>
<tr>
<td>DRAM</td>
<td>64MB – 1TB</td>
<td>&lt; 100ns</td>
<td>$</td>
</tr>
<tr>
<td>Disk</td>
<td>40GB – 1PB</td>
<td>&lt; 20ms</td>
<td>~0</td>
</tr>
</tbody>
</table>
Technology - Registers

Diagram showing the connections and functionality of registers in a digital circuit, including decoders and data paths.

Legend:
- C_Adx
- A_Adx
- B_Adx
- C = Write Port
- A.B = Read Ports
Technology – SRAM

“Word” Lines
-select a row

“Bit” Lines
-carry data in/out

Data_C(i)

A_A dx
select

B_A dx
select

C_A dx
select

Register Bit Slice

Data_A(i)

Data_B(i)

C = Write Port
A,B = Read Ports

13 May 2013
Computer Architecture@IIT Mandi

CADSL
Technology – DRAM

• Logically similar to SRAM
• Commodity DRAM chips
  – E.g. 1Gb
  – Standardized address/data/control interfaces
• Very dense
  – 1T per cell (bit)
  – Data stored in capacitor – decays over time
    • Must rewrite on read, refresh
• Density improving vastly over time
• Latency barely improving
Memory Timing – Read

- Latch-based SRAM or asynchronous DRAM (FPM/EDO)
  - Multiple chips/banks share address bus and tristate data bus
  - Enables are decoded from address to select bank
    - E.g. bbbbbbb0 is bank 0, bbbbbbb1 is bank 1
- Timing constraints: straightforward
  - $t_{EN}$ setup time from Addr stable to EN active (often zero)
  - $t_D$ delay from EN to valid data (10ns typical for SRAM)
  - $t_O$ delay from EN disable to data tristate off (nonzero)
Memory Timing - Write

- WR & EN triggers write of Data to ADDR
- Timing constraints: not so easy
  - $t_S$ setup time from Data & Addr stable to WR pulse
  - $t_P$ minimum write pulse duration
  - $t_H$ hold time for data/addr beyond write pulse end
- Challenge: WR pulse must start late, end early
  - $>t_S$ after Addr/Data, $>t_H$ before end of cycle
  - Requires multicycle control or glitch-free clock divider
Technology – Disk

• Bits stored as magnetic charge
• Still mechanical!
  – Disk rotates (3600-15000 RPM)
  – Head seeks to track, waits for sector to rotate to it
  – Solid-state replacements are becoming popular

• Glacially slow compared to DRAM (10-20ms)
• Density improvements astounding (100%/year)
Memory System

- Programmers want unlimited amounts of memory with low latency
- Fast memory technology is more expensive per bit than slower memory
- **Solution:** organize memory system into a hierarchy
  - Entire addressable memory space available in largest, slowest memory
  - Incrementally smaller and faster memories, each containing a subset of the memory below it, proceed in steps up toward the processor
- **Temporal and spatial locality** insures that nearly all references can be found in smaller memories
  - Gives the allusion of a large, fast memory being presented to the processor
Memory Hierarchy

- Registers
- On-Chip SRAM
- Off-Chip SRAM
- DRAM
- Disk

CAPACITY: Downward
SPEED and COST: Upward
Memory Hierarchy

(a) Memory hierarchy for server

- CPU
  - Registers
  - L1 Cache
    - Register reference
    - Size: 1000 bytes
    - Speed: 300 ps
    - Level 1 Cache
      - Cache reference
      - Size: 64 KB
      - Speed: 1 ns
      - Level 2 Cache Reference
      - Size: 256 KB
      - Speed: 3-10 ns
      - Level 3 Cache Reference
      - Size: 2-4 MB
      - Speed: 10-20 ns
      - Memory reference
      - Size: 4-16 GB
      - Speed: 50-100 ns
    - Memory bus
    - I/O bus
    - Disk storage
      - Size: 4-16 TB
      - Speed: 5-10 ms

(b) Memory hierarchy for a personal mobile device

- CPU
  - Registers
  - L1 Cache
    - Register reference
    - Size: 500 bytes
    - Speed: 500 ps
    - Level 1 Cache
      - Cache reference
      - Size: 64 KB
      - Speed: 2 ns
      - Level 2 Cache Reference
      - Size: 256 KB
      - Speed: 10-20 ns
      - Level 3 Cache Reference
      - Size: 256-512 MB
      - Speed: 50-100 ns
      - Memory reference
    - Memory bus
    - Storage
      - Size: 4-8 GB
      - Speed: 25-50 us
      - FLASH memory reference
Thank You