

Memory System

Virendra Singh

Associate Professor

Computer Architecture and Dependable Systems Lab

Department of Electrical Engineering
Indian Institute of Technology Bombay

<http://www.ee.iitb.ac.in/~viren/>

E-mail: viren@ee.iitb.ac.in

Computer Organization & Architecture



Lecture 21 (13 May 2013)

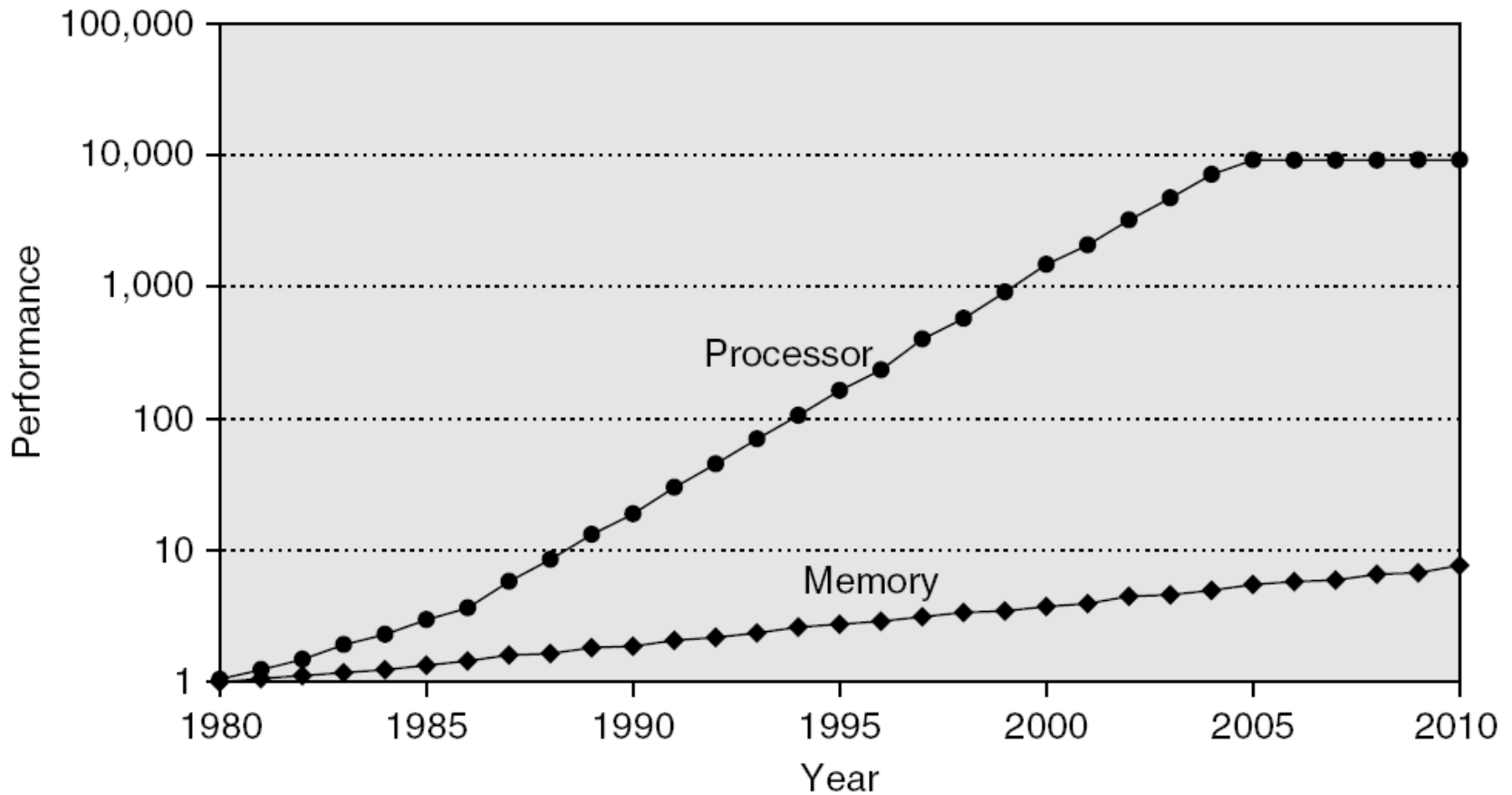
CADSL

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

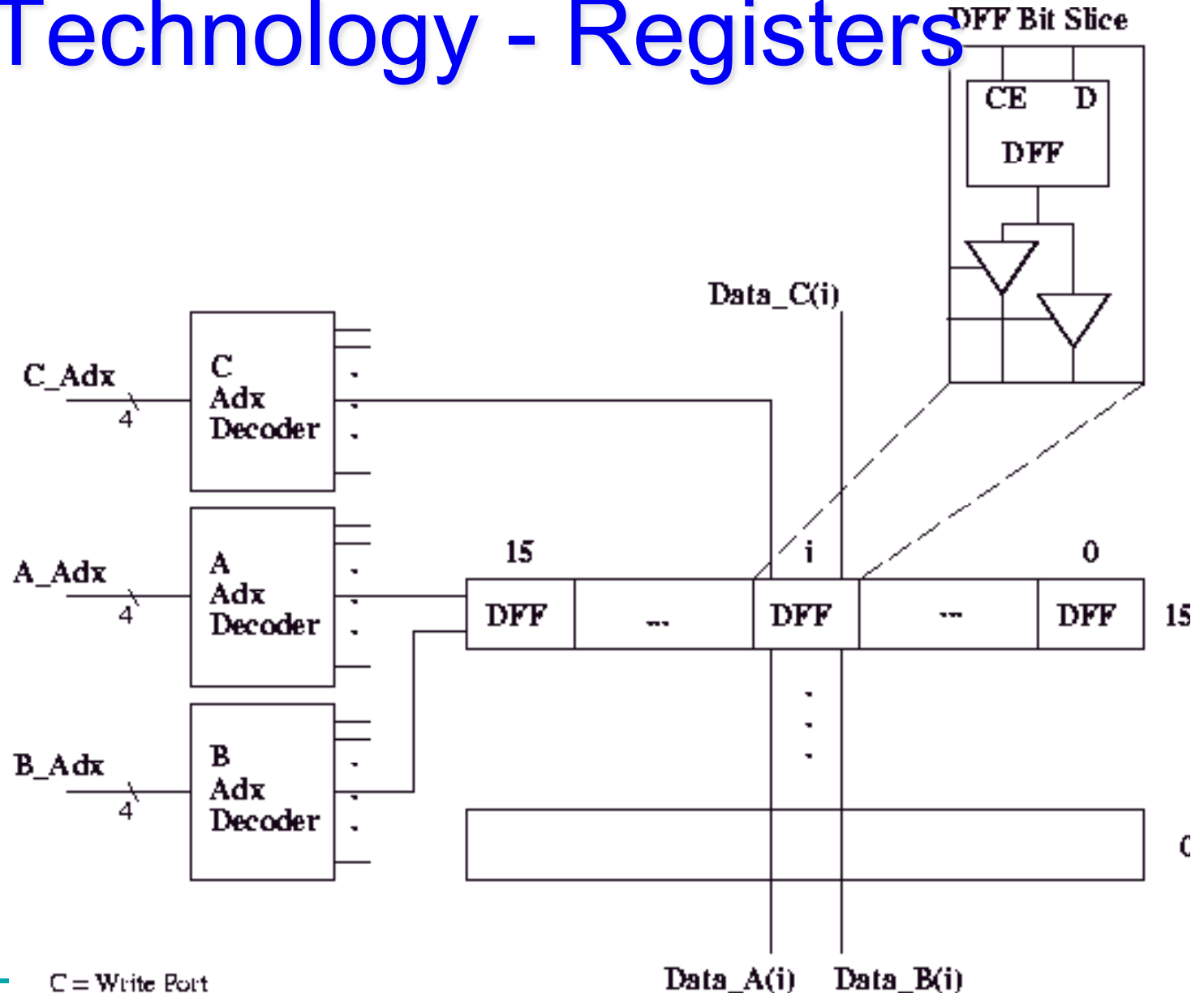


Types of Memory

Type	Size	Speed	Cost/bit
Register	< 1KB	< 1ns	\$\$\$\$
On-chip SRAM	8KB-6MB	< 10ns	\$\$\$
Off-chip SRAM	1Mb – 16Mb	< 20ns	\$\$
DRAM	64MB – 1TB	< 100ns	\$
Disk	40GB – 1PB	< 20ms	~0



Technology - Registers



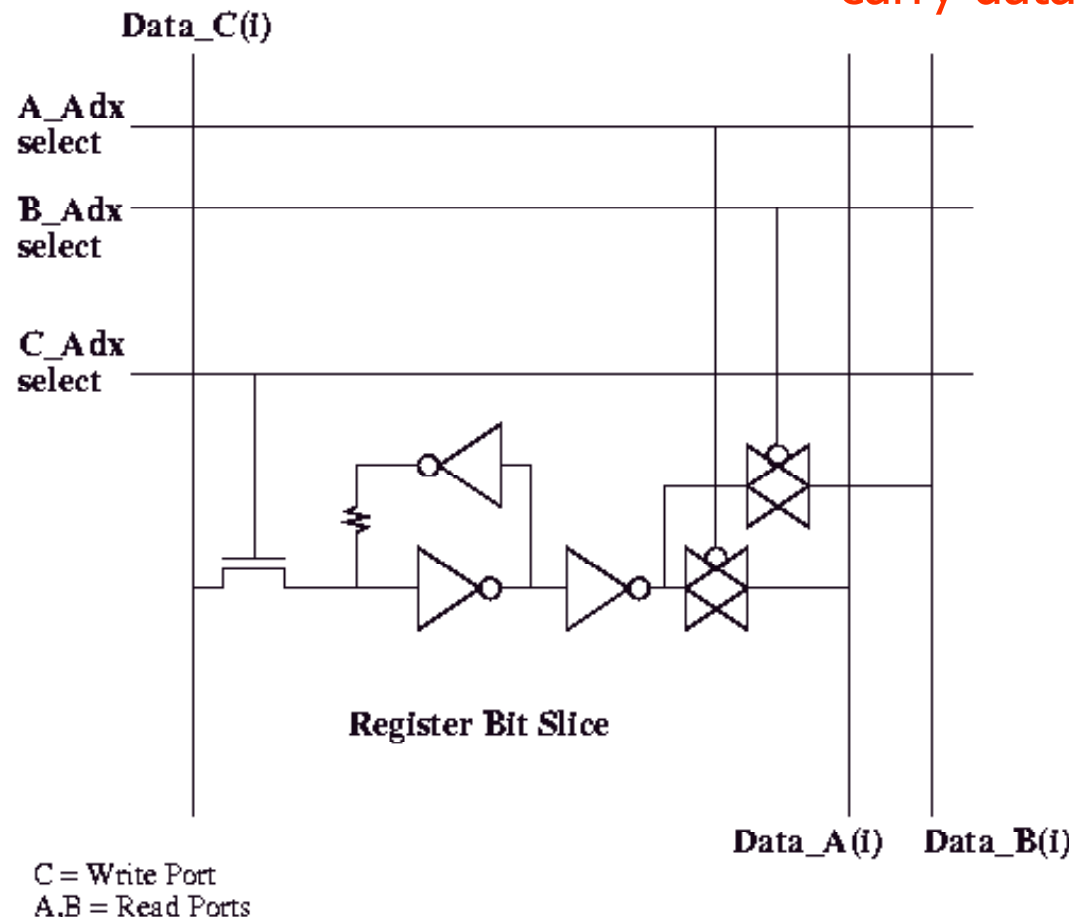
C = Write Port
A, B = Read Ports



Technology – SRAM

“Bit” Lines
-carry data in/out

“Word” Lines
-select a row

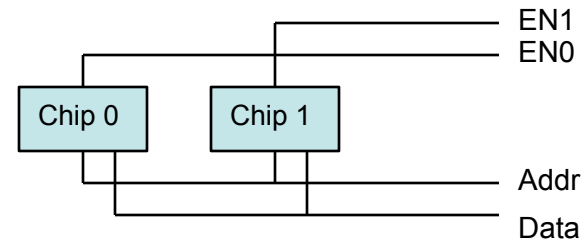
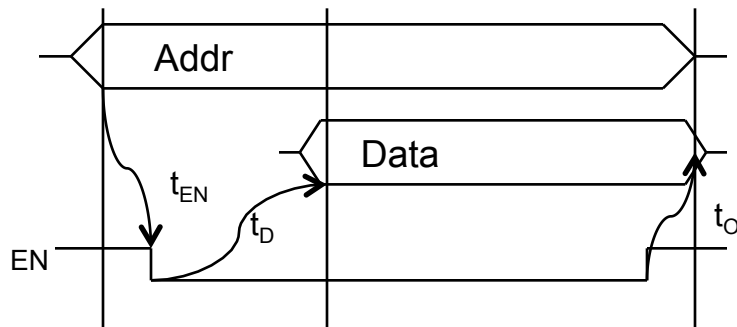


Technology – DRAM

- 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

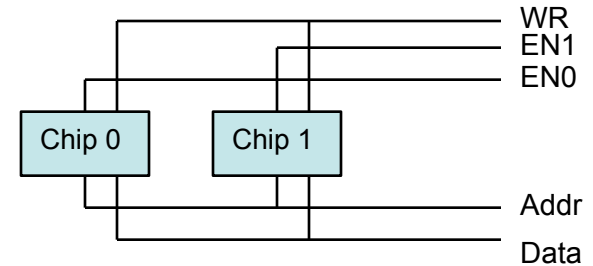
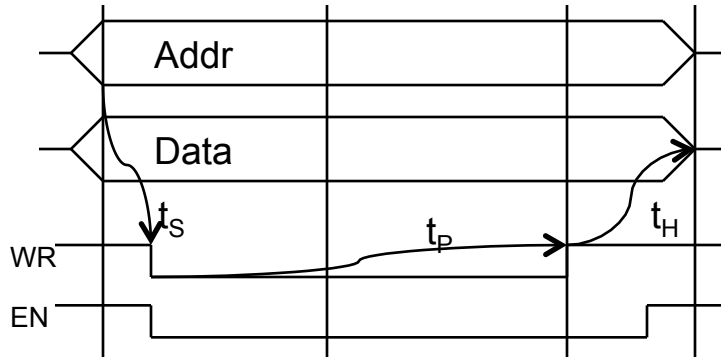


Memory Timing – Read



- 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)

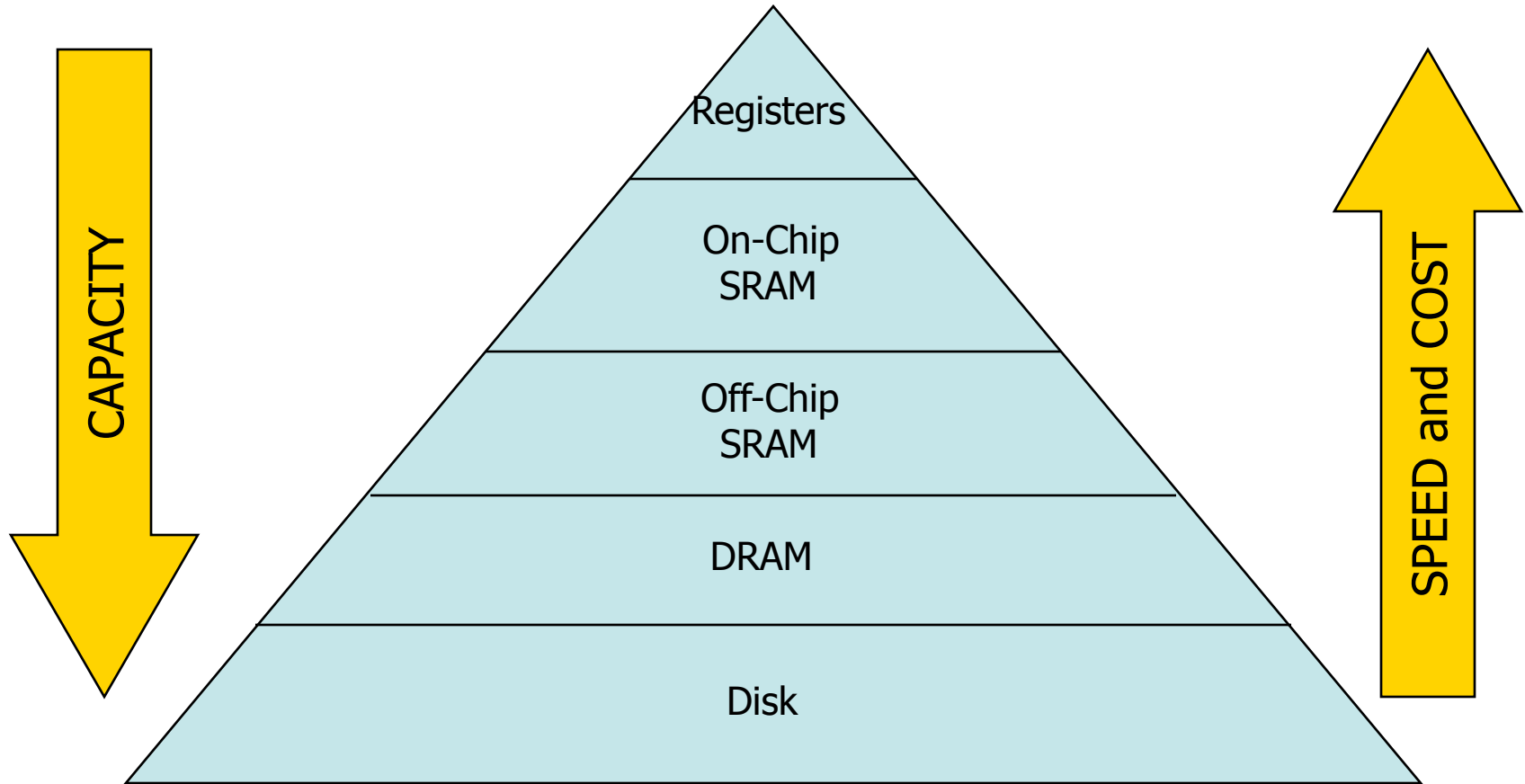


Memory System

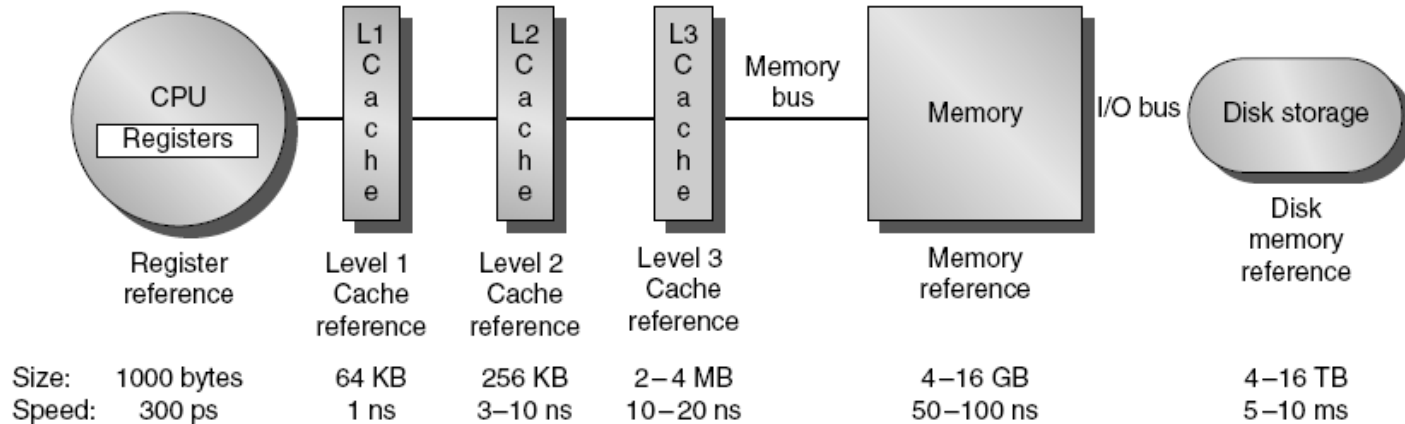
- Programmers want unlimited amounts of memory with low latency
- Fast memory technology is more expensive per bit than slower memory
- **Solution:** organize memory system into a hierarchy
 - Entire addressable memory space available in largest, slowest memory
 - Incrementally smaller and faster memories, 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 illusion of a large, fast memory being presented to the processor



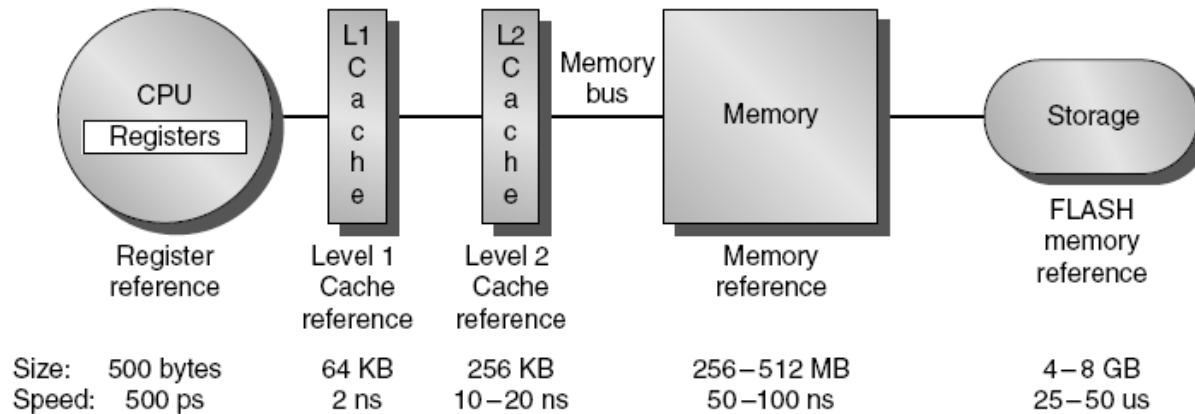
Memory Hierarchy



Memory Hierarchy



(a) Memory hierarchy for server



(b) Memory hierarchy for a personal mobile device

Thank You

