CSC 3210 Computer organization and programming

Chapter 5 Memory Hierarchy

Chunlan Gao

Georgia State University

Final

Appendix for final, print it and take it to the test:

Test_CSC_3210_Appendix

About the RAM :

- Stands for Random Access Memory
- Temporary, high-speed storage
- Volatile (data lost when power is off)
- Used to store active data and running programs

- Enables multitasking
- Determines performance
- Reduces load times
- Acts as a bridge between storage and CPU

13:30- 15:30

Types of RAM

- DRAM (Dynamic RAM): Main memory, inexpensive, high capacity, needs refreshing
- SRAM (Static RAM): Used in CPU caches, faster, expensive, no refreshing needed

How RAM Works

- Programs are loaded into RAM from disk
- CPU accesses data from RAM during execution
- RAM provides quick read/write access
- Data is cleared when the system shuts down

RAM Capacity and Speed

- Common sizes: 4GB, 8GB, 16GB, etc.
- Data bus: 64-bit systems = 8 bytes per cycle
- Frequency: e.g., DDR4-3200 = 3200 MT/s(DDR = Double Data Rate)
- (**MT/s** stands for **Mega Transfers per second**, which means **millions of data transfers per second**. It's commonly used to describe the **data rate** of memory (like DDR RAM) or high-speed buses.)

Unit	Meaning	Example	Notes
MHz	Million clock cycles per second	1600 MHz	Refers to clock frequency
MT/s	Million **data transfers** per second	3200 MT/s (DDR4)	Refers to the actual data rate

DDR (Double Data Rate) memory **transfers data twice per clock cycle**: once on the rising edge and once on the falling edge.

If your memory runs at **1600 MHz**, it delivers **3200 MT/s**.

Principle of Locality

- Programs access a small proportion of their address space at any time.

Analogy: Photographer's Studio

- A photographer works at a desk editing wedding photos.
- Photos are stored in a large archive (like a disk).
- Only a set of related photos (e.g., ceremony) are on the desk at one time.
- When switching topics (e.g., from ceremony to reception), they swap photo sets.
- Frequent photos remain on the desk—similar to frequently used data in RAM.

Comparison

- Desk = RAM or cache (fast, limited)
- Archive cabinet = disk or main memory (large, slower)
- Book/photo set = working set of data
- Going to shelves/archives = page swapping / memory access
- Temporal Locality= **frequently reuse the same set of photos** ((e.g., ceremony shots) while editing)
- Spatial Locality = **stored close together** in the archive (When you bring out one set (e.g., "ceremony"), you often **work with the surrounding ones** too.)

Principle of Locality

- Temporal locality
  - Items accessed recently are likely to be accessed again soon
  - e.g., instructions in a loop, induction variables
- Spatial locality
  - Items near those accessed recently are likely to be accessed soon
  - E.g., sequential instruction access, array data

Taking Advantage of Locality

- Memory hierarchy
- Store everything on disk
- Copy recently accessed (and nearby) items from disk to smaller DRAM memory
  - Main memory
- Copy more recently accessed (and nearby) items from DRAM to smaller SRAM memory
  - Cache memory attached to CPU

Memory Hierarchy

A memory hierarchy consists of multiple levels of memory with different speeds and sizes. The faster memories are more expensive per bit than the slower memories and thus are smaller.

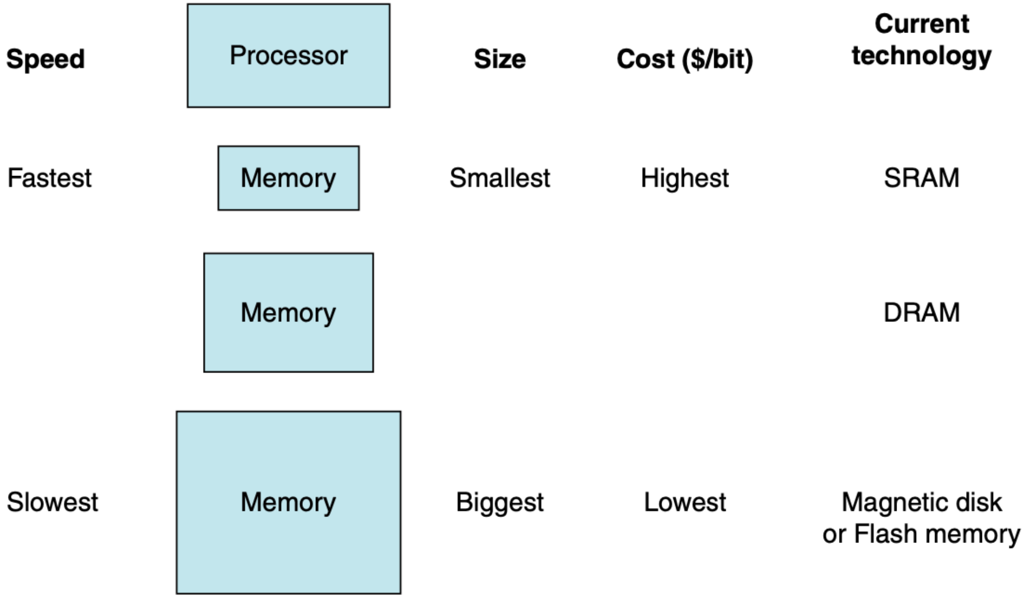
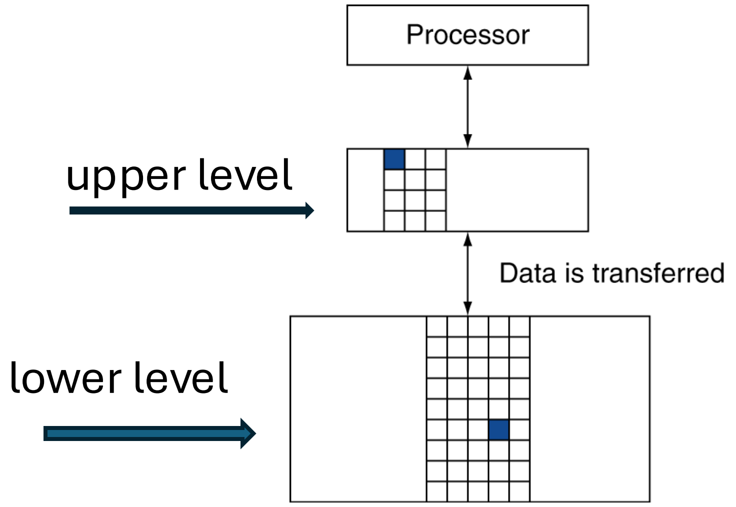


FIGURE 5.1 The basic structure of a memory hierarchy. By implementing the memory system as a hierarchy, the user has the illusion of a memory that is as large as the largest level of the hierarchy, but can be accessed as if it were all built from the fastest memory. Flash memory has replaced disks in many personal mobile devices, and may lead to a new level in the storage hierarchy for desktop and server computers; see Section 5.2.

Memory Hierarchy Levels

- Block (aka line): unit of copying
  - May be multiple words
- If accessed data is present in upper level
  - Hit: access satisfied by upper level
    - Hit ratio: hits/accesses
- If accessed data is absent
  - Miss: block copied from lower level
    - Time taken because of miss: miss penalty
    - Miss ratio: misses/accesses
      = 1 – hit ratio
  - Then accessed data supplied from upper level



Memory Accesses	Hits	Misses
1,000	920	80

Then:

- **Hit Rate** = 920 / 1,000 = **92%**
- **Miss Rate** = 1 – 0.92 = **8%**

5.2 Memory Technologies

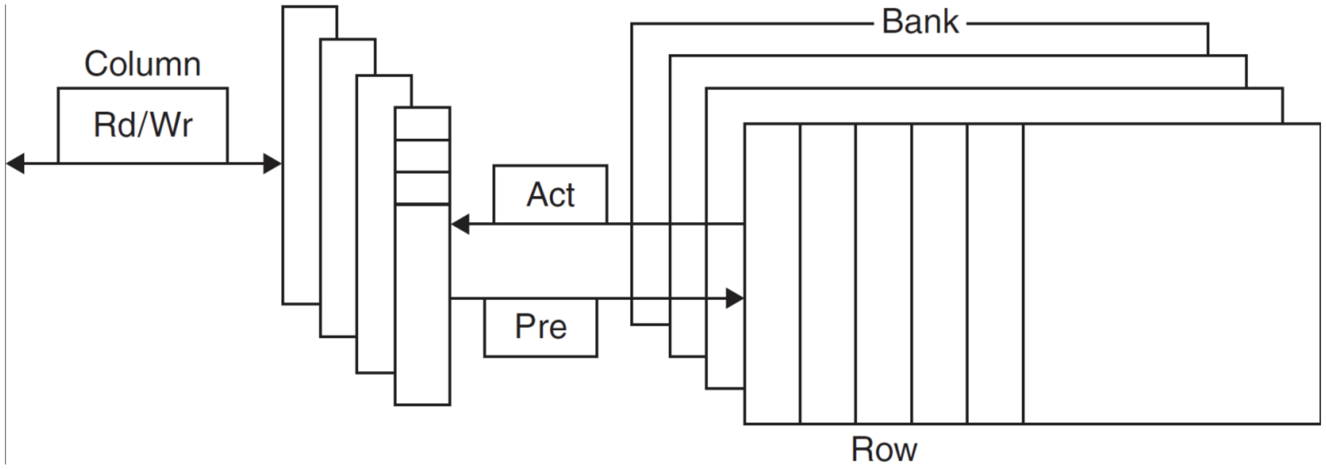
Memory technology	Typical access time	$ per GiB in 2020
SRAM semiconductor memory	0.5–2.5 ns	$500–$1000
DRAM semiconductor memory	50–70 ns	$3–$6
Flash semiconductor memory	5,000–50,000 ns	$0.06–$0.12
Magnetic disk	5,000,000–20,000,000 ns	$0.01–$0.02

Access time is the **time it takes** to read (or write) data from a memory or storage device after a request is made.

DRAM Technology

Must periodically be refreshed

- Read contents and write back
- Performed on a DRAM “row”



Analogy – DRAM as a Library

Think of DRAM like a big library (the memory):

•Each **bank** is a bookshelf.

•Each **row** is a book on that shelf.

•Each **column** is a specific page.

•You must:

- Close the previous book (Precharge)
- Open the new book (Activate)
- Go to the right page (Column)
- Read/write the content (Read/Write)

Advanced DRAM Organization

- Bits in a DRAM are organized as a rectangular array
  - DRAM accesses an entire row
  - Burst mode: supply successive words from a row with reduced latency
- Double data rate (DDR) DRAM
  - Transfer on rising and falling clock edges
- Quad data rate (QDR) DRAM
  - Separate DDR inputs and outputs

DRAM Generations

Year	Capacity	$/GB
1980	64 Kibibit	$6,480,000
1983	256 Kibibit	$1,980,000
1985	1 Mebibit	$720,000
1989	4 Mebibit	$128,000
1992	16 Mebibit	$30,000
1996	64 Mebibit	$9,000
1998	128 Mebibit	$900
2000	256 Mebibit	$840
2004	512 Mebibit	$150
2007	1 Gibibit	$40
2010	2 Gibibit	$13
2012	4 Gibibit	$5
2015	8 Gibibit	$7
2018	16 Gibibit	$6

DRAM Performance Factors

- Row buffer
  - Allows several words to be read and refreshed in parallel
- Synchronous DRAM
  - Allows for consecutive accesses in bursts without needing to send each address
  - Improves bandwidth
- DRAM banking
  - Allows simultaneous access to multiple DRAMs
  - Improves bandwidth

In computer architecture, **bandwidth** refers to the **amount of data** that can be **transferred per unit time** between memory and processor.

Bytes per second (B/s)

Megabytes per second (MB/s)

Gigabytes per second (GB/s)

Bandwidth

Assume DDR4-3200 memory, 64-bit (8-byte) bus:

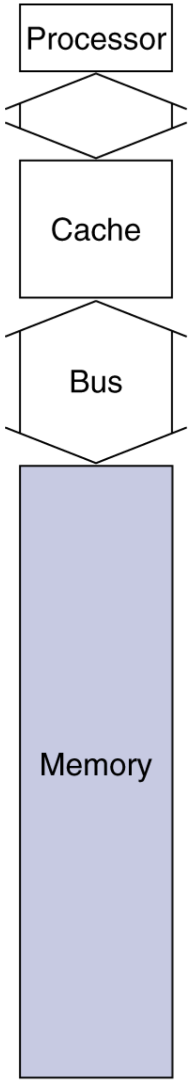
- Transfer rate: **3200 MT/s**
- Bus width: **64 bits = 8 bytes**
- Bandwidth:

$$8\,\text{Bytes} \times 3200 \times 10^6 = 25.6\,\text{GB/s}$$

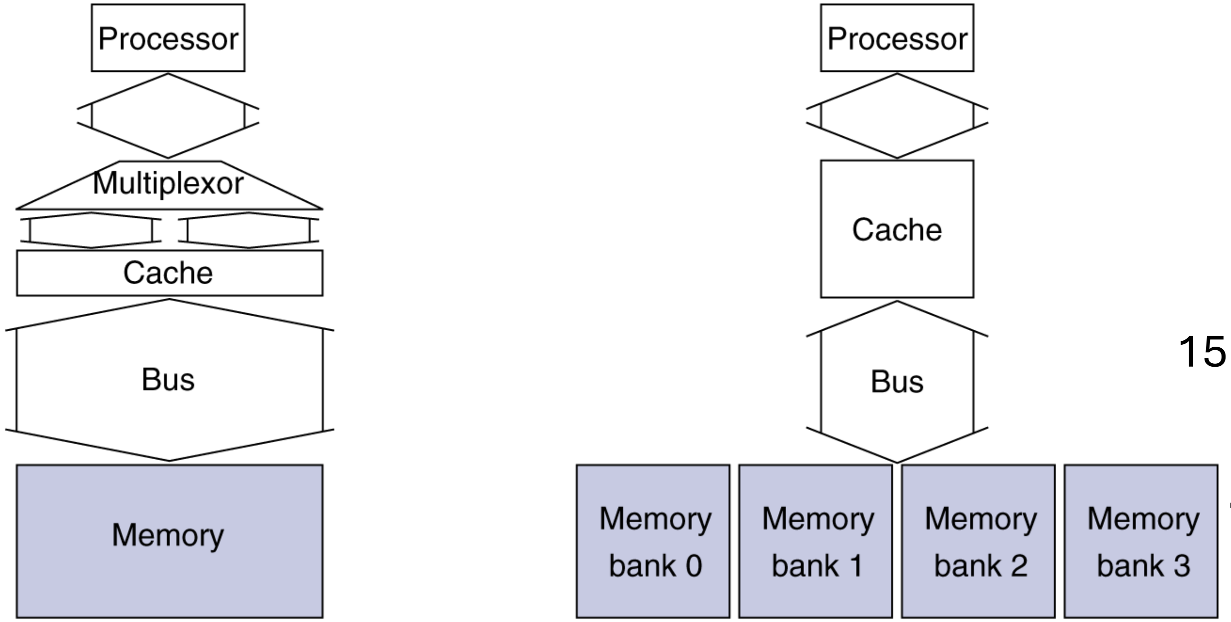
If you have **dual-channel**, that doubles to:

$$25.6 \times 2 = 51.2\,\text{GB/s}$$

Increasing Memory Bandwidth



a. One-word-wide memory organization



b. Wider memory organization

c. Interleaved memory organization

Term	Meaning
Miss penalty	How many **bus cycles** it takes to fetch the missed block from memory into cache
Bandwidth	How much **data (in bytes)** can be transferred **per bus cycle**
Word	Usually 4 bytes (so 4 words = 16 bytes)
Bus cycle	One memory access transfer opportunity (1 cycle = 1 slot for data transfer)

1 cycle to return 16 by

1 cycle to send address

15 cycles to access memory (latency)

- 4-word wide memory
  - Miss penalty = 1 + 15 + 1 = 17 bus cycles
  - Bandwidth = 16 bytes / 17 cycles = 0.94 B/cycle
- 4-bank interleaved memory
  - Miss penalty = 1 + 15 + 4×1 = 20 bus cycles
  - Bandwidth = 16 bytes / 20 cycles = 0.8 B/cycle

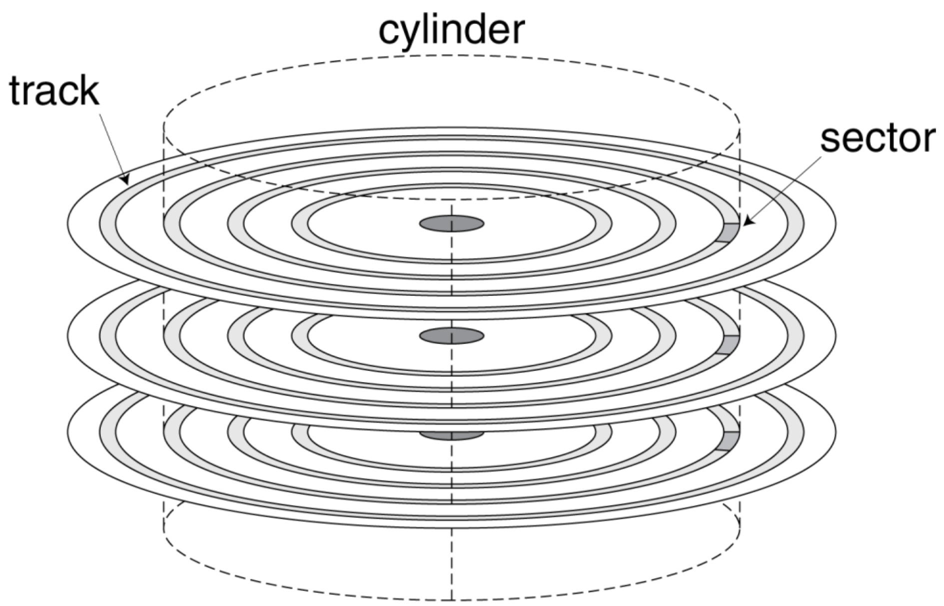
Flash Storage

- Nonvolatile semiconductor storage
  - 100× – 1000× faster than disk
  - Smaller, lower power, more robust
  - But more $/GB (between disk and DRAM)

Flash Types

- NOR flash: bit cell like a NOR gate
  - Random read/write access
  - Used for instruction memory in embedded systems
- NAND flash: bit cell like a NAND gate
  - Denser (bits/area), but block-at-a-time access
  - Cheaper per GB
  - Used for USB keys, media storage, …
- Flash bits wears out after 1000's of accesses
  - Not suitable for direct RAM or disk replacement
  - Wear leveling: remap data to less used blocks(Wear leveling = **smart remapping** of data in flash memory to **prevent certain blocks from wearing out early**, thereby **prolonging device life**.)

Disk Storage



Disk Sectors and Access

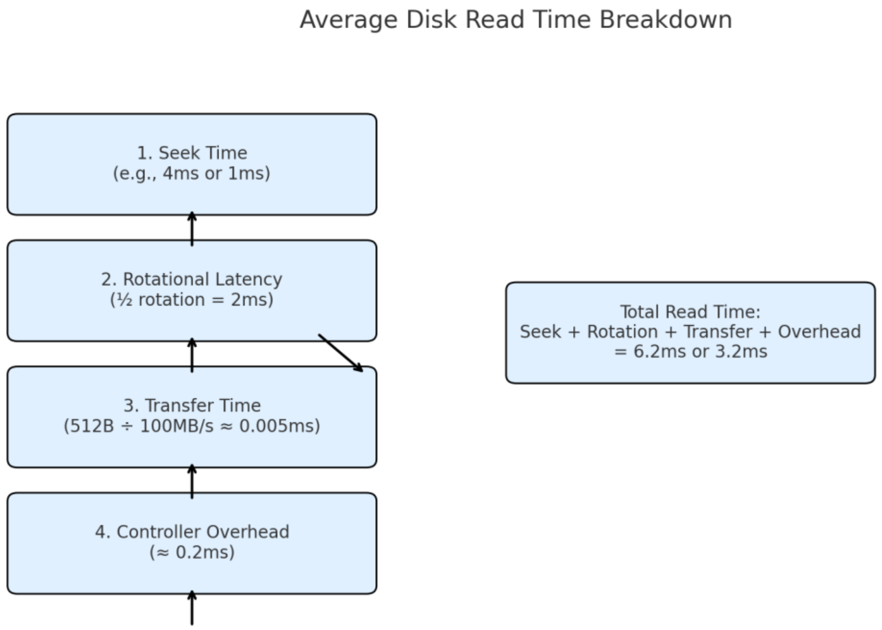
- Each sector records
  - Sector ID
  - Data (512 bytes, 4096 bytes proposed)
  - Error correcting code (ECC)
    - Used to hide defects and recording errors
  - Synchronization fields and gaps
- Access to a sector involves
  - Queuing delay if other accesses are pending
  - Seek: move the heads
  - Rotational latency
  - Data transfer
  - Controller overhead

Disk Access Example

- Given
  - 512B sector, 15,000rpm, 4ms average seek time, 100MB/s transfer rate, 0.2ms controller overhead, idle disk

15,000rpm

60s/m/15000rpm= 4ms/r



Disk Performance Issues

- Manufacturers quote average seek time
  - Based on all possible seeks
  - Locality and OS scheduling lead to smaller actual average seek times
- Smart disk controller allocate physical sectors on disk
  - Present logical sector interface to host
  - SCSI, ATA, SATA
- Disk drives include caches
  - Prefetch sectors in anticipation of access
  - Avoid seek and rotational delay

Disk Performance issues

Protocol	Full Name	Usage
SCSI	Small Computer System Interface	Mainly used in servers and workstations
ATA	Advanced Technology Attachment	Traditional hard drive interface (now replaced by SATA)
SATA	Serial ATA	One of the most common disk interface standards today