Master of Computer Applications

# CAPOL403R01: Computer Organization & Architecture

## Unit III: Lecture 1
## Memory System Overview

**Dr. D. MURALIDHARAN**
**School of Computing**
**SASTRA Deemed to be University**

# Characteristics of memory

- Location
- Capacity
- Unit of transfer
- Access method
- Performance
- Physical type
- Physical characteristics
- Organization

# Location

- Internal
  - Internal memory has a direct communication with the processor
  - Usually, they are always attached to the processor
  - Example: processor register, cache memory and internal (primary) memory
- External
  - External memories are communicated with the processor through IO controller.
  - These are slow, high capacity memories.
  - Usually they are attached to the computer when needed and detached from it when not needed.
  - Eg Optical disks, pen drives

# Capacity

- Words
  - Internal memory capacity is expressed in terms of words (1, 2 or 4 bytes)
- Bytes
  - The capacity of peripheral devices is expressed in terms of bytes

# Unit of transfer

- Word
  - The communication between processor and internal memory is in terms of words (bytes/words)
- Block
  - The communication between processor and external memory is in terms of blocks (blocks of bytes)

# Accessing method

- Sequential access
  - The read/write head is moved from current location to the desired location in order and reject the addresses (or records) in between.
  - The accessing time will vary depends on the 'distance' between the current location of the header and the desired location
  - Eg: Tape units

- Direct access
  - Individual blocks or records have a unique address based on physical location
  - Access is accomplished by direct access to reach a general vicinity plus sequential searching, counting, or waiting to reach the final location
  - Access time is variable
  - Eg: Disk units

# Accessing method…

- Random access
  - Each addressable location in memory has a unique, physically wired-in addressing mechanism
  - Same, constant amount of time is needed to access all locations
  - Eg: Main memory

- Associative access
  - This is a special kind of random access
  - It enables one to make a comparison of desired bit locations within a word for a specified match
  - This comparison will be done for all words simultaneously
  - Thus, a word is retrieved based on a portion of its contents rather than its address
  - Eg: Cache memory

# Accessing method…

- Random access
  - Each addressable location in memory has a unique, physically wired-in addressing mechanism
  - Same, constant amount of time is needed to access all locations
  - Eg: Main memory

- Associative access
  - This is a special kind of random access
  - It enables one to make a comparison of desired bit locations within a word for a specified match
  - This comparison will be done for all words simultaneously
  - Thus, a word is retrieved based on a portion of its contents rather than its address
  - Eg: Cache memory

# Performance

- Access time (Latency)
  - This is the time from the instant that an address is presented to the memory to the instant that data have been stored or made available for use
  - For non-random-access memory, access time is the time it takes to position the read–write mechanism at the desired location
- Memory cycle time
  - This time is the sum of access time and an additional time
  - The additional time is needed to die-out the previous signals from bus
  - Memory cycle time is concerned with the system bus, not the processor
  - In random access memories, this is the minimum expected time between two successive memory accesses

# Performance…

- Transfer rate:
  - This is the rate at which data can be transferred into or out of a memory unit
  - It is equal to 1/cycle time for random access memories
  - For non-random-access memories, it is calculated using the formula:

$T_n = T_A + (n/R)$

where,

Tn = Average time to read or write n bits

$T_A$ = Average access time

n = Number of bits

R = Transfer rate, in bits per second (bps)

# Physical type

- Semiconductor memories
- Magnetic surface (disks & tapes)
- Optical disks
- Magneto-optical disks

# Physical Characteristics

- Volatile
  - Information is lost after the power is off due to leakage
  - Usually refresh units are necessary to keep the information
- Non-volatile
  - No electrical power is needed to keep the information
- Erasable
  - The memory content can be erased using UV rays or by applying electrical pulses
  - The memory can be re-written
- Non-erasable
  - They are only one time written and then used to read

# Characteristics of Memory systems

**Location**
- Internal (e.g., processor registers, cache, main memory)
- External (e.g., optical disks, magnetic disks, tapes)

**Capacity**
- Number of words
- Number of bytes

**Unit of Transfer**
- Word
- Block

**Access Method**
- Sequential
- Direct
- Random
- Associative

**Performance**
- Access time
- Cycle time
- Transfer rate

**Physical Type**
- Semiconductor
- Magnetic
- Optical
- Magneto-optical

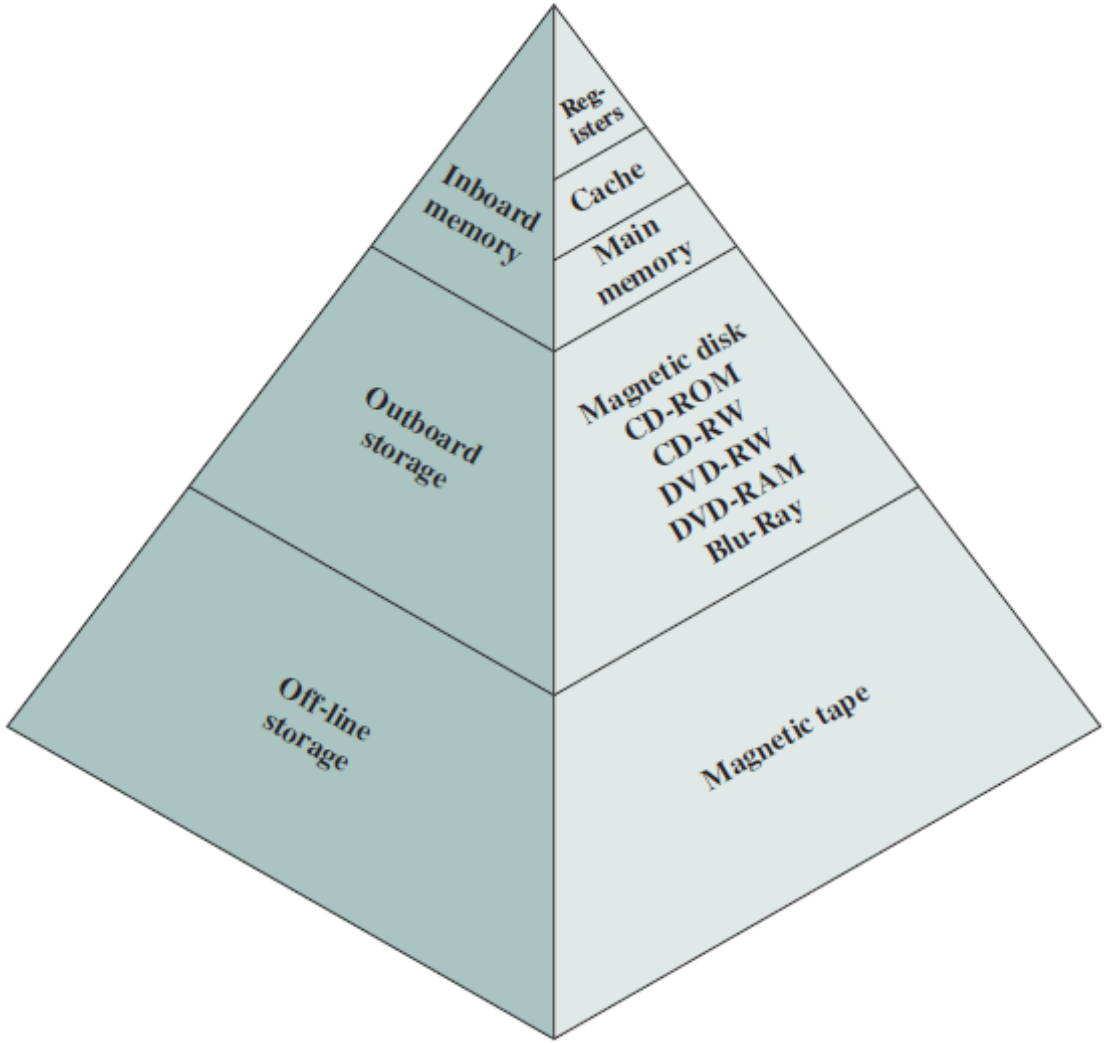
**Physical Characteristics**
- Volatile/nonvolatile
- Erasable/nonerasable

**Organization**
- Memory modules

# Memory Hierarchy

- Design constraints
  - Capacity
  - Performance
  - Cost
- Relationship
  - Faster access time – expensive storage (cost per bit is high)
  - Greater capacity – cheaper storage (cost per bit is low)
  - Greater capacity – low performance (slow access time)

# Memory Hierarchy

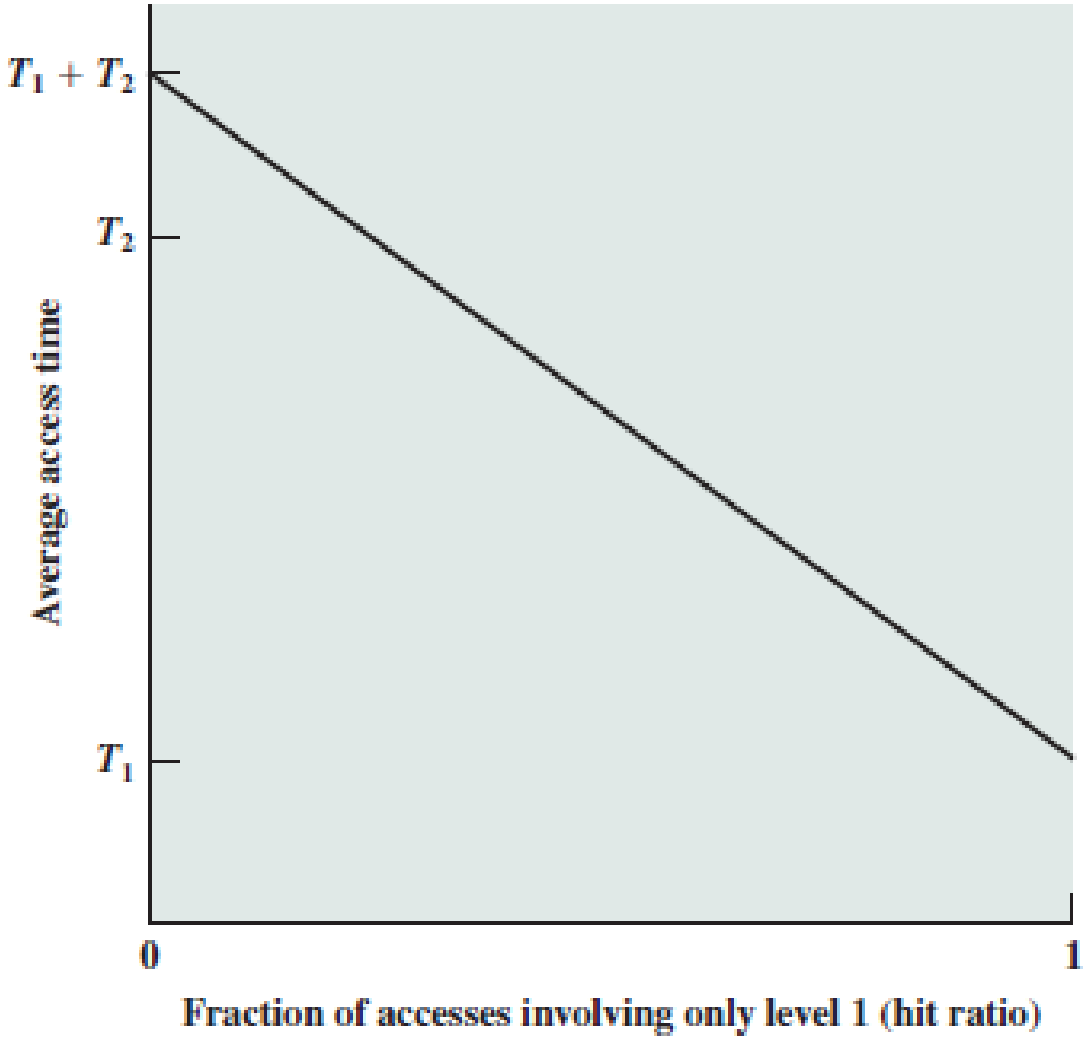



- Hierarchy (top to bottom)
  - Decreasing cost per bit
  - Increasing capacity
  - Increasing access time
  - Decreasing frequency of access of the memory by the processor
    - locality of reference

# A “toy” example

- The processor has access to two levels of memory
- L1 has 1000 words with access time of 0.01 microsecond
- L2 contains 100,000 words with access time of 0.1 microsecond
- If the word is in L1, it is accessed directly by the processor
- If the word is in L2, it is transferred to L1 and is accessed from L1
- Ignore the time to determine whether a word is in L1 or in L2
- Hit (H): Fraction of all memory accesses that are found in L1
- When H=95%, then average time to access memory is
  (0.95*0.01)+(0.05*0.11)=0.0095+0.0055=0.015 microsecond
- The average time is closer to L1 access time than L2 access time

# Performance of access




- T1 is the access time of faster memory
- T2 is the access time of slower memory

*Thank you*