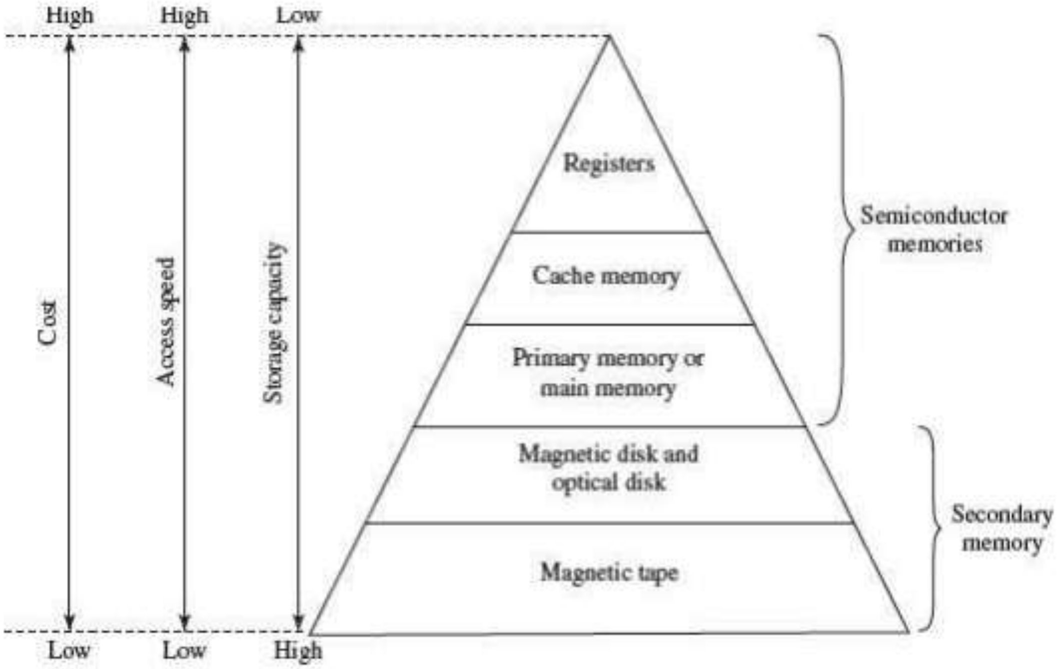
COMP101TH
Problem Solving Using Computer
Unit: 1 (Computer Fundamentals)
Lecture: 5
Memory Hierarchy

- The computer uses a hierarchy of memory that is organized in a manner to enable the fastest speed and largest capacity of memory.
- The memory is characterized on the basis of two key factors—capacity and access time.
    - Capacity is the amount of information (in bits) that a memory can store.
    - Access time is the time interval between the read/ write request and the availability of data.
- The lesser the access time, the faster is the speed of memory.
- Ideally, we want the memory with fastest speed and largest capacity. However, the cost of fast memory is very high.
- **Main objectives of Memory Hierarchy:**
    - 1. To minimize execution time of executing programs
    - 2. To maximize the throughput of the computer
    - 3. To minimize response time

      subject to the constraint that high-speed memory is limited in size.
- The Memory Hierarchy is based on the idea that the faster the memory, the more costly to build and therefore the smaller in capacity, and conversely, the larger the memory, the slower to access and less costly
- **Two Principles of Locality:**
    - Temporal Locality: If an item is referenced, it will tend to be referenced again in the near future.
    - Spatial Locality: If an item is referenced, items whose addresses are close will tend to be referenced soon.
        - Empirical and theoretical justification:
            - programs tend to stay in loops, so instructions and their data are repeatedly accessed (temporal)
            - instructions tend to be executed sequentially (spatial)

  - data tends to be accessed sequentially, as in array accesses (spatial)

- **Memory Hierarchy:**
  - The memory hierarchy is an enhancement to organize the memory such that it can minimize the access time. The memory hierarchy was developed based on a program behavior known as locality of references.
  - Fig. below demonstrates the different levels of memory hierarchy:



  - The five hierarchies in the memory are registers, cache, main memory, magnetic discs and magnetic tapes.
  - The first three hierarchies are volatile memories which mean when there is no power, they will automatically lose their stored data. While the last two hierarchies are non-volatile i.e. they store the data permanently**.**

- **The internal memory and external memory are the two broad categories of memory used in the computer.**
  - The internal memory consists of the CPU registers, cache memory and primary memory. The internal memory is used by the CPU to perform the computing tasks.
  - The external memory is also called the secondary memory. The secondary memory is used to store the large amount of data and the software.

- **Internal Memory—The key features of internal memory are—**
    - limited storage capacity,
    - temporary storage,
    - fast access, and
    - high cost.
- **Registers, cache memory, and primary memory constitute the internal memory.**