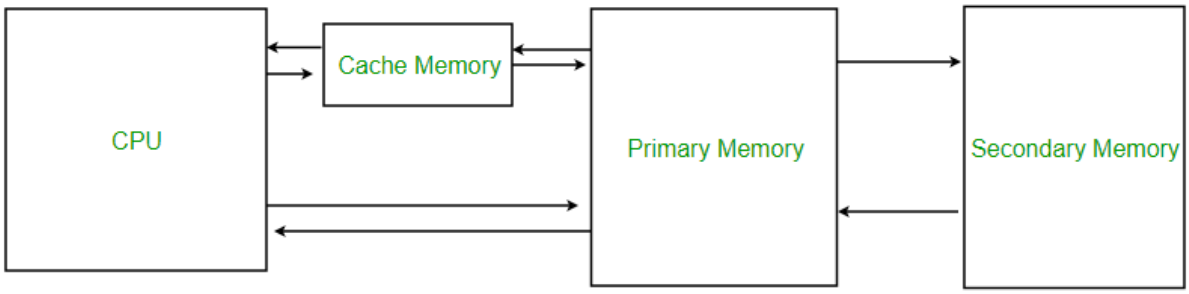
COMP101TH
Problem Solving Using Computer
Unit: 1 (Computer Fundamentals)
Lecture: 6
Registers and Cache

- **Registers:**
  - Registers are high-speed storage areas within the CPU, but have the least storage capacity. Registers are not referenced by their address, but are directly accessed and manipulated by the CPU during instruction execution.
  - Registers store data, instructions, addresses and intermediate results of processing. Registers are often referred to as the CPU's working memory.
  - The data and instructions that require processing must be brought in the registers of CPU before they can be processed.
  - For example, if two numbers are to be added, both numbers are brought in the registers, added and the result is also placed in a register.
  - Registers are the top of the memory hierarchy and are the fastest way for the system to manipulate data.
  - Registers are normally measured by the number of bits they can hold, for example, an 8-bit register means it can store 8 bits of data or a 32-bit register means it can store 32 bit of data.
  - 32-bit processor and 64-bit processor are the terms used to refer to the size of the registers.
  - Following are some commonly used registers:
    - **Accumulator:** This is the most common register, **used to store data taken out from the memory.**
    - **General Purpose Registers**: This is used to store data, intermediate results during program execution. It can be accessed via assembly programming.
    - **Special Purpose Registers**: Users do not access these registers. These registers are for Computer system,
      - **MAR: Memory Address Register** are those registers that holds the address for memory unit.
    - **MBR: Memory Buffer Register** stores instruction and data received from the memory and sent from the memory.
    - **PC: Program Counter** points to the next instruction to be executed.
    - **IR: Instruction Register** holds the instruction to be executed.

- **Cache Memory:**
  - A Cache is a **small and very fast temporary storage memory**. It is designed to speed up the transfer of data and instructions. It is located inside or close to the CPU chip. It is faster than RAM and the data/instructions that are most recently or most frequently used by CPU are stored in cache.
  - The data and instructions are **retrieved from RAM when CPU uses them for the first time**. A copy of that data or instructions is **stored in cache**. The next time the CPU needs that data or instructions, it first looks in cache. If the required data is found there, it is retrieved from cache memory instead of main memory. It **speeds up the working of CPU.**
  - Cache memory is costlier than main memory or disk memory but economical than CPU registers. Cache memory is an extremely fast memory type that acts as a **buffer between RAM and the CPU**. It holds frequently requested data and instructions so that they are immediately available to the CPU when needed.
  - Cache memory is used to **reduce the average time to access data from the Main memory.**



- **Purpose of Cache:**
  - The purpose of cache memory is **to store program instructions and data that are used repeatedly** in the operation of programs or information that the **CPU is likely to need next**.
  - The computer **processor can access this information quickly from the cache** rather than having to get it from computer's main memory.
  - Fast access to these instructions **increases the overall speed of the program.**
  - As the microprocessor processes data, it looks first in the cache memory. If it finds the instructions or data it's looking for there from a previous reading of data, it **does not have to perform a more time-consuming reading of data from larger** main memory or other data storage devices.
  - Cache memory is **responsible for speeding up computer operations and processing.**

- **How Cache Works?**
  - The ability of cache memory to improve a computer's performance relies on the **concept of locality of reference.** Locality describes various situations that make a **system more predictable**, such as where the same storage location is repeatedly accessed**, creating a pattern of memory access that the cache memory relies upon.**
  - There are several types of locality. **Two key ones for cache are temporal and spatial**.
  - Temporal locality is when the same resources are accessed repeatedly in a short amount of time.
  - Spatial locality refers to accessing **various data or resources that are in close proximity to each other**.
- The following are the different levels of Cache Memory.
  - **Level 1 (L1) Cache**
    - It is also called **primary or internal cache.** It is built directly into the processor chip. It has small capacity from 8 kB to 128 kB.
  - **Level 2 (L2) Cache**
    - It is **slower than L1 cache**. The current processors contain advanced transfer cache on processor chip that is a type of L2 cache. The common size of this cache is from 512 kB to 8 MB.
  - **Level 3 (L3) Cache**
    - It exists on the computer that uses L2 advanced transfer cache. It is slower than L1 and L2 cache. The personal computer often has up to 8 MB of L3 cache.