



Logical Organization of Computer

Memory Devices

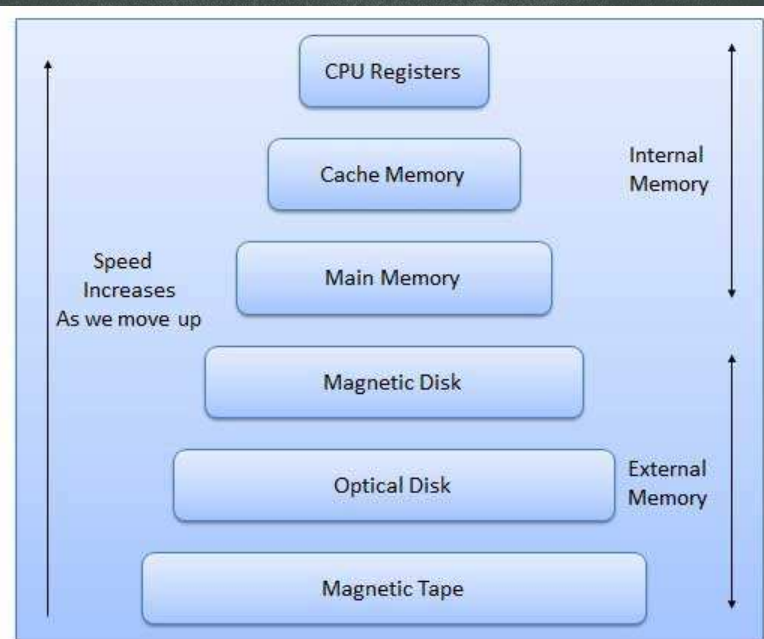
Developed By:
Vivek Vyas
Department of MCA
DDU

□ Introduction

- A memory is used to store data and instruction. Computer memory is the storage space in computer where data is to be processed and instructions required for processing are stored.
- The memory is divided into large number of small parts. Each part is called a cell. Each location or cell has a unique address which varies from zero to memory size minus one.
- For example if computer has 64k words, then this memory unit has $64 * 1024 = 65536$ memory location. The address of these locations varies from 0 to 65535.

❑ Introduction

- Memory is primarily of two types
- **Internal Memory** – cache memory and primary/main memory
- **External Memory** – magnetic disk / optical disk etc



❑ Introduction

- **Characteristics of Memory Hierarchy are following when we go from top to bottom.**
- Capacity in terms of storage increases.
- Cost per bit of storage decreases.
- Frequency of access of the memory by the CPU decreases.
- Access time by the CPU increases.

❑ RAM

- A RAM constitutes the internal memory of the CPU for storing data, program and program result. It is read/write memory. It is called random access memory (RAM).
- Since access time in RAM is independent of the address to the word that is, each storage location inside the memory is as easy to reach as other location & takes the same amount of time.
- We can reach into the memory at random & extremely fast but can also be quite expensive.

❑ RAM

- RAM is volatile, i.e. data stored in it is lost when we switch off the computer or if there is a power failure.
- Hence, a backup uninterruptible power system (UPS) is often used with computers. RAM is small, both in terms of its physical size and in the amount of data it can hold.
- RAM is of two types
 - Static RAM (SRAM)
 - Dynamic RAM (DRAM)

❑ RAM

▪ **Static RAM (SRAM)**

- The word **static** indicates that the memory retains its contents as long as power remains applied. However, data is lost when the power gets down due to volatile nature.
- SRAM chips use a matrix of 6-transistors and no capacitors. Transistors do not require power to prevent leakage, so SRAM need not have to be refreshed on a regular basis.
- Because of the extra space in the matrix, SRAM uses more chips than DRAM for the same amount of storage space, thus making the manufacturing costs higher.
- Static RAM is used as cache memory needs to be very fast and small.

❑ RAM

▪ **Dynamic RAM (DRAM)**

- DRAM, unlike SRAM, must be continually **refreshed** in order for it to maintain the data.
- This is done by placing the memory on a refresh circuit that rewrites the data several hundred times per second.
- DRAM is used for most system memory because it is cheap and small. All DRAMs are made up of memory cells. These cells are composed of one capacitor and one transistor.

□ ROM

- ROM stands for Read Only Memory. The memory from which we can only read but cannot write on it. This type of memory is non-volatile. The information is stored permanently in such memories during manufacture.
- A ROM, stores such instruction as are required to start computer when electricity is first turned on, this operation is referred to as bootstrap.

□ ROM

- **MROM (Masked ROM)**
 - The very first ROMs were hard-wired devices that contained a pre-programmed set of data or instructions. These kind of ROMs are known as masked ROMs. It is inexpensive ROM.
- **PROM (Programmable Read Only Memory)**
 - PROM is read-only memory that can be modified only once by a user. The user buys a blank PROM and enters the desired contents using a PROM programmer.
 - It can be programmed only once and is not erasable.

❑ ROM

- **EPROM (Erasable and Programmable Read Only Memory)**

- The EPROM can be erased by exposing it to ultra-violet light for a duration of upto 40 minutes.

- **EEPROM (Electrically Erasable and Programmable Read Only Memory)**

- The EEPROM is programmed and erased electrically. It can be erased and reprogrammed about ten thousand times.
- Both erasing and programming take about 4 to 10 ms (millisecond). In EEPROM, any location can be selectively erased and programmed.

❑ Serial Access Memory

- Sequential access means the system must search the storage device from the beginning of the memory address until it finds the required piece of data.
- Memory device which supports such access is called a Sequential Access Memory or Serial Access Memory. Magnetic tape is an example of serial access memory.

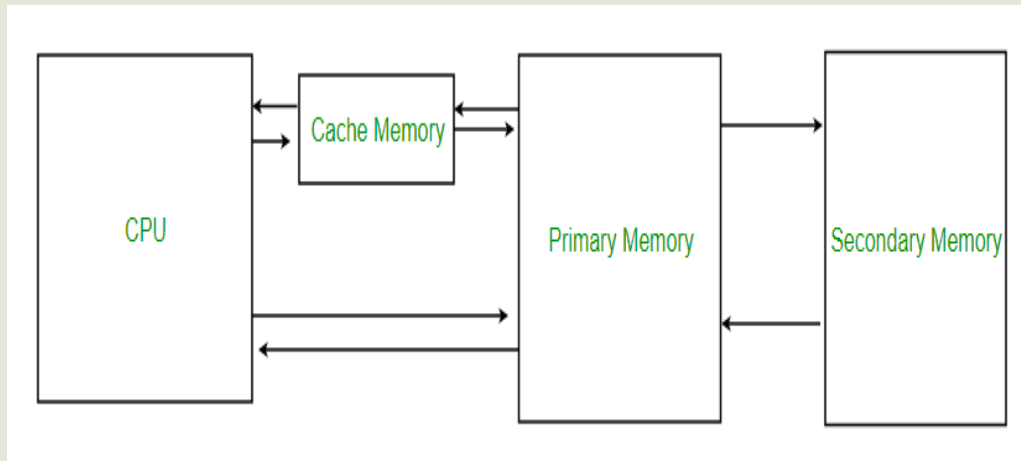
❑ Direct Access Memory

- Direct access memory or Random Access Memory, refers to conditions in which a system can go directly to the information that the user wants.
- Memory device which supports such access is called a Direct Access Memory. Magnetic disks, optical disks are examples of direct access memory.

❑ Cache Memory

- As CPU has to fetch instruction from main memory speed of CPU depending on fetching speed from main memory. CPU contains register which has fastest access but they are limited in number as well as costly.
- Cache is cheaper so we can access cache. Cache memory is a very high speed memory that is placed between the CPU and main memory, to operate at the speed of the CPU.
- It is used to reduce the average time to access data from the main memory. The cache is a smaller and faster memory which stores copies of the data from frequently used main memory locations.
- Most CPUs have different independent caches, including instruction and data.

❑ Cache Memory



❑ Cache Memory

▪ Cache Performance

- When the processor needs to read or write a location in main memory, it first checks for a corresponding entry in the cache.
 - If the processor finds that the memory location is in the cache, a **cache hit** has occurred and data is read from cache
 - If the processor **does not** find the memory location in the cache, a **cache miss** has occurred. For a cache miss, the cache allocates a new entry and copies in data from main memory, then the request is fulfilled from the contents of the cache.

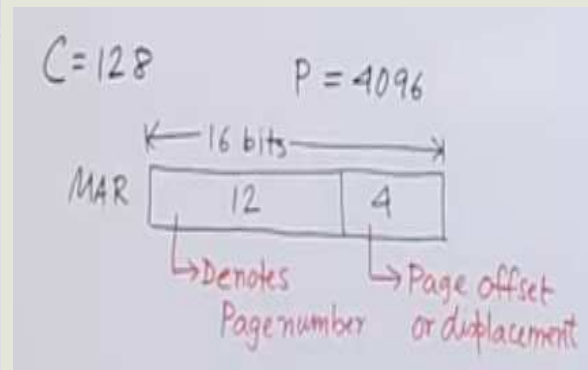
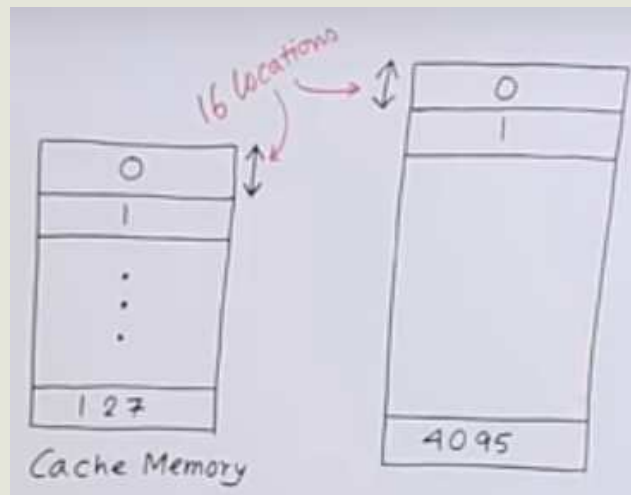
❑ Cache Memory

- The performance of cache memory is frequently measured in terms of a quantity called Hit ratio.
 - $\text{Hit ratio} = \text{hit} / (\text{hit} + \text{miss}) = \text{no. of hits} / \text{total accesses}$

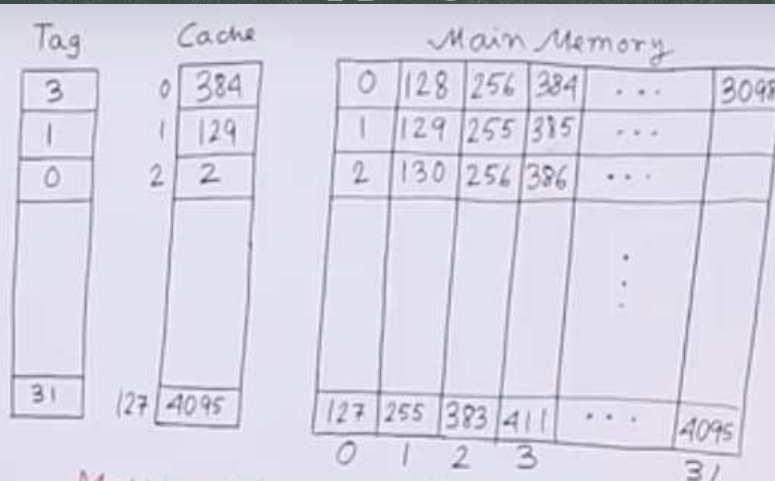
❑ Cache Mapping

- The three different types of mapping used for the purpose of cache memory are as follow,
 - Direct mapping
 - Associative mapping,
 - Set-Associative mapping.

□ Direct Mapping



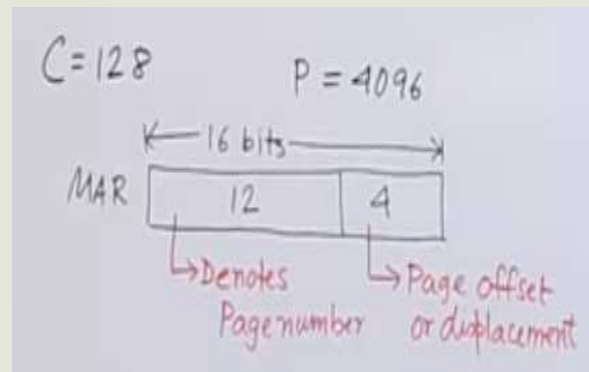
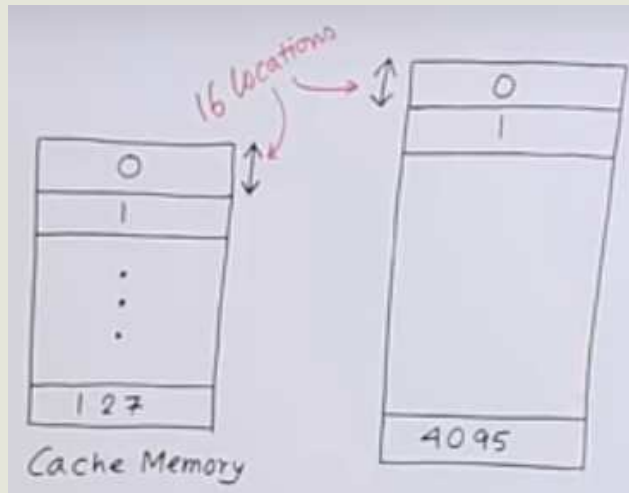
□ Cache Mapping



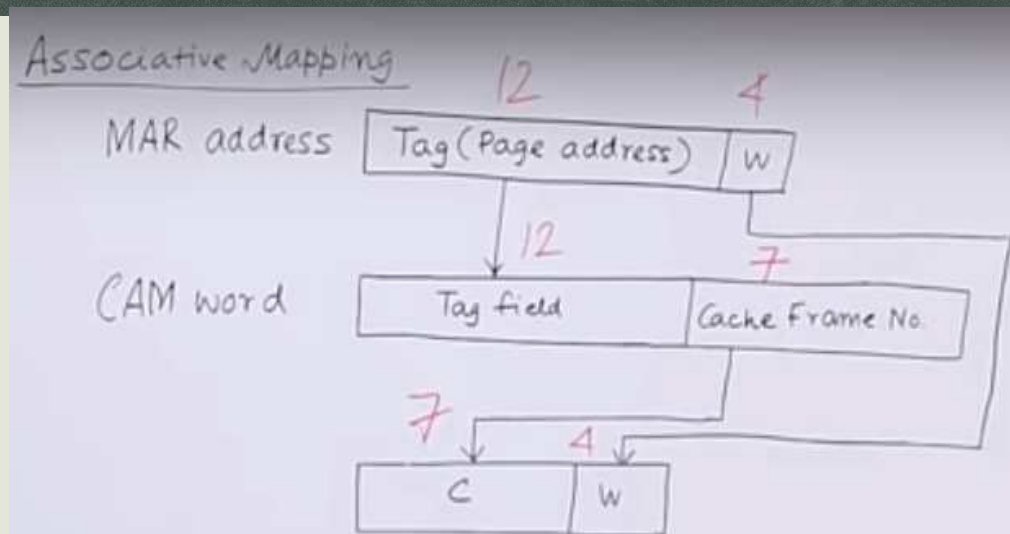
Mapping main memory blocks to cache block

$$\frac{4096}{128} = \frac{2^{12}}{2^7} = 2^5 = 32 \text{ (0 to 31)}$$

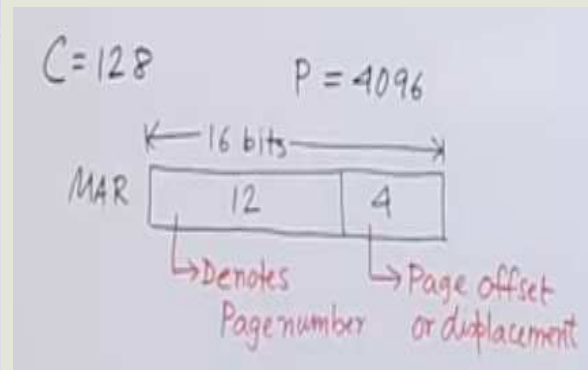
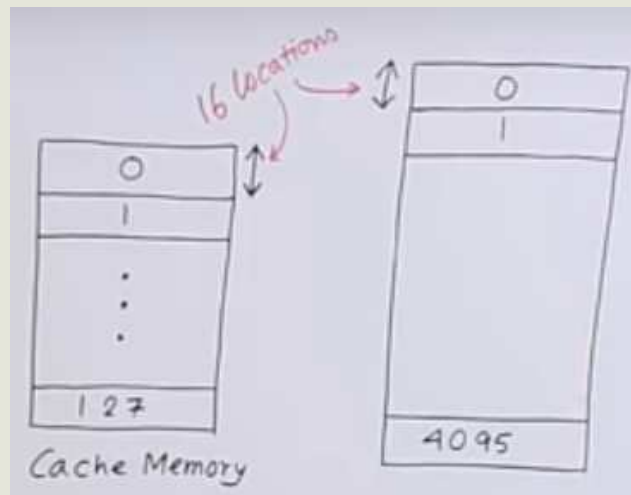
❑ Associative Mapping



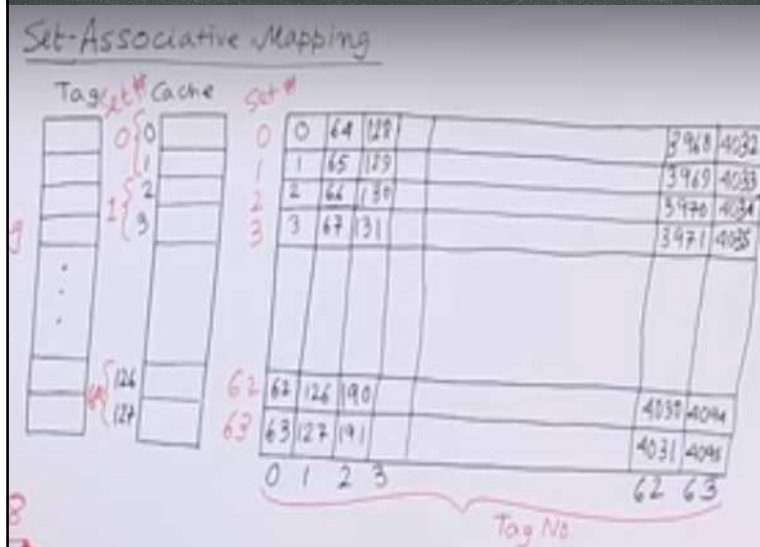
❑ Associative Mapping



□ Set Associative Mapping

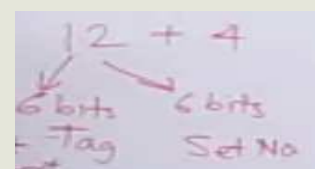


□ Set Associative Mapping



$$\frac{128}{2} = 64 \text{ Sets (6 bits)}$$

$$\frac{4096}{64} = \frac{2^{12}}{2^6} = 2^6$$



❑ Set Associative Mapping

Handwritten calculation for Set Associative Mapping:

$$64 \overline{) 4033} \begin{array}{r} 63 \\ \underline{384} \\ 193 \\ \underline{192} \\ 1 \end{array}$$

The result shows a quotient of 63 and a remainder of 1. The quotient 63 is labeled as the Tag No. and the remainder 1 is labeled as the Set No.

Binary representation of the Tag No. (63): 111111_2

Binary representation of the Set No. (1): 000001_2

The binary representation of the Set No. is further divided into two parts: TN (Tag Number) and SN (Set Number).

❑ Locality of reference

- Since size of cache memory is less as compared to main memory. So to check which part of main memory should be given priority and loaded in cache is decided based on locality of reference.
- **1. Spatial Locality of reference:** If the storage has been accessed then likelihood of accessing the storage nearby that, is high.
- **2. Temporal Locality of reference:** It tells us whether memory locations in a program are likely to be accessed again in the near future. A method has high temporal locality if it is called repeatedly in a short period of time.

❑ Virtual Memory

- Virtual Memory is a storage scheme that provides user an illusion of having a very big main memory. This is done by treating a part of secondary memory as the main memory.
- In this scheme, User can load the bigger size processes than the available main memory by having the illusion that the memory is available to load the process.
- Instead of loading one big process in the main memory, the Operating System loads the different parts of more than one process in the main memory.
- By doing this, the degree of multiprogramming will be increased and therefore, the CPU utilization will also be increased.

❑ Virtual Memory

▪ How Virtual Memory Works?

- In this scheme, whenever some pages needs to be loaded in the main memory for the execution and the memory is not available for those many pages, then in that case, instead of stopping the pages from entering in the main memory, the OS search for the RAM area that are least used in the recent times or that are not referenced and copy that into the secondary memory to make the space for the new pages in the main memory.

❑ Virtual Memory

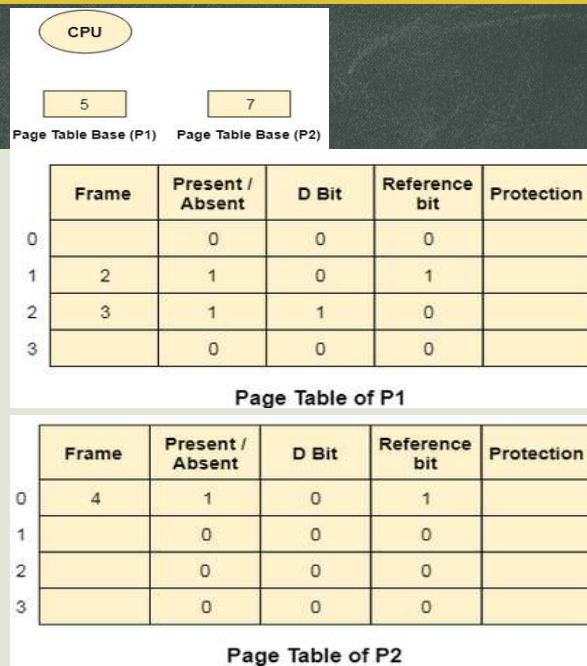
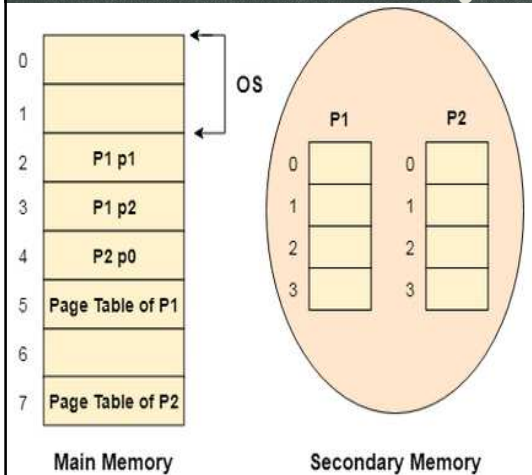
▪ Virtual Memory Management System

- Let us assume 2 processes, P1 and P2, contains 4 pages each. Each page size is 1 KB. The main memory contains 8 frame of 1 KB each. The OS resides in the first two partitions.
- In the third partition, 1st page of P1 is stored and the other frames are also shown as filled with the different pages of processes in the main memory.
- The page tables of both the pages are 1 KB size each and therefore they can be fit in one frame each. The page tables of both the processes contain various information that is also shown in the image.

❑ Virtual Memory

- The CPU contains a register which contains the base address of page table that is 5 in the case of P1 and 7 in the case of P2.
- This page table base address will be added to the page number of the Logical address when it comes to accessing the actual corresponding entry.

❑ Virtual Memory



❑ Auxiliary Memory

- Auxiliary memory is much larger in size than main memory but is slower. It normally stores system programs, instruction and data files. It is also known as secondary memory.
- Secondary memories cannot be accessed directly by a processor. First the data/information of auxiliary memory is transferred to the main memory and then that information can be accessed by the CPU.

□ Auxiliary Memory

▪ Characteristics of Auxiliary Memory are following.

- **Non-volatile memory** – Data is not lost when power is cut off.
- **Reusable** – The data stays in the secondary storage on permanent basis until it is not overwritten or deleted by the user.
- **Reliable** – Data in secondary storage is safe because of high physical stability of secondary storage device.
- **Convenience** – With the help of a computer software, authorized people can locate and access the data quickly.
- **Capacity** – Secondary storage can store large volumes of data in sets of multiple disks.
- **Cost** – It is much lesser expensive to store data on a tape or disk than primary memory.