


Module 5

Memory Organization

Introduction

- Memory is a computer chip or device that is used to hold data and instructions which are used by computer during the processing.
 - These programs and data needs to be transferred between CPU and memory.
 - Faster the transfer rate, better is the performance of computer system.
- 

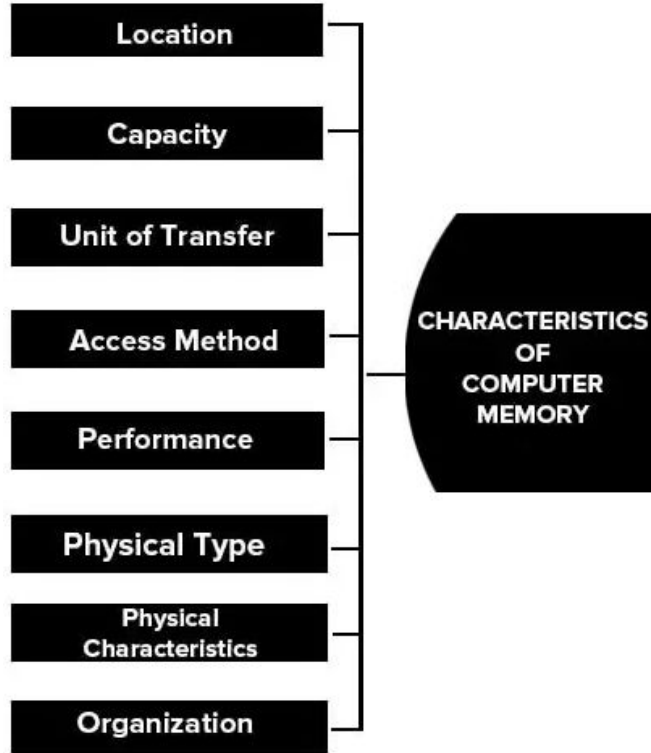
Memory parameters

Given below are the important parameters in choosing computer memory:

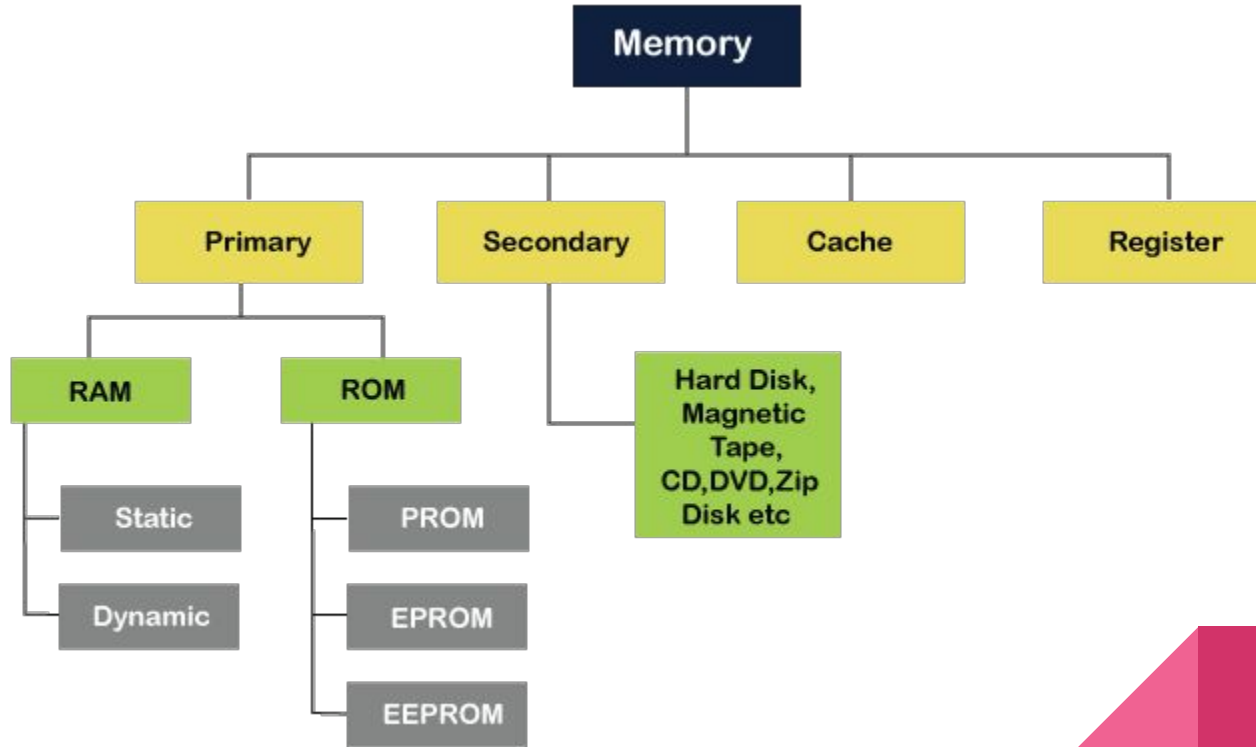
- Capacity
- Bandwidth or data transfer rate
- Speed



Memory characteristic



Classification of Memory



Primary Memory or main memory

- It directly communicates with the CPU of the computer.
- Main memory contains only those programs, data or information that is required by CPU for current execution.
- It occupies the central position in memory hierarchy as it is able to communicate directly to both CPU and secondary memory through I/O processor.
- It is further divided into two parts:
 1. RAM (Random Access Memory)
 2. ROM (Read Only Memory)



RAM

- We can read from it and also can write there. So this memory is read and writes memory.
- This memory is also known as Volatile Memory.
- When the computer is switched off than the entry is vanished.
- There are two main types of RAM:
 - a. Static RAM
 - b. Dynamic RAM.




Static RAM

- It is a type of RAM used to store static data in the memory.
- It means to store data in SRAM remains active as long as the computer system has a power supply.
- However, data is lost in SRAM when power failures have occurred.

Characteristics:


1. It does not require to refresh.
2. It is faster than DRAM
3. It is expensive.
4. High power consumption
5. Longer life
6. Large size
7. Uses as a cache memory

Dynamic RAM

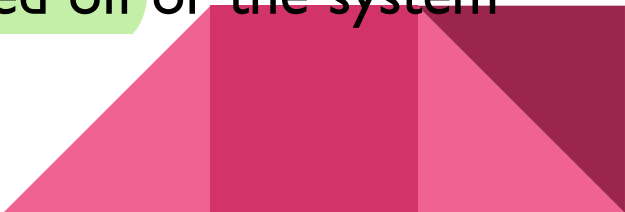
- It is a type of RAM that is used for the dynamic storage of data.
 - In DRAM, each cell carries one-bit information. The cell is made up of two parts: **a capacitor and a transistor.**
 - The size of the capacitor and the transistor is so small, requiring millions of them to store on a single chip.
 - Hence, a DRAM chip can hold more data than an SRAM chip of the same size.
 - DRAM is volatile. Thus, If the power is switched off, the data store in memory is lost.
- 

Dynamic RAM

Characteristic:

1. It requires continuously refreshed to retain the data.
 2. It is slower than SRAM
 3. It holds a large amount of data
 4. It is the combination of capacitor and transistor
 5. It is less expensive as compared to SRAM
 6. Less power consumption
- 

ROM

- It is a memory device that is used to permanently store information inside a chip.
 - It is a read-only memory that can only read stored information, data or programs, but we cannot write or modify anything.
 - A ROM contains some important instructions or program data that are required to start or boot a computer.
 - It is a non-volatile memory; it means that the stored information cannot be lost even when the power is turned off or the system is shut down.
- 

Types of ROM


1. **PROM** (Programmable Read Only Memory): In this, user can write any type of information or program only once.
2. **EPROM** (Erasable Programmable Read Only Memory): In this, To erase stored data and re-programmed it, first, pass the ultraviolet light for 40 minutes; after that, the data is re-created.
3. **EEPROM** (Electrically Erasable Programmable Read Only Memory): In this, the stored data can be erased and reprogrammed up to 10 thousand times using a high voltage electrical charge and re-programmed it.

Secondary memory or External memory

- It is used to store data and information just like primary memory.
- They are slow-speed and low-cost memory devices used to provide backup storage.
- Examples:

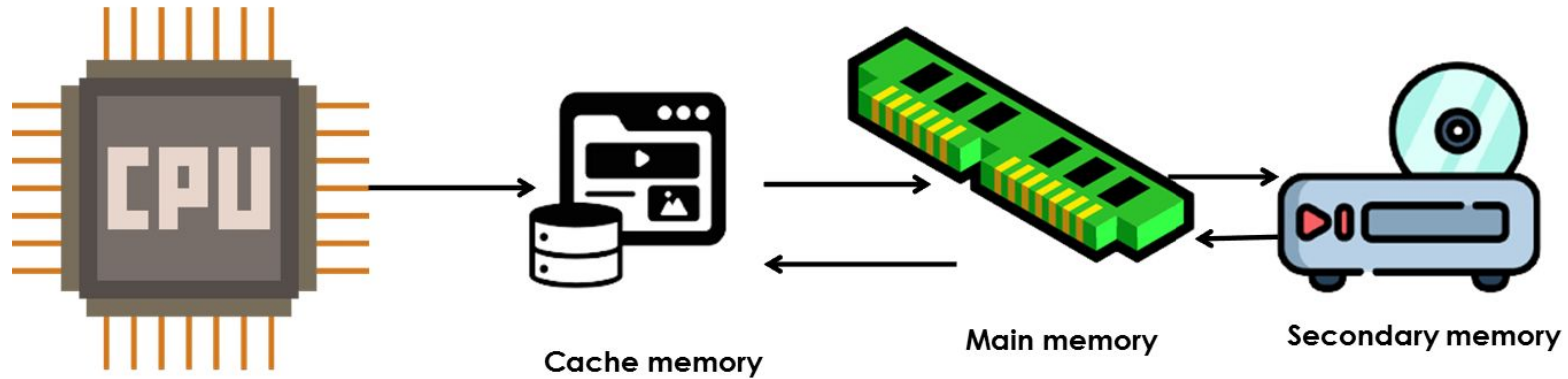


Cache memory

- It is a small-sized chip-based computer memory that lies between the CPU and the main memory.
 - It is a faster, high performance and temporary memory to enhance the performance of the CPU.
 - It stores all the data and instructions that are often used by computer CPUs.
 - It is faster than the main memory, and sometimes, it is also called CPU memory because it is very close to the CPU chip.
- 

Cache memory

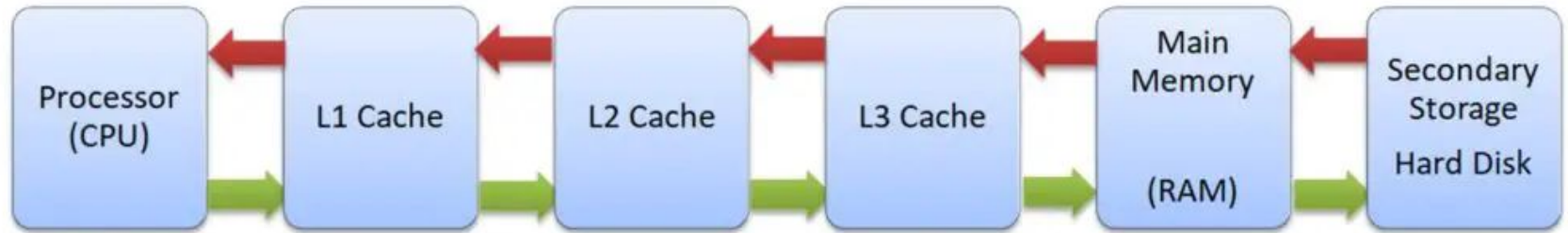
It acts as buffer between CPU and main memory.



Levels of Cache memory

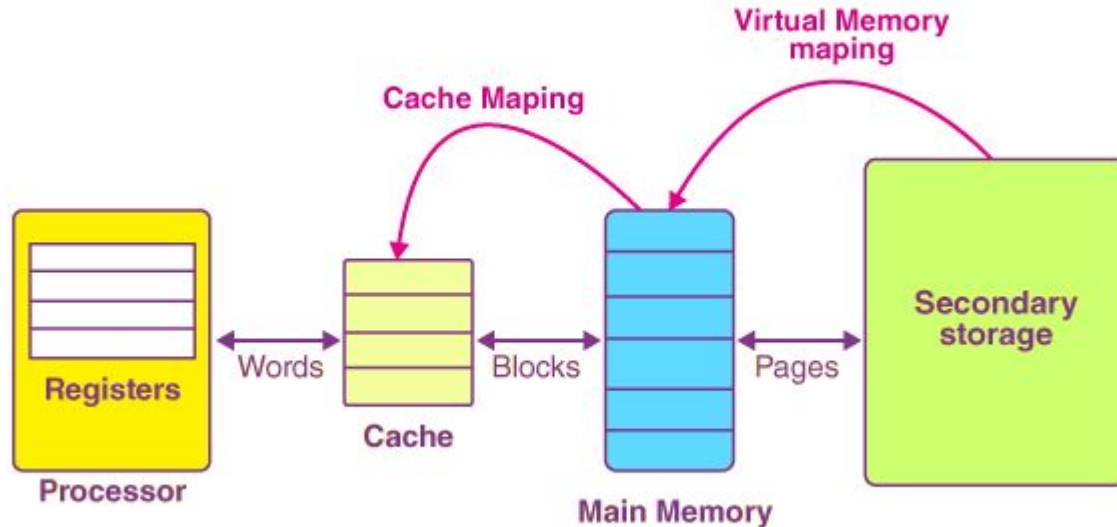
1. **L1 Cache:** It is primary cache. It is built with the help of the CPU. Its speed is very high. Its size varies from 8 KB to 128 KB.
2. **L2 Cache:** It is also known as external or secondary cache, which requires fast access time to store temporary data. It is built into a separate chip in a motherboard, not built into the CPU like the L1 level. Its size may be 128 KB to 1 MB.
3. **L3 Cache:** It is used with high performance and capacity of the computer. It is built into a motherboard. Its speed is very slow, and the maximum size up to 8 MB.

Levels of Cache memory



Cache memory mapping

The process of cache mapping helps us define how a certain block in the main memory gets mapped to the memory of a cache in the case of any cache miss.

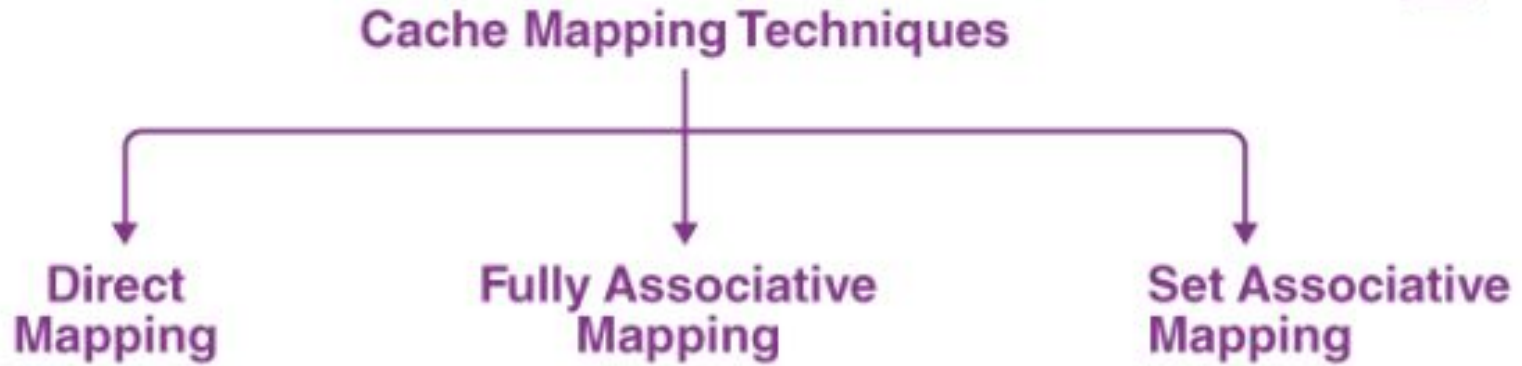


Cache memory mapping

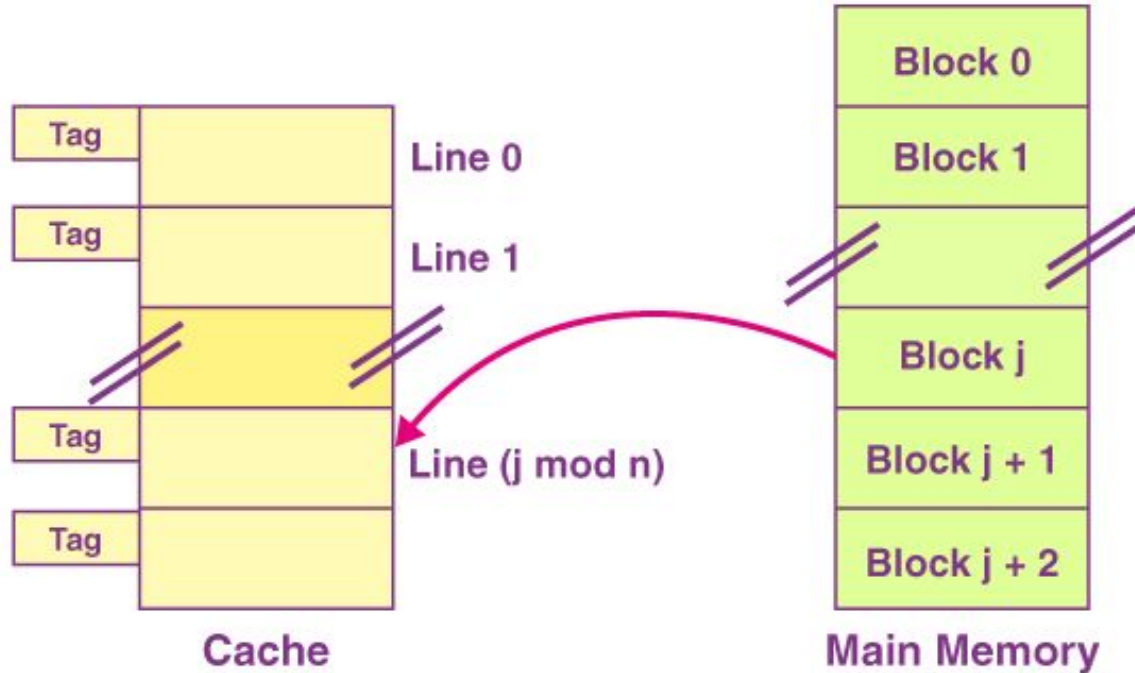
- The main memory gets divided into multiple partitions of equal size, known as the frames or blocks.
- The cache memory is actually divided into various partitions of the same sizes as that of the blocks, known as lines.
- The main memory block is copied simply to the cache during the process of cache mapping, and this block isn't brought at all from the main memory.



Cache memory mapping techniques



I. Direct mapping



I. Direct mapping

In direct mapping,

- A particular block of main memory can map only to a particular line of the cache.
- The line number of cache to which a particular block can map is given by-

Cache line number = (Main Memory Block Address) Modulo
(Number of lines in Cache)



I. Direct mapping

In the case of direct mapping,

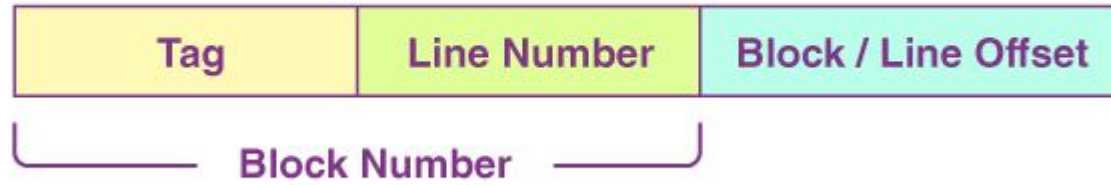
- There is no requirement for a replacement algorithm.
- It is because the block of the main memory would be able to map to a certain line of the cache only.
- Thus, the incoming (new) block always happens to replace the block that already exists, if any, in this certain line.



I. Direct mapping

Division of Physical Address

In the case of direct mapping, the division of the physical address occurs as follows:



Division of Physical Address in Direct Mapping

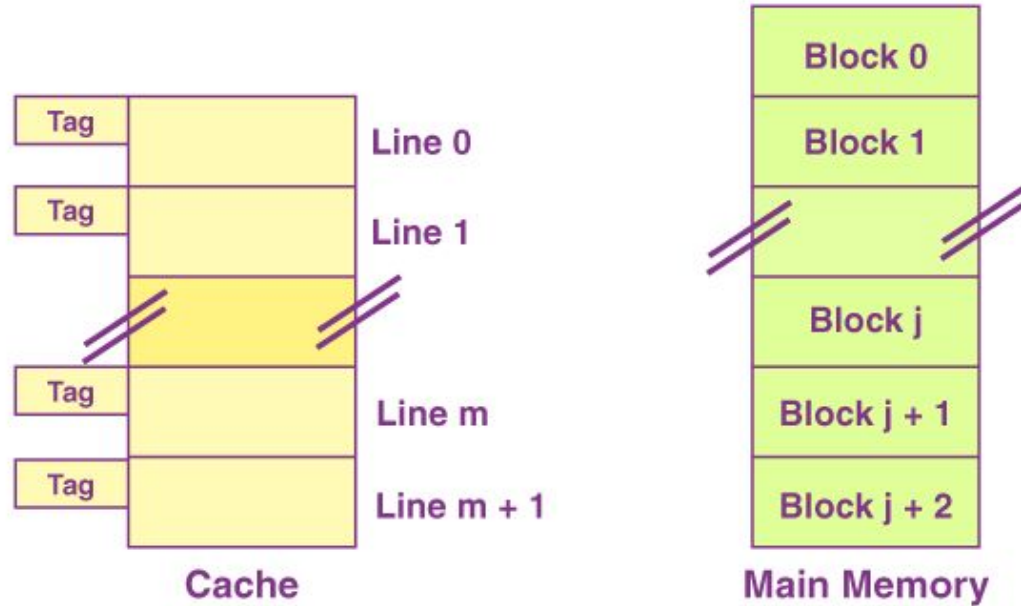
2. Fully associative mapping

In the case of fully associative mapping,

- The main memory block is capable of mapping to any given line of the cache that's available freely at that particular moment.
- It helps us make a fully associative mapping comparatively more flexible than direct mapping.



2. Fully associative mapping



2. Fully associative mapping

In the case of fully associative mapping,

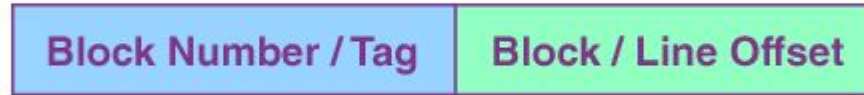
- The replacement algorithm is always required.
- The replacement algorithm suggests a block that is to be replaced whenever all the cache lines happen to be occupied.
- So, replacement algorithms such as LRU Algorithm, FCFS Algorithm, etc., are employed.



2. Fully associative mapping

Division of Physical Address

In the case of fully associative mapping, the division of the physical address occurs as follows:



Division of Physical Address in Fully Associative Mapping

3. Set associative mapping

In the case of k-way set associative mapping,

- The grouping of the cache lines occurs into various sets where all the sets consist of k number of lines.
- Any given main memory block can map only to a particular cache set.
- However, within that very set, the block of memory can map any cache line that is freely available.



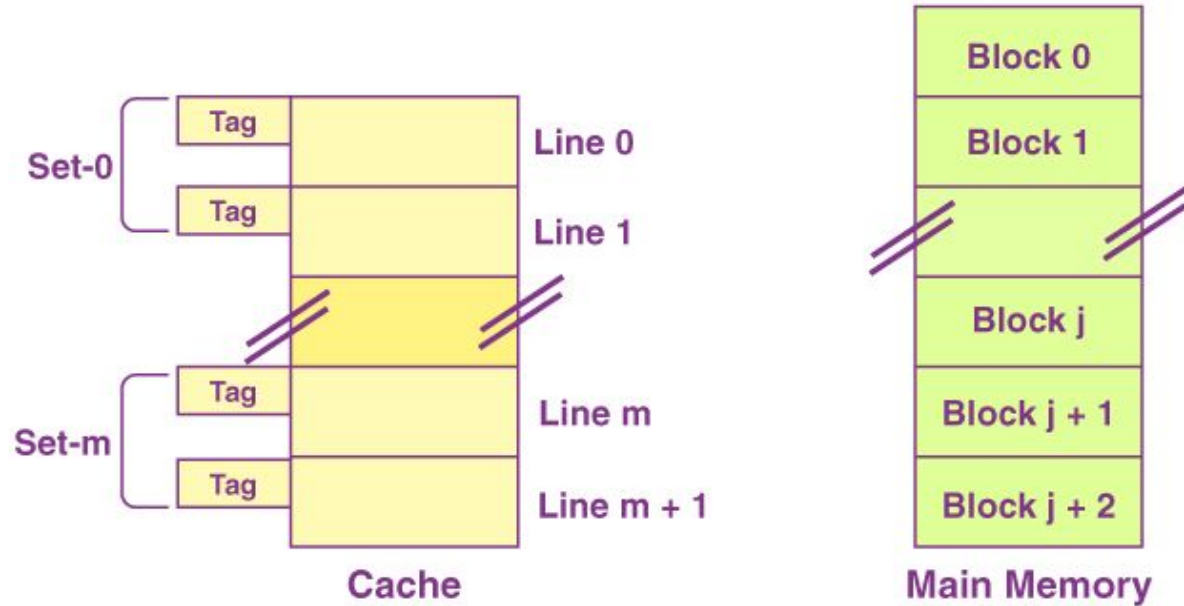
3. Set associative mapping

- The cache set to which a certain main memory block can map is basically given as follows:

Cache set number = (Block Address of the Main Memory) Modulo
(Total Number of sets present in the Cache)



3. Set associative mapping



3. Set associative mapping

In the case of k-way set associative mapping,

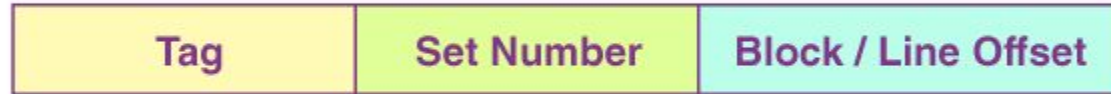
- The k-way set associative mapping refers to a combination of the direct mapping as well as the fully associative mapping.
- It makes use of the fully associative mapping that exists within each set.
- Therefore, the k-way set associative mapping needs a certain type of replacement algorithm.



3. Set associative mapping

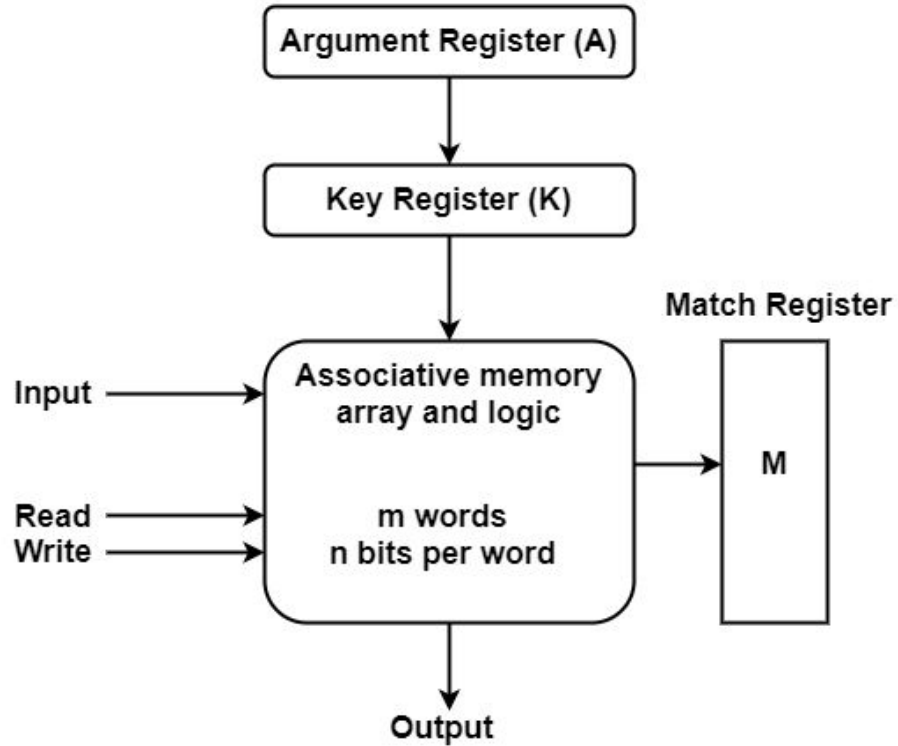
Division of Physical Address

In the case of fully k-way set mapping, the division of the physical address occurs as follows:



Division of Physical Address in K-way Set Associative Mapping

Associative memory



Associative memory

- An associative memory can be treated as a memory unit whose saved information can be recognized for approach by the content of the information itself instead of by an address or memory location.
- Associative memory is also known as Content Addressable Memory (CAM).



Associative memory

- It includes a memory array and logic for m words with n bits per word. The argument register A and key register K each have n bits, one for each bit of a word.
- The match register M has m bits, one for each memory word. Each word in memory is related in parallel with the content of the argument register.



Associative memory


- The words that connect the bits of the argument register set an equivalent bit in the match register. After the matching process, those bits in the match register that have been set denote the fact that their equivalent words have been connected.
- Reading is proficient through sequential access to memory for those words whose equivalent bits in the match register have been set.



Interleaved memory

The memory system in which successive addresses are evenly spread across memory bank to compensate for the relatively slow speed of DRAM.

The contiguous memory reads and writes are using each bank in term, resulting in higher memory throughputs due to reduced waiting for memory banks to become ready for desired operations.



Types of Interleaving

1. Low order Interleaving

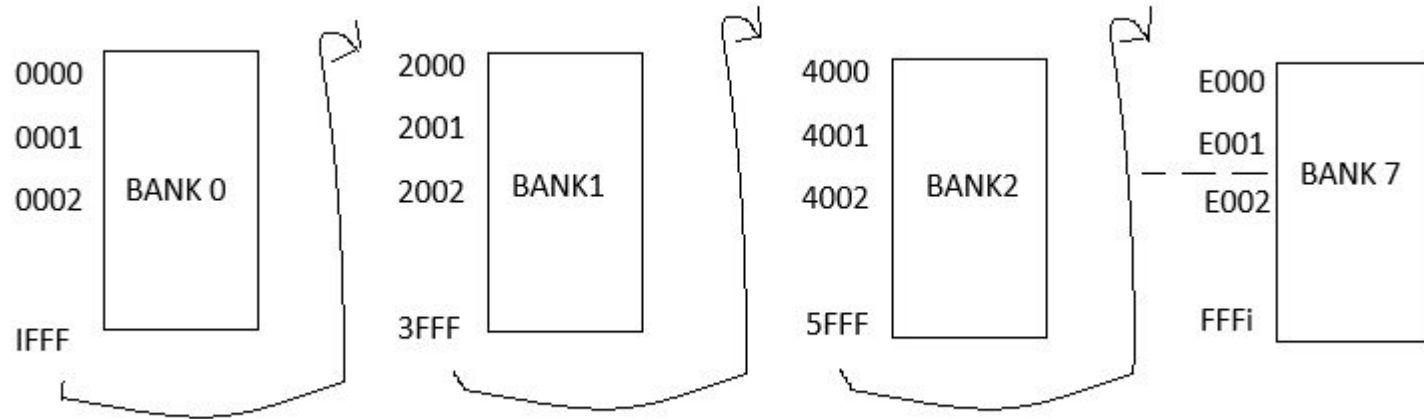
In this lower order address lines eg. A0-A12 used to identify location in each bank whereas higher order address lines are decoded to generate chip select (CS) of individual memory chip.

In this low order interleaving successive address will get allocated in the same memory chip as shown below.



Types of Interleaving

1. Low order Interleaving



Interleaved memory

2. High order Interleaving.

In this case higher order address lines, eg, A3 - A15 used to identify location in each bank where as lower order address lines eg. A0, A1, A2 will be decoded to generate CS 07 individual chip.

This will allocate successive addresses in the different memory bank as shown below.



Interleaved memory

2. High order Interleaving.

