Memory Organization

Module 6

Outline

- Introduction to memory and memory parameters
- Classification of primary and secondary memories
- Types of RAM and ROM
- Allocation policies
- Memory hierarchy and characteristics
- Cache Memory: Concepts, architecture (L1,L2,L3) ,mapping techniques
- Cache coherency
- Interleaved and Associative Memory

Computer Memory

- Memory unit is essential component of digital computer since it is **used for storing programs and data** that are required to perform a specific task.
- For CPU to operate at its maximum speed, it required an uninterrupted and high speed access to these memories that contain programs and data.
- Memory unit that **communicates directly with CPU is called Main memory.**
- Only programs and data currently needed by processor reside in the main memory.

Memory Hierarchy

- Devices that provide **backup storage** is called **auxiliary memory.**
- All other information is stored in auxiliary memory and transferred to main memory when needed.
- Some of the criteria need to be taken into consideration while deciding which memory is to be used:
  - Cost
  - Speed
  - Memory access time
  - Data transfer rate
  - Reliability

Key Characteristics of computer memory system

- Location
- Capacity
- Unit of Transfer
- Access Method
- Performance
- Physical type
- Physical characteristics
- Organization

Location:

- It deals with the location of the memory device in the computer system. There are three possible locations:

1. CPU : This is often in the form of CPU registers and small amount of cache

2. Internal or main: This is the main memory like RAM or ROM. The CPU can directly access the main memory.

3. External or secondary: It comprises of secondary storage devices like hard disks, magnetic tapes. The CPU doesn't access these devices directly. It uses device controllers to access secondary storage devices.

Capacity

- The capacity of any memory device is expressed in terms of:
- **i)word size ii)Number of words**
- **Word size:** Words are expressed in bytes (8 bits). A word can however mean any number of bytes. **Commonly used word sizes are 1 byte (8 bits), 2bytes (16 bits) and 4 bytes (32 bits).**
- **Number of words:** This specifies the number of words available in the particular memory device. For example, if a memory device is given as 4K x 16. This means the device has a word size of 16 bits and a total of 4096(4K) words in memory.

Unit of Transfer:

- It is the maximum number of bits that can be read or written into the memory at a time.
- In case of **main memory**, it is mostly equal to **word size**.
- In case of **external memory, unit of transfer is** not limited to the word size; it is often larger and is referred to as **blocks.**

Access Methods:

- It is a fundamental characteristic of memory devices. It is the sequence or order in which memory can be accessed. There are three types of access methods:
- **Random Access:** If storage locations in a particular memory device can be accessed in any order and access time is independent of the memory location being accessed. Such memory devices are said to have a random access mechanism. RAM (Random Access Memory) IC's use this access method.

Access Methods:

- **Serial Access:** If memory locations can be accessed only in a certain predetermined sequence, this access method is called serial access.
  - Magnetic Tapes, CD-ROMs employ serial access methods.
- **Semi random Access:** Memory devices such as Magnetic Hard disks use this access method. Here each track has a read/write head thus each track can be accessed randomly but access within each track is restricted to a serial access.

Performance:

- The performance of the memory system is determined using three parameters
- **Access Time:** In random access memories, it is the time taken by memory to complete the read/write operation from the instant that an address is sent to the memory. For non-random access memories, it is the time taken to position the read write head at the desired location. Access time is widely used to measure performance of memory devices.
- **Memory cycle time:** It is defined only for Random Access Memories and is **the sum of the access time and the additional time required before the second access can commence**.
- **Transfer rate:** It is defined as **the rate at which data can be transferred into or out of a memory unit.**

Physical type

- Memory devices can be either **semiconductor memory** (like RAM) or **magnetic surface memory** (like Hard disks).

Organization:

Erasable/Non-erasable:

- The memories in which data once programmed cannot be erased are called **Non-erasable memories.**
- Memory devices in which data in the memory can be erased is called **erasable memory**.
- E.g. RAM(erasable), ROM(non-erasable).

Memory Parameters

- choosing a suitable memory becomes necessary for improved performance.

- **Parameter in choosing memory**
  - Capacity
  - Bandwidth
  - Speed

Memory Parameters

Capacity

- The size of computer depends on its memory capacity.
- Memory can be seen as a storage unit containing x number of locations, each of which stores y number of bits.
- The total capacity of **memory can be calculated as x*y-bit or x-word memory.**

Bandwidth

- Bandwidth of the memory indicates the maximum amount of information that can be transferred to or from the memory per unit time.
- It is expressed as number of bytes or words per second.

Memory Parameters

Speed

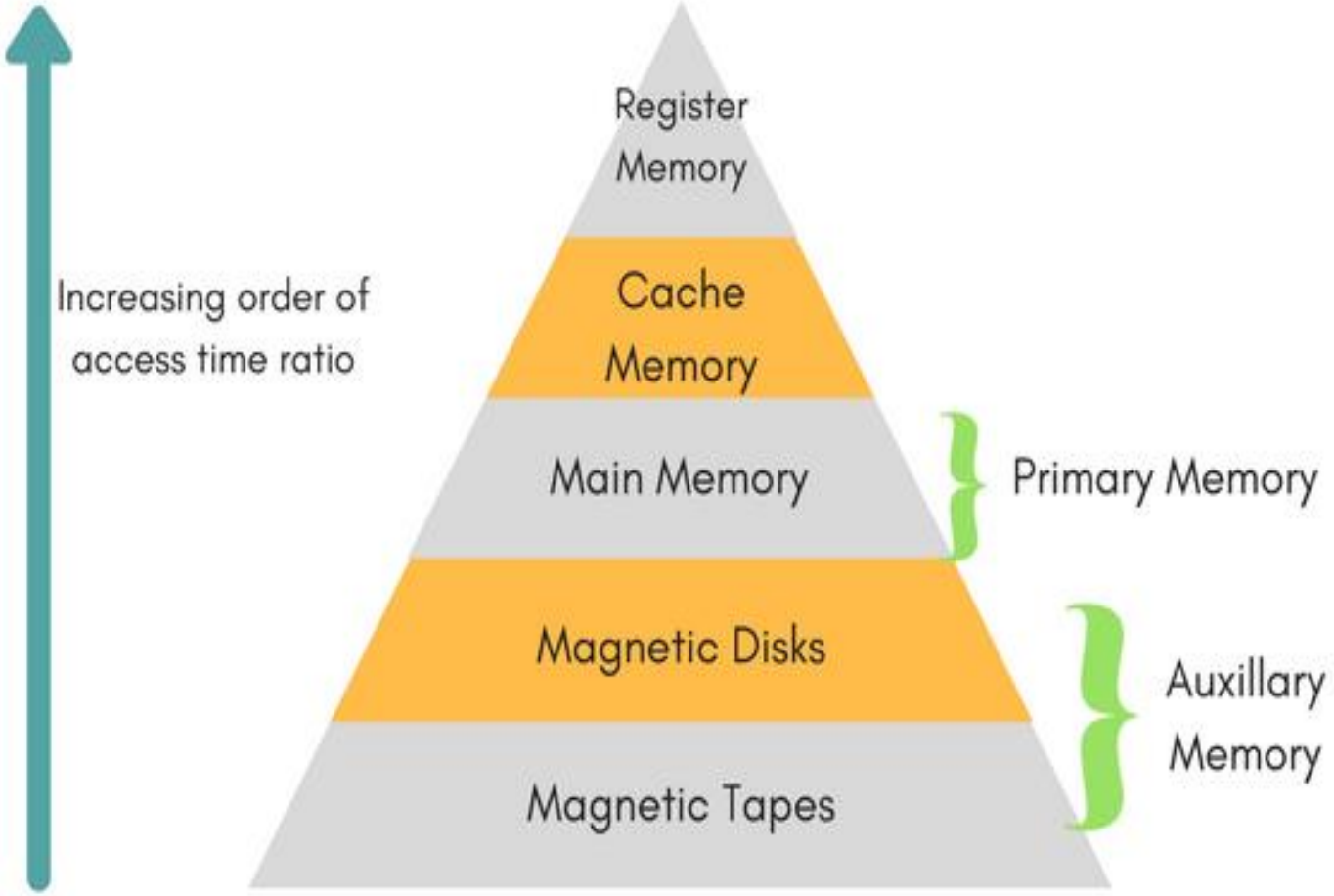
- The speed of operation of the memory is very important parameter.
- The speed simply indicates the time between start of an operation and end of that operation.
- Speed of memory is measured in two parameters:
  - access time (ta)
  - cycle time (tc)
- All the three parameters capacity, bandwidth and speed needs to be considered while choosing a memory while designing computer architecture.

Memory Hierarchy

A memory unit can be classified broadly into two categories:

- The memory unit that establishes direct communication with the CPU is called **Main Memory**. The main memory is often referred to as RAM (Random Access Memory).
- The memory units that provide backup storage are called **Auxiliary Memory**. For example, magnetic disks and magnetic tapes are the most commonly used auxiliary memories.
- Apart from the basic classifications of a memory unit, the memory hierarchy consists all of the storage devices available in a computer system ranging from the slow but high-capacity auxiliary memory to relatively faster main memory.

Memory Hierarchy



Memory Hierarchy

- Memory hierarchy system consist of all storage devices from auxiliary memory to main memory to cache memory
- As one goes down the hierarchy :
- Cost per bit decreases
- Capacity increases
- Access time increases
- Frequency of access by the processor decreases.

Memory Hierarchy

- **Auxiliary memory is having** access time is generally **1000 times** that of the main memory, hence it is at the bottom of the hierarchy.
- The **main memory** occupies the central position because it is equipped to communicate directly with the CPU and with auxiliary memory devices through Input/output processor (I/O).
- When the program not residing in main memory is needed by the CPU, they are brought in from auxiliary memory.
- Programs not currently needed in main memory are transferred into auxiliary memory to provide space in main memory for other programs that are currently in use.
- The **cache memory** **is used to store program data which is currently being executed in the CPU**. Approximate access time ratio between cache memory and main memory is about **1 to 7~10**

Memory Classification

Memory is primarily of three types –

- Cache Memory
- Primary Memory/Main Memory
- Secondary Memory

Main Memory

- It is the memory used to store programs and data during the computer operation.
- The principal technology is based on semiconductor integrated circuits.
- It consists of **RAM and ROM chips.**
- RAM chips are available in two form static and dynamic.

Main Memory

Primary Memory (Main Memory)

- Primary memory holds only those data and instructions on which the computer is currently working.
- It has a limited capacity
- Data is lost when power is switched off.
- Made up of semiconductor device.
- These memories are not as fast as registers.
- The data and instruction required to be processed resides in the main memory.
- It is divided into two subcategories **RAM and ROM.**

Primary Memory (Main Memory)

Characteristics of Main Memory

- These are semiconductor memories.
- Usually volatile memory.
- Data is lost in case power is switched off.
- It is the **working memory of the computer.**
- Faster than secondary memories.
- **A computer cannot run a program without the primary memory**

Main Memory

- The primary technology used for the main memory is based on semiconductor integrated circuits.
- The integrated circuits for the main memory are classified into two major units.
- RAM (Random Access Memory) integrated circuit chips
- ROM (Read Only Memory) integrated circuit chips

RAM integrated circuit chips

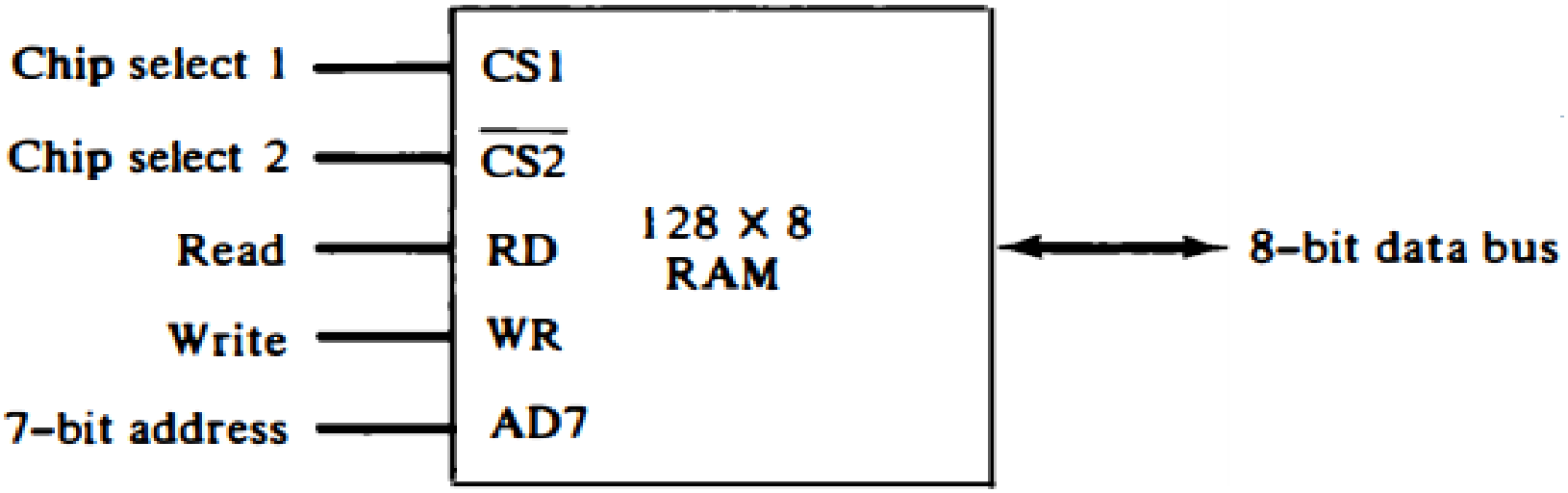
- The RAM integrated circuit chips are further classified into two possible operating modes, **static** and **dynamic**.
- The primary compositions of a static RAM are *flip-flops* that store the binary information.
- The nature of the stored information is volatile.
- The static RAM is easy to use and takes less time performing read and write operations as compared to dynamic RAM.

RAM integrated circuit chips

- The dynamic RAM exhibits the binary information **in the form of electric charges that are applied to capacitors.**
- The capacitors are integrated inside the chip by MOS transistors.
- The dynamic RAM consumes less power and provides large storage capacity in a single memory chip.
- RAM chips are available in a variety of sizes and are used as per the system requirement.

BASIS FOR COMPARISON	SRAM	DRAM
Speed	Faster	Slower
Size	Small	Large
Cost	Expensive	Cheap
Used in	Cache memory	Main memory
Density	Less dense	Highly dense
Construction	Complex and uses transistors and latches.	Simple and uses capacitors and very few transistors.
Single block of memory requires	6 transistors	Only one transistor.
Charge leakage property	Not present	Present hence require power refresh circuitry
Power consumption	Low	High

Figure 2 Typical RAM chip.



(a) Block diagram

CS1	$\overline{CS2}$	RD	WR	Memory function	State of data bus
0	0	x	x	Inhibit	High-impedance
0	1	x	x	Inhibit	High-impedance
1	0	0	0	Inhibit	High-impedance
1	0	0	1	Write	Input data to RAM
1	0	1	x	Read	Output data from RAM
1	1	x	x	Inhibit	High-impedance

(b) Function table

RAM integrated circuit chips

- 128 * 8 RAM chip has a memory capacity of 128 words of eight bits (one byte) per word.
- This requires a 7-bit address and an 8-bit bidirectional data bus.
- The 8-bit bidirectional data bus allows the transfer of data either from memory to CPU during a **read** operation or from CPU to memory during a **write** operation.

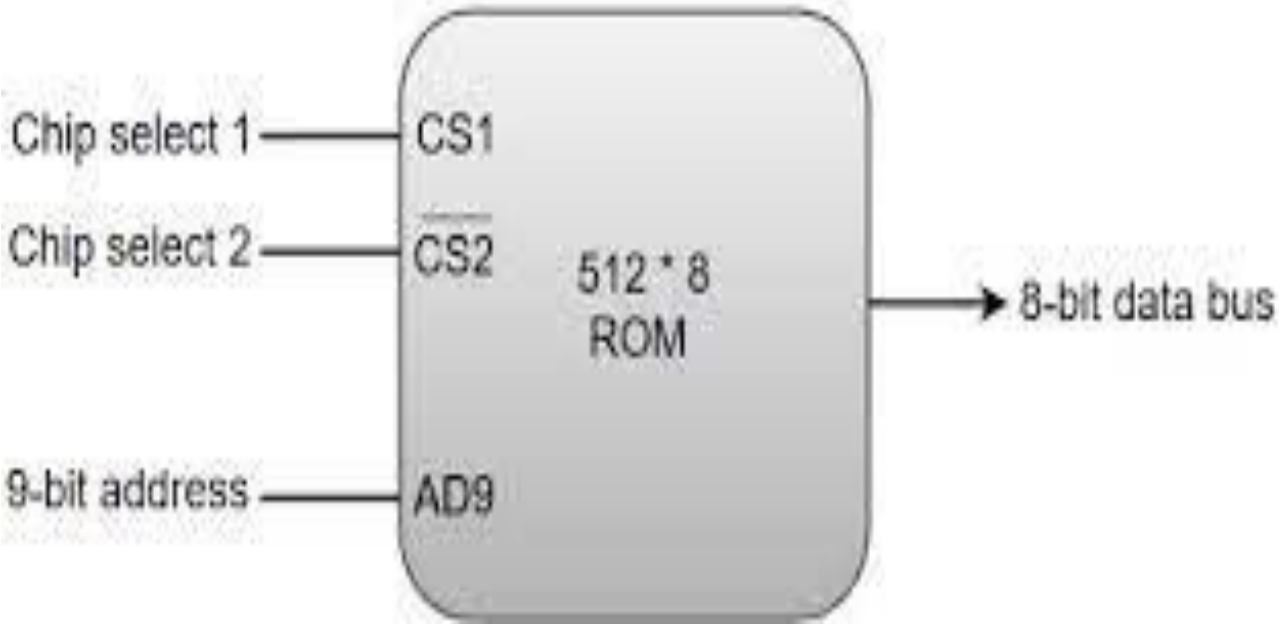
RAM integrated circuit chips

- The **read** and **write** inputs specify the memory operation, and the two chip select (CS) control inputs are for enabling the chip only when the microprocessor selects it.
- The bidirectional data bus is constructed using **three-state buffers**.
- The output generated by three-state buffers can be placed in one of the three possible states which include a signal equivalent to logic 1, a signal equal to logic 0, or a high-impedance state.

ROM integrated circuit chips

- The primary component of the main memory is RAM integrated circuit chips, but a portion of memory may be constructed with ROM chips.
- A ROM memory is used for keeping programs and data that are permanently resident in the computer.

Typical ROM chip:



ROM integrated circuit chips

- ROM is used for storing an initial program called a **bootstrap loader**.
- The primary function of the **bootstrap loader** program is to start the computer software operating when power is turned on.
- ROM chips are also available in a variety of sizes and are also used as per the system requirement.

ROM integrated circuit chips

- A ROM can only perform read operation; the data bus can only operate in an output mode.
- The 9-bit address lines in the ROM chip specify any one of the 512 bytes stored in it.
- The value for chip select 1 and chip select 2 must be 1 and 0 for the unit to operate. Otherwise, the data bus is said to be in a high-impedance state.

Secondary Memory

Auxiliary Memory/Secondary Memory

- An Auxiliary memory is known as the lowest-cost, highest-capacity and slowest-access storage in a computer system.
- The most common examples of auxiliary memories are **magnetic tapes and magnetic disks.**

Secondary Memory

- This type of memory is also known as external memory or non-volatile.
- It is slower than the main memory.
- These are used for storing data/information permanently.
- **CPU directly does not access these memories.**
- For example, disk, CD-ROM, DVD, etc.

Characteristics of Secondary Memory

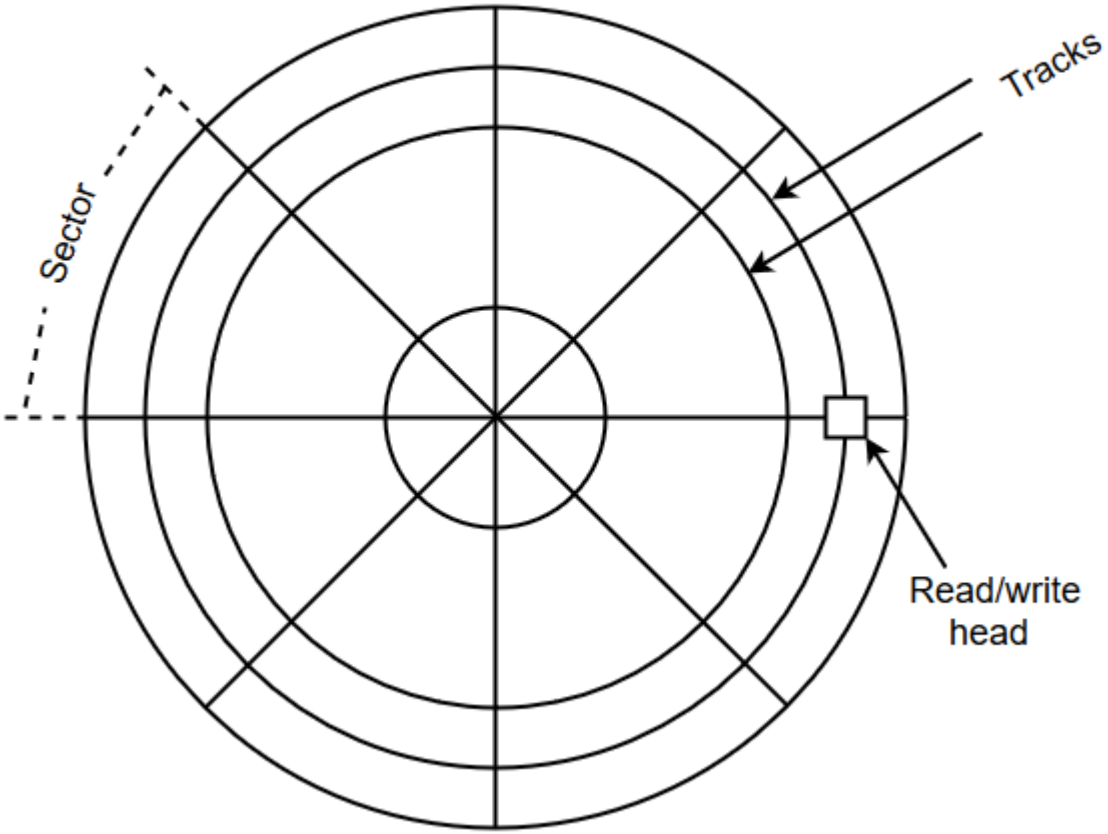
- These are magnetic and optical memories.
- It is known as the backup memory.
- It is a non-volatile memory.
- Data is permanently stored even if power is switched off.
- It is used for storage of data in a computer.
- Computer may run without the secondary memory.
- Slower than primary memories.

Auxiliary Memory/Secondary Memory

Magnetic Disks

- A magnetic disk is constructed using **a circular plate of metal or plastic coated with magnetized materials.**
- Usually, both sides of the disks are used to carry out read/write operations.
- However, several disks may be stacked on one spindle with read/write head available on each surface.

Magnetic disks

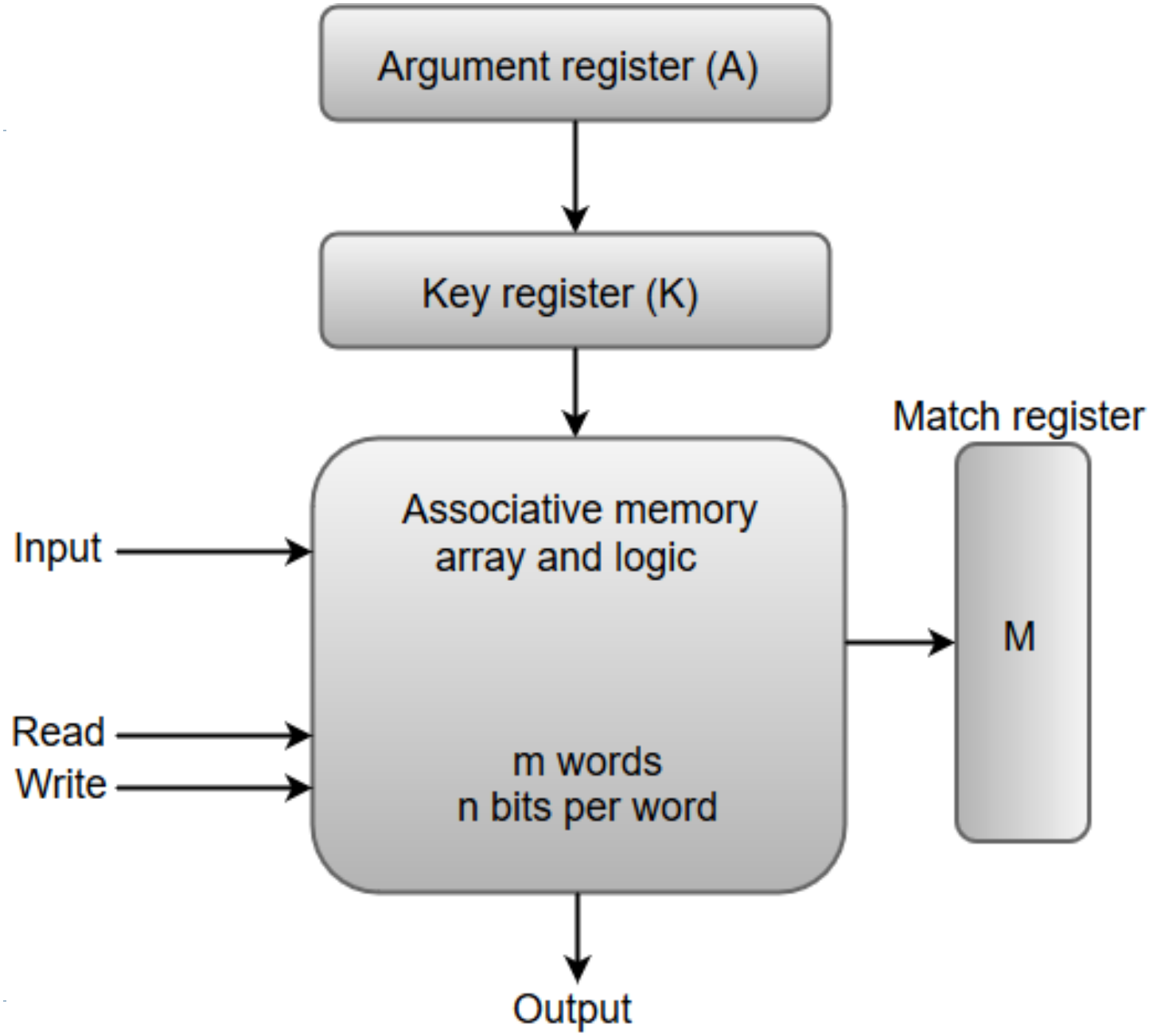


Magnetic Disks

- The memory bits are stored in the magnetized surface in spots along the concentric circles called **tracks.**
- The concentric circles (tracks) are commonly divided into sections called **sectors**.

Associative Memory

- An associative memory is a memory unit whose stored data can be identified for access by the content of the data itself rather than by an address or memory location.
- Associative memory is often referred to as **Content Addressable Memory (CAM)**.
- When a write operation is performed on associative memory, no address or memory location is given to the word.
- The memory itself is capable of finding an empty unused location to store the word.


Argument register (A)
Key register (K)
Match register
Associative memory
array and logic
Input
Read
Write
m words
n bits per word
M
Output

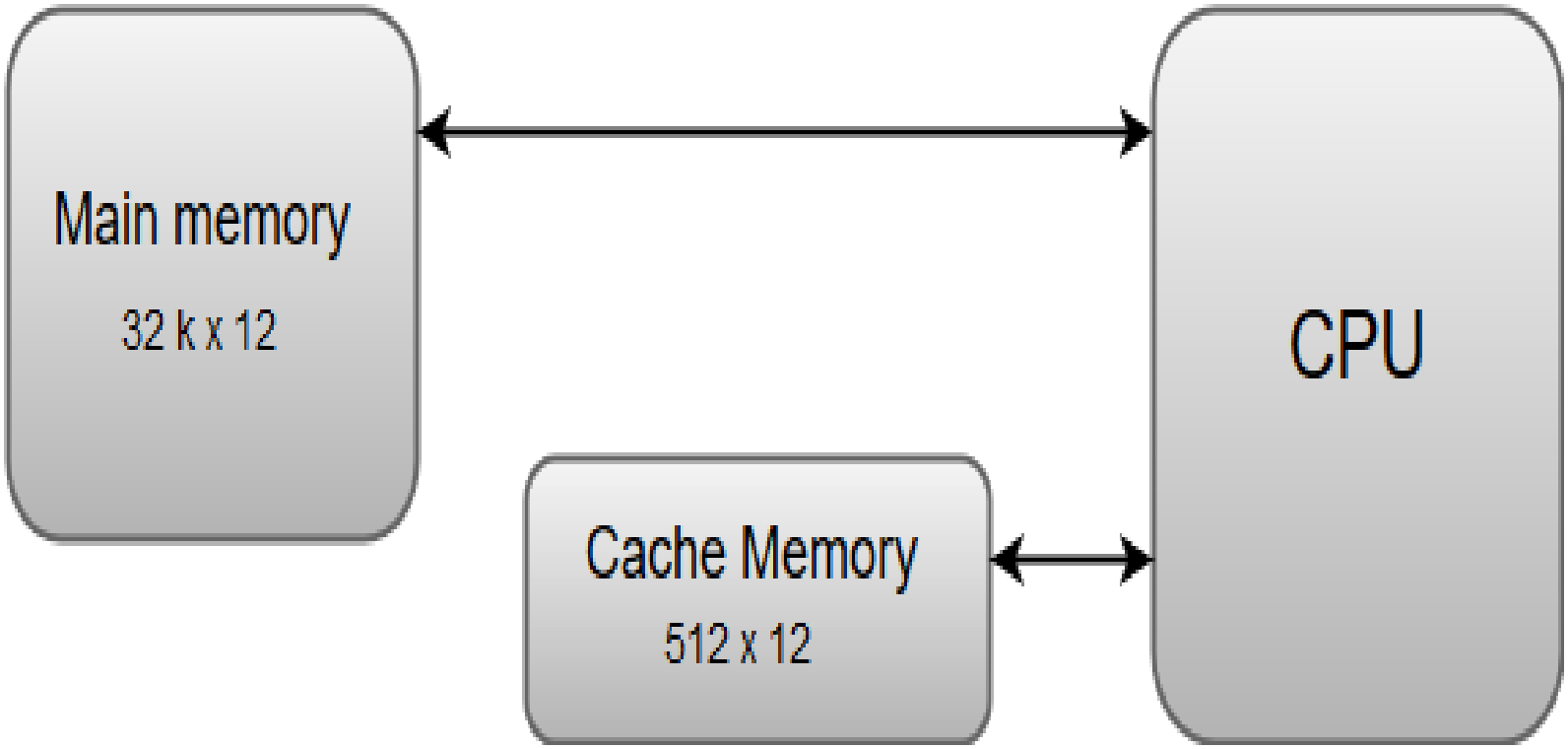
- It consists memory array of m words with n bits per words.
- Argument register A and key register K have n bits one for each bit of word.
- Match register has m bits, one for each memory word. •
- Each word in memory is compared in parallel with the content of the A register. For the word that match corresponding bit in the match register is set.

- The key register (K) provides a mask for choosing a particular field or key in the argument word.
- If the key register contains a binary value of all 1's, then the entire argument is compared with each memory word. Otherwise, only those bits in the argument that have 1's in their corresponding position of the key register are compared.
- Thus, the key provides a mask for identifying a piece of information which specifies how the reference to memory is made.

Cache Memory

- The data or contents of the main memory that are used frequently by CPU are stored in the cache memory so that the processor can easily access that data in a shorter time.
- Whenever the CPU needs to access memory, it first checks the cache memory. If the data is not found in cache memory, then the CPU moves into the main memory.
- Cache memory is placed between the CPU and the main memory.

Cache Memory



Cache Memory

- The cache is the fastest component in the memory hierarchy.

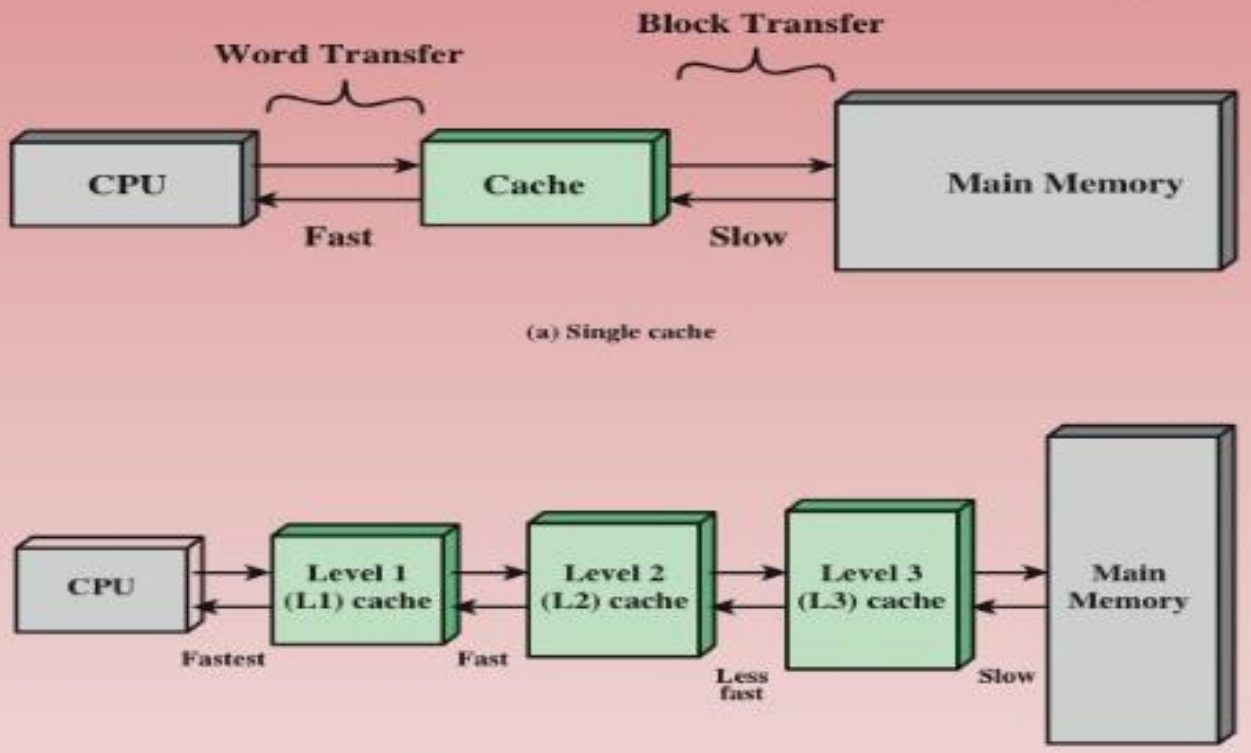
The basic operation of a cache memory

- When the CPU needs to access memory, the cache is examined. If the word is found in the cache, it is read from the fast memory.
- If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word.
- The performance of the cache memory is frequently measured in terms of **hit ratio**.

The basic operation of a cache memory

- When the CPU refers to memory and finds the word in cache, it is said to produce a **hit**.
- If the word is not found in the cache, it is in main memory and it counts as a **miss**.
- The ratio of the number of hits divided by the total CPU references to memory (hits plus misses) is the **hit ratio**.

Cache and Main Memory



Memory subsystem

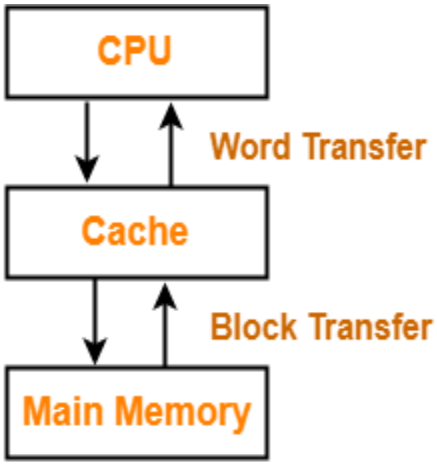
Memory Management system : cache memory

Recap

- Cache memory is a Random Access Memory.
- The main advantage of cache memory is its very fast speed.
- It can be accessed by the CPU at much faster speed than main memory.
- SRAM (Cache), DRAM (Main), and flash (Nonvolatile)

Cache memory :Location

- Cache memory is placed between the CPU and the main memory.
- It facilitates the transfer of data between the processor and the main memory at the speed which matches to the speed of the processor.



Cache and Main Memory

Cache memory :Location

- Data is transferred in the form of ***words between the cache memory and the CPU.***
- Data is transferred in the form of ***blocks or pages between the cache memory and the main memory.***

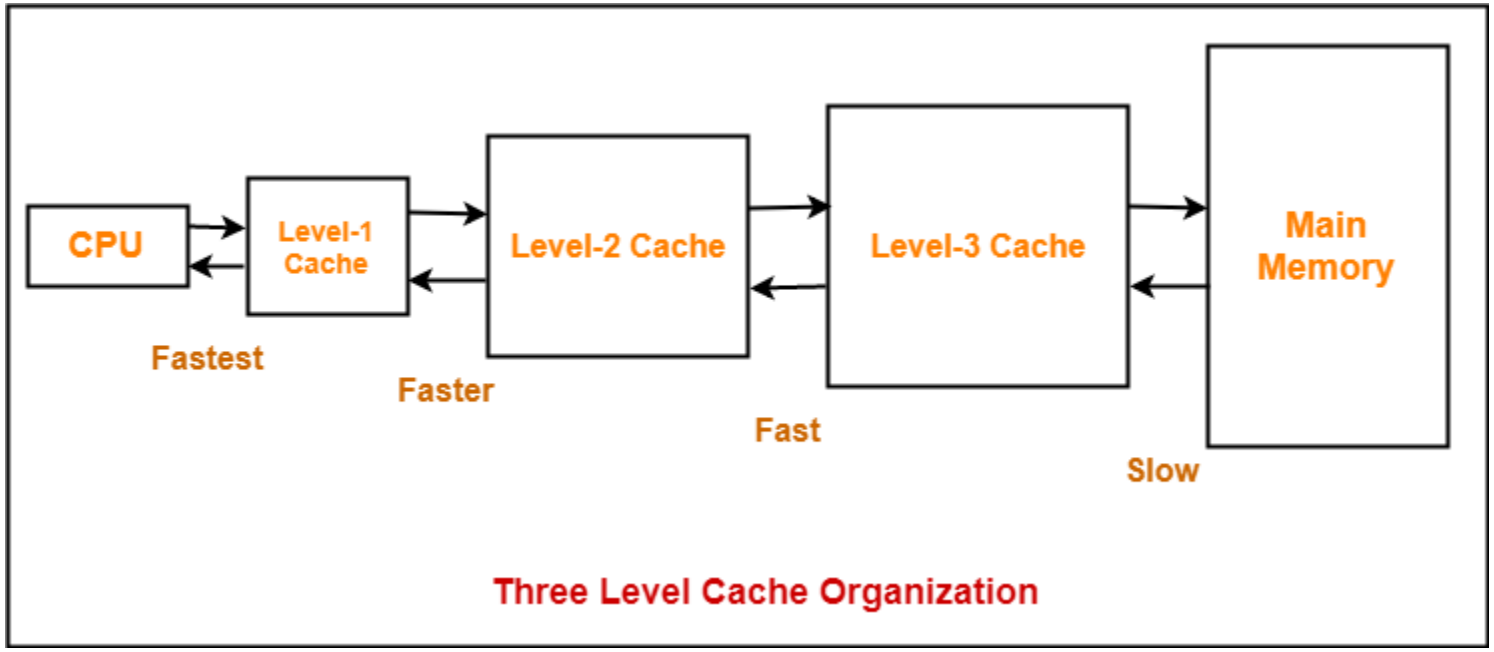
Why to use cache memory?

- The fast speed of the cache memory makes it extremely useful.
- It is used for bridging the speed mismatch between the fastest CPU and the main memory.
- It does not let the CPU performance suffer due to the slower speed of the main memory.

Multilevel Cache Organization

- A multilevel cache organization is an organization where cache memories of different sizes are organized at multiple levels to increase the processing speed to a greater extent.
- The smaller the size of cache, the faster its speed.
- The smallest size cache memory is placed closest to the CPU.
- This helps to achieve better performance in terms of speed.

Multilevel Cache Organization



Multilevel Cache Organization

- **Size (L1 Cache) < Size (L2 Cache) < Size (L3 Cache) < Size (Main Memory)**

cache mapping techniques

When cache hit occurs,

- The required word is present in the cache memory.
- The required word is delivered to the CPU from the cache memory.

When cache miss occurs,

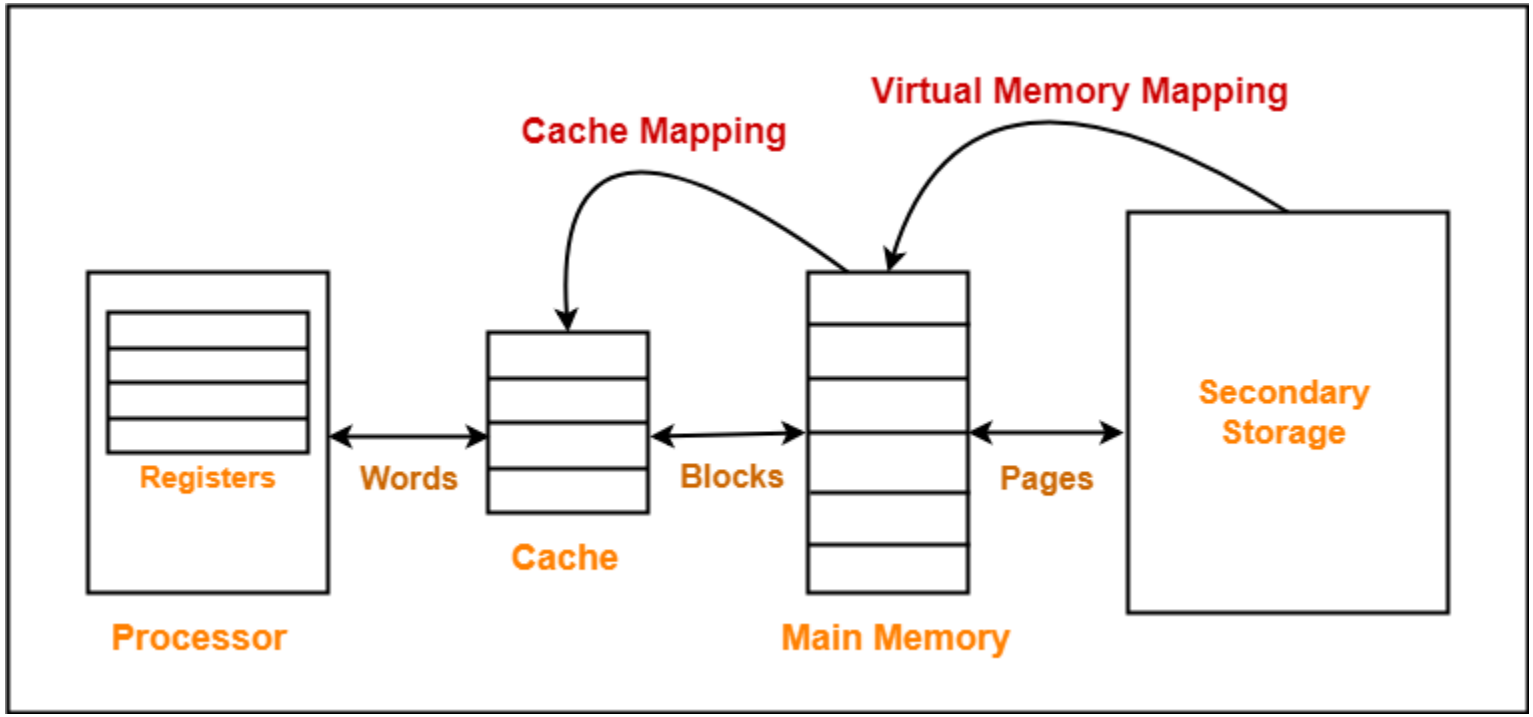
- The required word is not present in the cache memory.
- The page containing the required word has to be mapped from the main memory.
- This mapping is performed using cache mapping techniques.
-

cache mapping techniques

Cache Mapping-

- Cache mapping defines how a block from the main memory is mapped to the cache memory in case of a cache miss.
- **OR**
- Cache mapping is a technique by which the contents of main memory are brought into the cache memory.

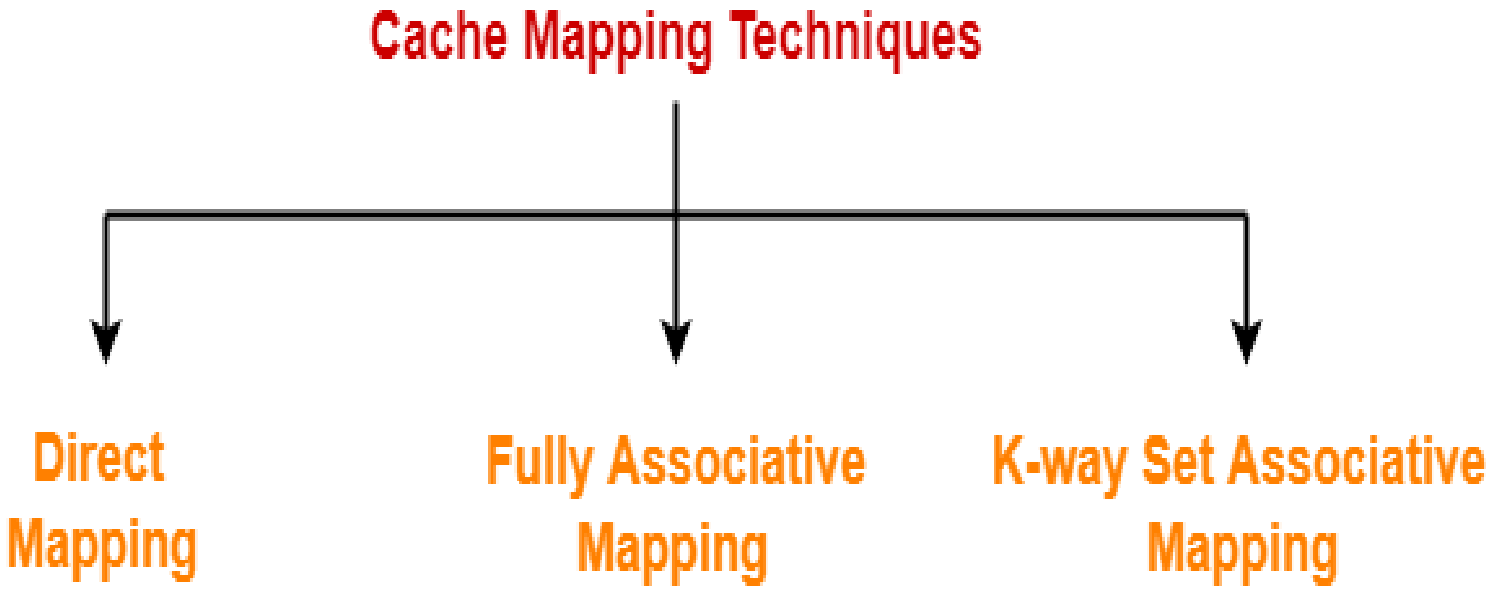
Cache Mapping



Cache Mapping

- Main memory is divided into equal size partitions called as **blocks** or **frames**.
- Cache memory is divided into partitions having same size as that of blocks called as **lines**.
- During cache mapping, block of main memory is simply copied to the cache and the block is not actually brought from the main memory.

Cache mapping is performed using three different techniques



1.Direct Mapping

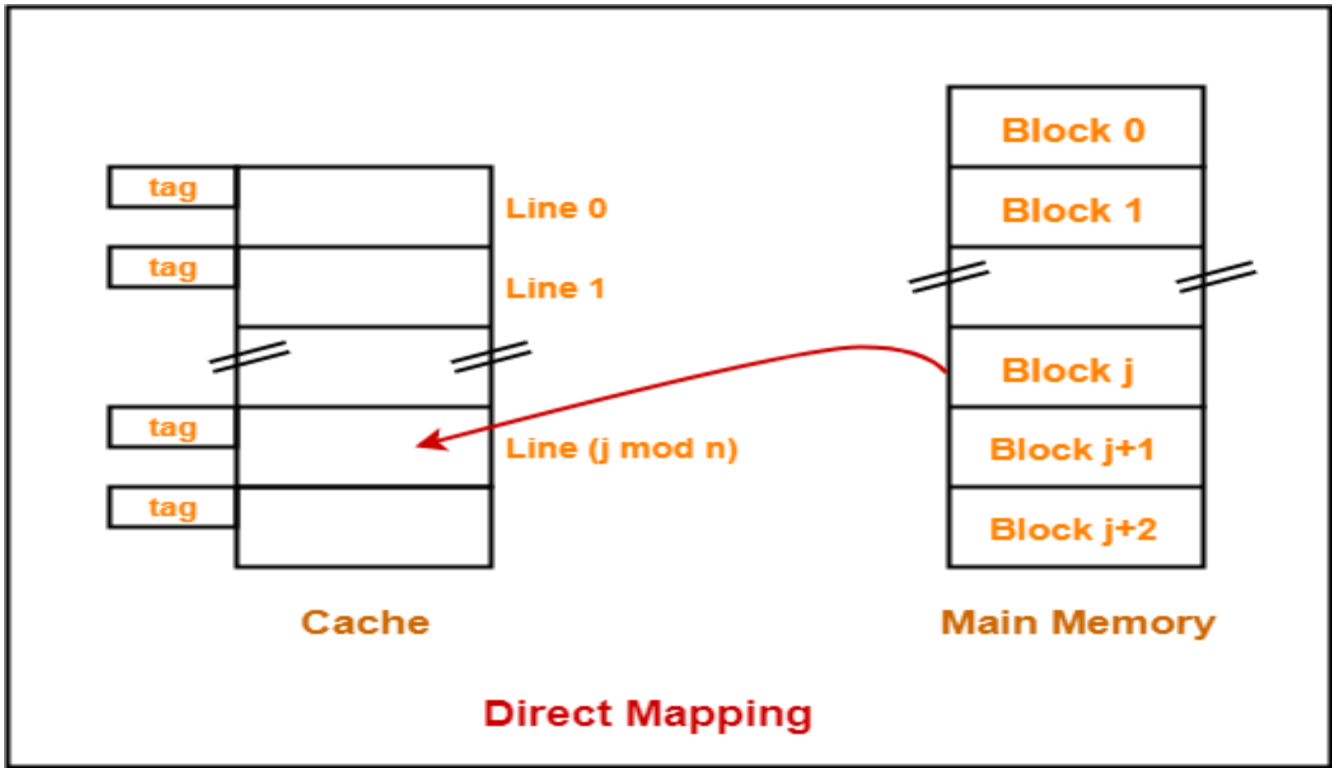
In direct mapping,

- 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)**

1.Direct Mapping

- Consider cache memory is divided into ‘n’ number of lines.
- Then, block ‘j’ of main memory can map to line number (j mod n) only of the cache.

1.Direct Mapping



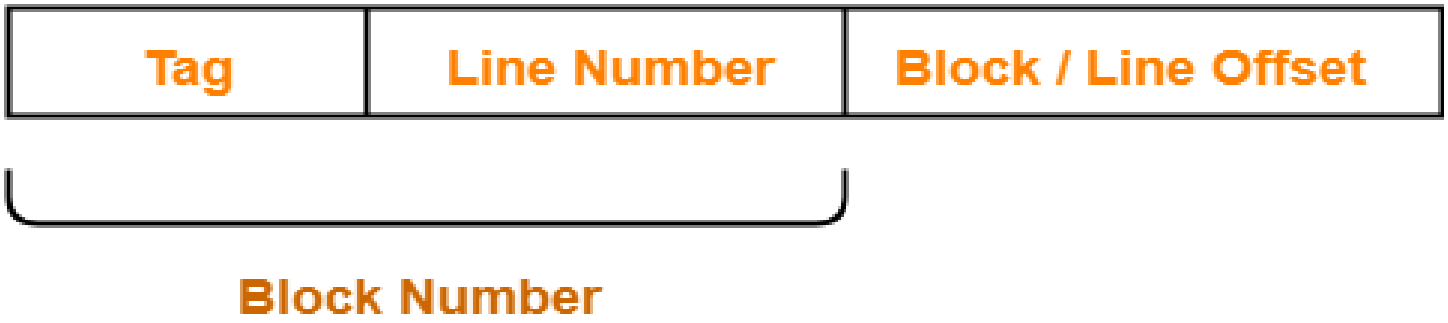
1.Direct Mapping

In direct mapping,

- There is ***no need of any replacement algorithm***.
- This is because a main memory block can map only to a particular line of the cache.
- Thus, the new incoming block will always replace the existing block (if any) in that particular line.

1.Direct Mapping

In direct mapping, the physical address is divided as



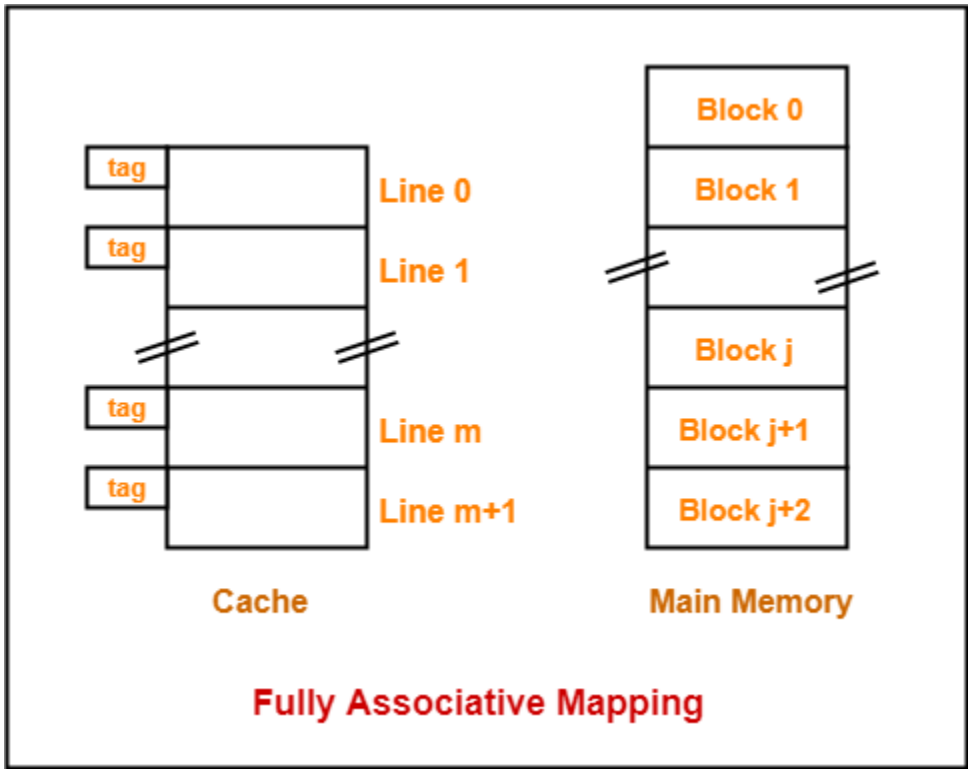
Division of Physical Address in Direct Mapping

2. Fully Associative Mapping

In fully associative mapping,

- A block of main memory can map to any line of the cache that is freely available at that moment.
- This makes fully associative mapping more flexible than direct mapping.

2. Fully Associative Mapping



2. Fully Associative Mapping

- All the lines of cache are freely available.
- Thus, any block of main memory can map to any line of the cache.
- Had all the cache lines been occupied, then one of the existing blocks will have to be replaced.

2. Fully Associative Mapping

In fully associative mapping,

- A replacement algorithm is required.
- Replacement algorithm suggests the block to be replaced if all the cache lines are occupied.
- Thus, replacement algorithm like FCFS Algorithm, LRU Algorithm etc is employed.

2. Fully Associative Mapping

In fully associative mapping, the physical address is divided as

Division of Physical Address in Fully Associative Mapping

3. K-way Set Associative Mapping

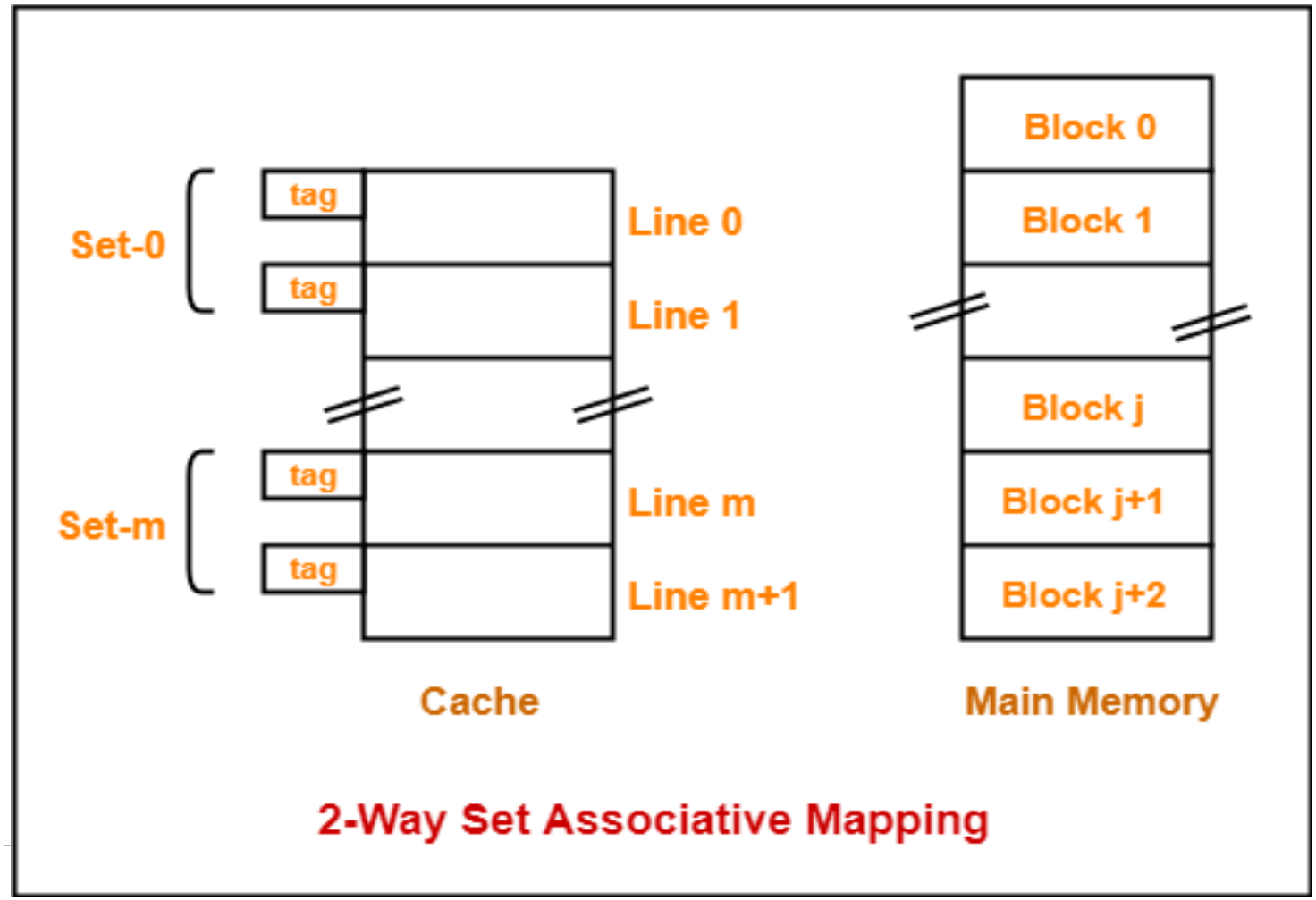
In k-way set associative mapping,

- Cache lines are grouped into sets where each set contains k number of lines.
- A particular block of main memory can map to only one particular set of the cache.
- However, within that set, the memory block can map any cache line that is freely available.
- The set of the cache to which a particular block of the main memory can map is given by-

Cache set number

= (Main Memory Block Address) Modulo (Number of sets in Cache)

3. K-way Set Associative Mapping



3. K-way Set Associative Mapping

- k = 2 suggests that each set contains two cache lines.
- Since cache contains 6 lines, so number of sets in the cache = 6 / 2 = 3 sets.
- Block 'j' of main memory can map to set number (j mod 3) only of the cache.
- Within that set, block 'j' can map to any cache line that is freely available at that moment.
- If all the cache lines are occupied, then one of the existing blocks will have to be replaced.

3. K-way Set Associative Mapping

- Set associative mapping is a combination of direct mapping and fully associative mapping.
- It uses fully associative mapping within each set.
- Thus, set associative mapping requires a replacement algorithm.
-
- In set associative mapping, the physical address is divided as

Division of Physical Address in K-way Set Associative Mapping

3. K-way Set Associative Mapping

- If k = 1, then k-way set associative mapping becomes direct mapping i.e.

1-way Set Associative Mapping ≡ Direct Mapping

If k = Total number of lines in the cache, then k-way set associative mapping becomes fully associative mapping.

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.

Types of Locality of reference

- **Spatial Locality of reference**
  There is a chance that element will be present in the close proximity to the reference point and next time if again searched then more close proximity to the point of reference.

Types of Locality of reference

- **Temporal Locality of reference**
  In this Least recently used algorithm will be used. Whenever there is page fault occurs within a word will not only load word in main memory but complete page fault will be loaded because spatial locality of reference rule says that if you are referring any word next word will be referred in its register that's why we load complete page table so the complete block will be loaded.

Concept

- **Locality of reference** refers to a phenomenon in which a computer program tends to access same set of memory locations for a particular time period.

- The property of locality of reference is mainly shown by loops and subroutine calls in a program.

Locality of reference

- In case of loops in program control processing unit repeatedly refers to the set of instructions that constitute the loop.
- In case of subroutine calls, everytime the set of instructions are fetched from memory.
- References to data items also get localized that means same data item is referenced again and again.

Locality of reference

- when the CPU wants to read or fetch the data or instruction ,First it will access the cache memory as it is near to it and provides very fast access.
- If the required data or instruction is found, it will be fetched. This situation is known as a cache hit.
- But if the required data or instruction is not found in the cache memory then this situation is known as a cache miss.
- Now the main memory will be searched for the required data or instruction that was being searched and if found will go through one of the two ways
-

Locality of reference

- First way is that the CPU should fetch the required data or instruction and use it and that's it but what, when the same data or instruction is required again. CPU again has to access the same main memory location for it and we already know that main memory is the slowest to access.
- The second way is to store the data or instruction in the cache memory so that if it is needed soon again in the near future it could be fetched in a much faster way.

Cache Operation

It is based on the principle of locality of reference. There are two ways with which data or instruction is fetched from main memory and get stored in cache memory.

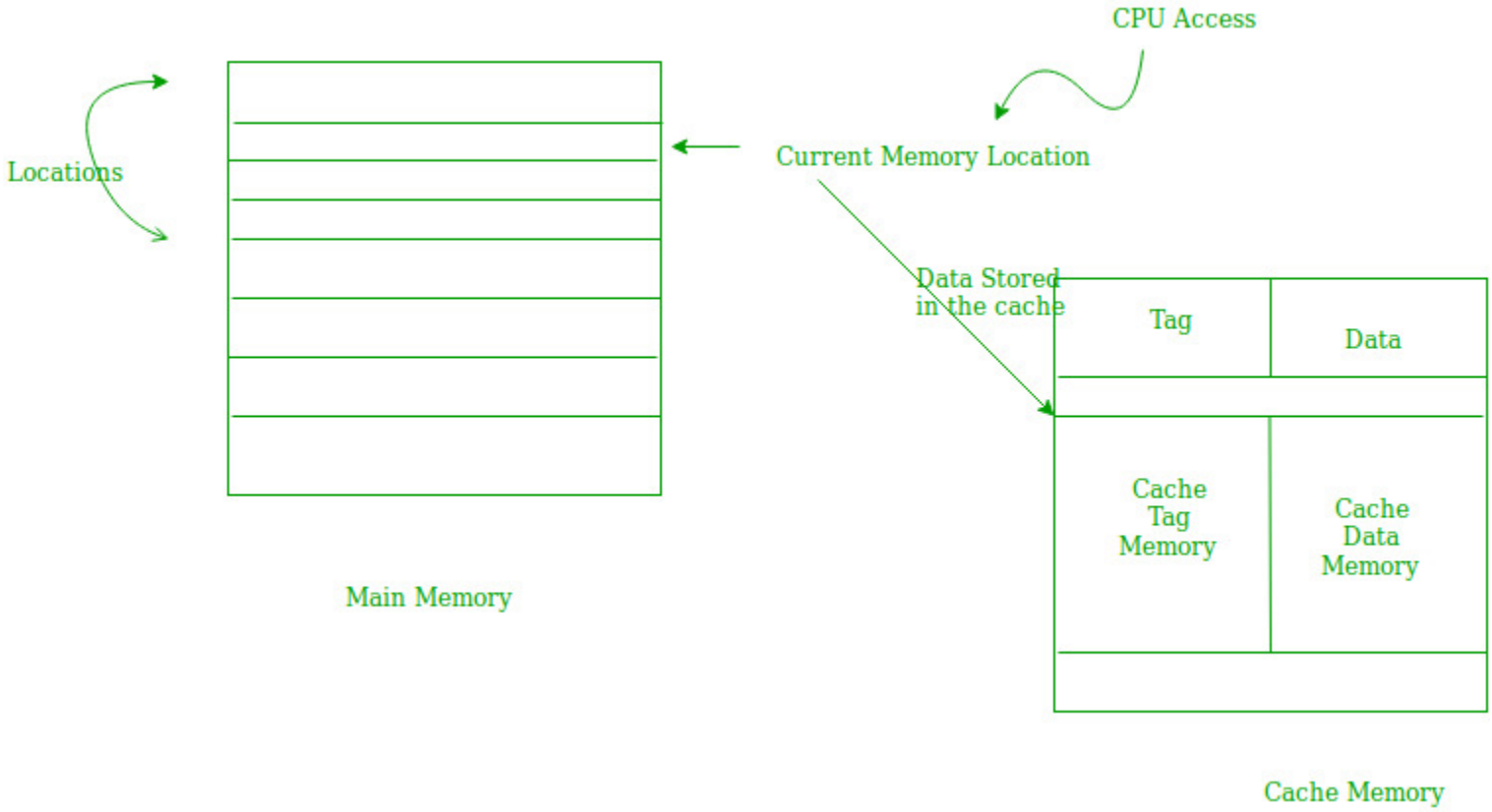
These two ways to do so.

Temporal Locality

Temporal locality means current data or instruction that is being fetched may be needed soon.

So we should store that data or instruction in the cache memory so that we can avoid again searching in main memory for the same data.

Temporal Locality



Spatial Locality

Spatial locality means instruction or data near to the current memory location that is being fetched, may be needed soon in the near future.

This is slightly different from the temporal locality.

Here we are talking about nearly located memory locations while in temporal locality we were talking about the actual memory location that was being fetched.

Spatial Locality