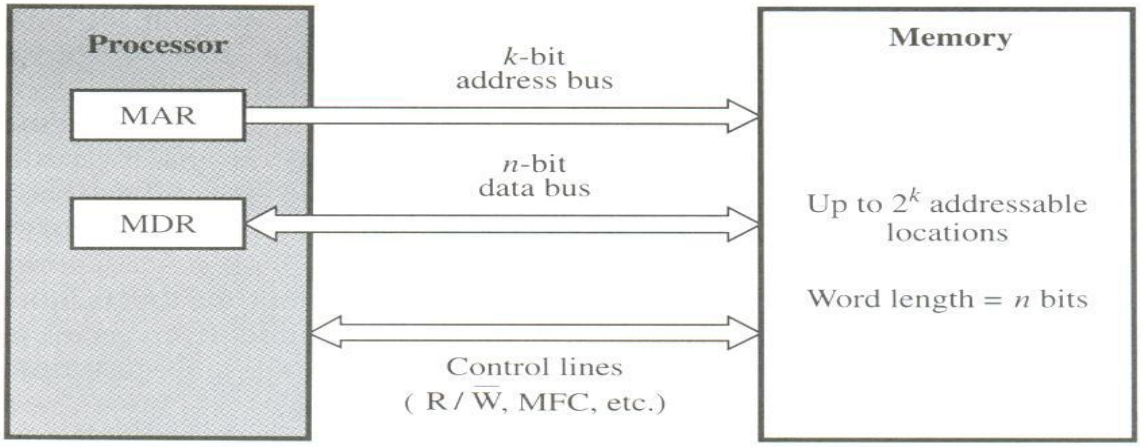
UNIT-V

5.1 SOME BASIC MEMORY CONCEPTS

The maximum size of the memory that can be used in any computer is determined by the addressing scheme.

Address	**memory location**
16 bit	2^{16} = 64 K
32 bit	2^{32} = 4G (Giga)
40 bit	2^{40} = IT (Tera)

Fig: Connection of Memory to Processor



1. If MAR is k bits long and MDR is n bits long, then the memory may contain upto 2K addressable locations and the n-bits of data are transferred between the memory and processor. This transfer takes place over the processor bus.
2. The processor bus has,
   - ➢ Address Line
   - ➢ Data Line
   - ➢ Control Line (R/W, MFC – Memory Function Completed)
3. The control line is used for coordinating data transfer.
4. The processor reads the data from the memory by loading the address of the required memory location into MAR and setting the R/W line to 1.
5. The memory responds by placing the data from the addressed location onto the data lines and confirms this action by asserting MFC signal.
6. Upon receipt of MFC signal, the processor loads the data onto the data lines into MDR register.
7. The processor writes the data into the memory location by loading the address of this location into MAR and loading the data into MDR sets the R/W line to 0.

Memory Access Time → It is the time that elapses between the initiation of an Operation and the completion of that operation.
Memory Cycle Time → It is the minimum time delay that required between the initiation of the two successive memory operations.

5.2 MEMORY SYSTEM CONSIDERATIONS

To reduce the number of pins, the dynamic memory chips use multiplexed address inputs.

The address is divided into two parts. They are,

- **High Order Address Bit**(Select a row in cell array, are provided first and latched into memory chips under the control of RAS signal).
- **Low Order Address Bit**(Selects a column and they are provided on same address pins and latched using CAS signals).

The Multiplexing of address bit is usually done by **Memory Controller Circuit.**

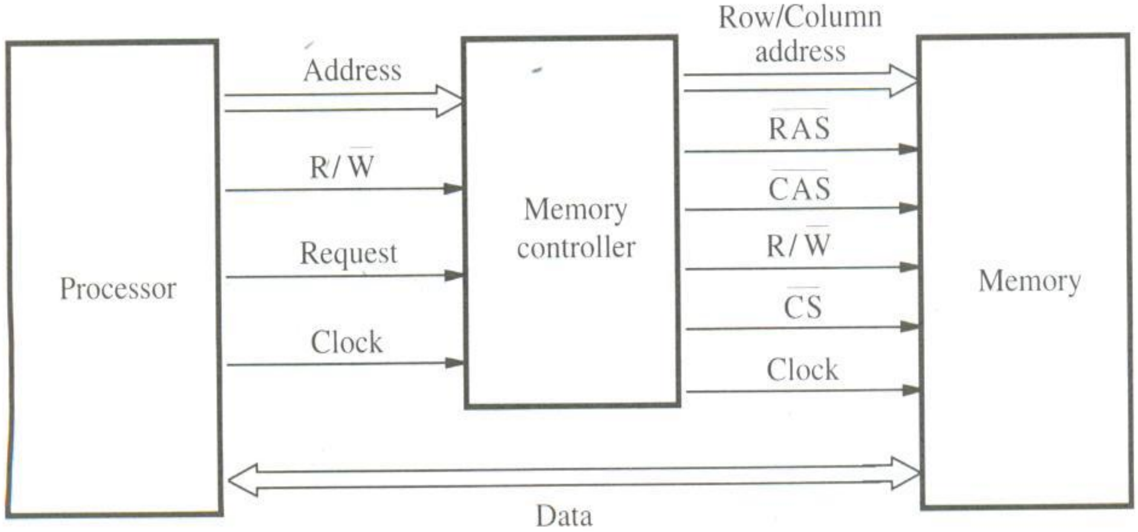


Figure: Use of Memory Controller

The Controller accepts a complete address & R/W signal from the processor, under the control of a Request signal which indicates that a memory access operation is needed. The Controller then forwards the row & column portions of the address to the memory and generates RAS &CAS signals. It also sends R/W &CS signals to the memory. The CS signal is usually active low, hence it is shown as CS.

Refresh Overhead:

All dynamic memories have to be refreshed.

In DRAM ,the period for refreshing all rows is 16ms whereas 64ms in SDRAM.

Example: Given a cell array of 8K(8192).
Clock cycle=4
Clock Rate=133MHZ
No of cycles to refresh all rows =8192*4 =32,768
Time needed to refresh all rows=32768/133*10-6
=246*10-6 sec
=0.246sec
Refresh Overhead =0.246/64
Refresh Overhead =0.0038

5.3 READ ONLY MEMORY

1. Both SRAM and DRAM chips are volatile, which means that they lose the stored information if power is turned off.
2. Many application requires Non-volatile memory (which retain the stored information if power is turned off).

   Example : Operating System software has to be loaded from disk to memory which requires the program that boots the Operating System i.e. It requires non-volatile memory.
3. Non-volatile memory is used in embedded system.
4. Since the normal operation involves only reading of stored data ,a memory of this type is called ROM.

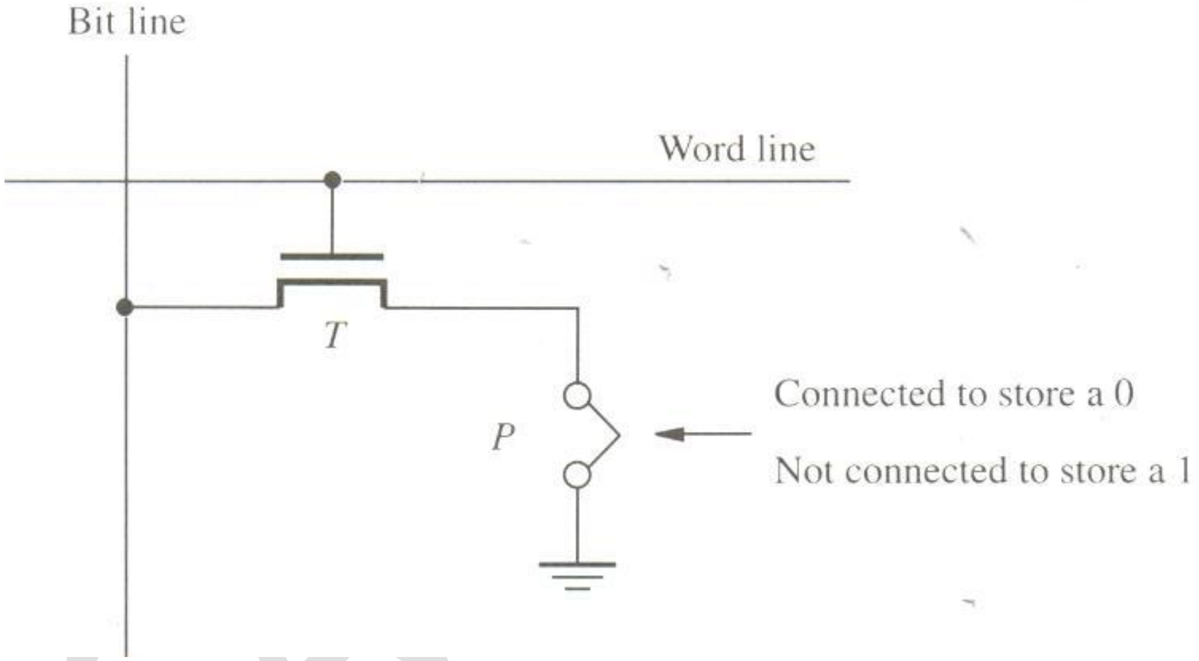


Figure: ROM cell

At Logic value '0' → Transistor(T) is connected to the ground point(P). Transistor switch is closed & voltage on bit line nearly drops to zero.

At Logic value '1' → Transistor switch is open. The bit line remains at high voltage. To read the state of the cell, the word line is activated. A Sense circuit at the end of the bit line generates the proper output value.

Types of ROM

Different types of non-volatile memory are,

- PROM
- EPROM
- EEPROM
- Flash Memory

PROM:-Programmable ROM:

PROM allows the data to be loaded by the user. Programmability is achieved by inserting a “fuse‟ at point P in a ROM cell. Before it is programmed, the memory contains all” 0‟s .The user can insert “1‟s at the required location by burning out the fuse at these locations using high-current pulse. This process is irreversible.

Merit:

- It provides flexibility.
- It is faster.
- It is less expensive because they can be programmed directly by the user.

EPROM:-Erasable reprogrammable ROM

EPROM allows the stored data to be erased and new data to be loaded. In an EPROM cell, a connection to ground is always made at “P‟ and a special transistor is used, which has the ability to function either as a normal transistor or as a disabled transistor that is always turned “off‟. This transistor can be programmed to behave as a permanently open switch, by injecting charge into it that becomes trapped inside. Erasure requires dissipating the charges trapped in the transistor of memory cells. This can be done by exposing the chip to ultra-violet light, so that EPROM chips are mounted in packages that have transparent windows.

Merits:

- It provides flexibility during the development phase of digital system.
- It is capable of retaining the stored information for a long time.

Demerits:

- The chip must be physically removed from the circuit for reprogramming and its entire contents are erased by UV light.

EEPROM:-Electrically Erasable ROM:

Merits:

- It can be both programmed and erased electrically.
- It allows the erasing of all cell contents selectively.

Demerits:

- It requires different voltage for erasing ,writing and reading the stored data.

Flash Memory:

- In EEPROM, it is possible to read & write the contents of a single cell.
- In Flash device, it is possible to read the contents of a single cell but it is only possible to write the entire contents of a block.
- Prior to writing, the previous contents of the block are erased.

  Example. In MP3 player, the flash memory stores the data that represents sound.
- Single flash chips cannot provide sufficient storage capacity for embedded system application.
- There are 2 methods for implementing larger memory modules consisting of number of chips. They are,
  - Flash Cards
  - Flash Drives.

Merits:

- Flash drives have greater density which leads to higher capacity & low cost per bit.

- It requires single power supply voltage & consumes less power in their operation.

Flash Cards

One way of constructing larger module is to mount flash chips on a small card. Such flash card have standard interface. The card is simply plugged into a conveniently accessible slot. Its memory size are of 8,32,64MB.

Example :A minute of music can be stored in 1MB of memory. Hence 64MB flash cards can store an hour of music.

Flash Drives:

Larger flash memory module can be developed by replacing the hard disk drive. The flash drives are designed to fully emulate the hard disk. The flash drives are solid state electronic devices that have no movable parts.

Merits:

- They have shorter seek and access time which results in faster response.
- They have low power consumption which makes them attractive for battery driven application.
- They are insensitive to vibration.

Demerit:

- The capacity of flash drive (<1GB) is less than hard disk(>1GB).
- It leads to higher cost per bit.
- Flash memory will deteriorate after it has been written a number of times(typically at least 1 million times.)

5.4 CACHE MEMORIES

The effectiveness of cache mechanism is based on the property of "**Locality of reference'.**

Locality of Reference

Many instructions in the localized areas of the program are executed repeatedly during some time period and remainder of the program is accessed relatively infrequently. It manifests itself in 2 ways. They are,

- **Temporal**(The recently executed instruction are likely to be executed again very soon.)
- **Spatial(**The instructions in close proximity to recently executed instruction are also likely to be executed soon.)

If the active segment of the program is placed in cache memory, then the total execution time can be reduced significantly. The term Block refers to the set of contiguous address locations of some size. The cache line is used to refer to the cache block.

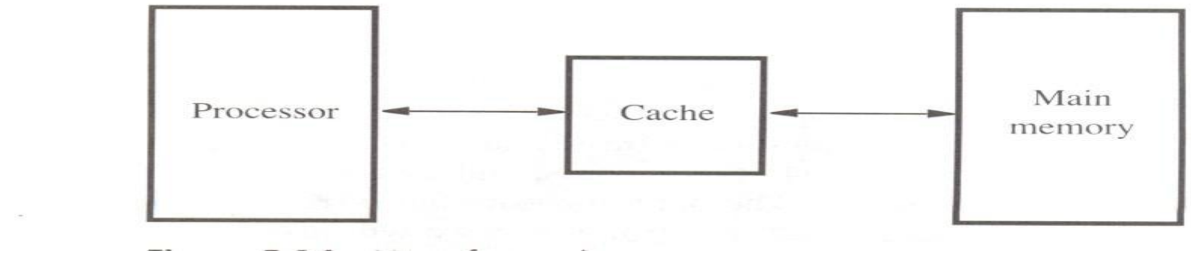


Figure: Use of Cache Memory

The Cache memory stores a reasonable number of blocks at a given time but this number is small compared to the total number of blocks available in Main Memory. The correspondence between main memory block and the block in cache memory is specified by a mapping function. The Cache control hardware decide that which block should be removed to create space for the new block that contains the referenced word. The collection of rule for making this decision is called the **replacement algorithm.** The cache control circuit determines whether the requested word currently exists in the cache. If it exists, then Read/Write operation will take place on appropriate cache location. In this case **Read/Write hit** will occur. In a Read operation, the memory will not involve.

The write operation is proceed in 2 ways. They are,

- ➢ Write-through protocol
- ➢ Write-back protocol

Write-through protocol:

Here the cache location and the main memory locations are updated simultaneously.

Write-back protocol:

This technique is to update only the cache location and to mark it as with associated flag bit called **dirty/modified bit.** The word in the main memory will be updated later, when the block containing this marked word is to be removed from the cache to make room for a new block. If the requested word currently not exists in the cache during read operation, then **read miss** will occur. To overcome the read miss **Load –through / Early restart protocol** is used.

Read Miss:

The block of words that contains the requested word is copied from the main memory into cache.

Load –through:

After the entire block is loaded into cache, the particular word requested is forwarded to the processor. If the requested word not exists in the cache during write operation, then **Write Miss** will occur. If Write through protocol is used, the information is written directly into main memory. If Write back protocol is used then block containing the addressed word is first brought into the cache and then the desired word in the cache is over-written with the new information.

Mapping Function

There are three cache mapping techniques

1. **Direct Mapping**
2. **Associative Mapping**
3. **Set-Associative Mapping**

1.Direct Mapping:

- It is the simplest technique in which block j of the main memory maps onto block „j‟ modulo 128 of the cache.
- Thus whenever one of the main memory blocks 0,128,256 is loaded in the cache, it is stored in block 0.
- Block 1,129,257 are stored in cache block 1 and so on.

  The contention may arise when,

  - ➢ When the cache is full
  - ➢ When more than one memory block is mapped onto a given cache block position.
- The contention is resolved by allowing the new blocks to overwrite the currently resident block.
- Placement of block in the cache is determined from memory address.

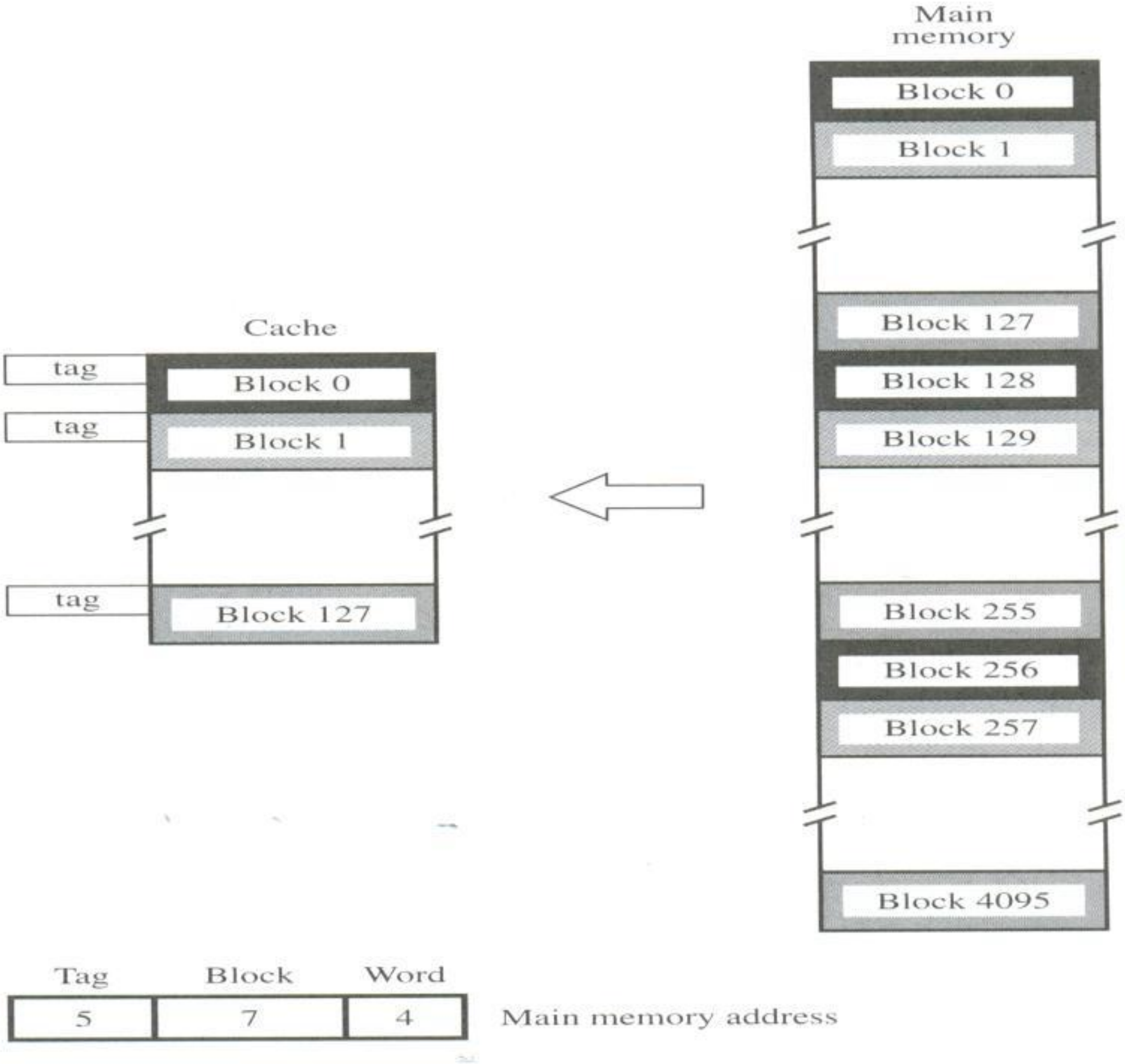


Figure: Direct Mapped Cache

The memory address is divided into 3 fields. They are,

- **Low Order 4 bit field(word)**:Selects one of 16 words in a block.
- **7 bit cache block field**: When new block enters cache,7 bit determines the cache position in which this block must be stored.
- **5 bit Tag field:** The high order 5 bits of the memory address of the block is stored in 5 tag bits associated with its location in the cache.
- As execution proceeds, the high order 5 bits of the address is compared with tag bits associated with that cache location.
- If they match, then the desired word is in that block of the cache.
- If there is no match, then the block containing the required word must be first read from the main memory and loaded into the cache.

Merit:

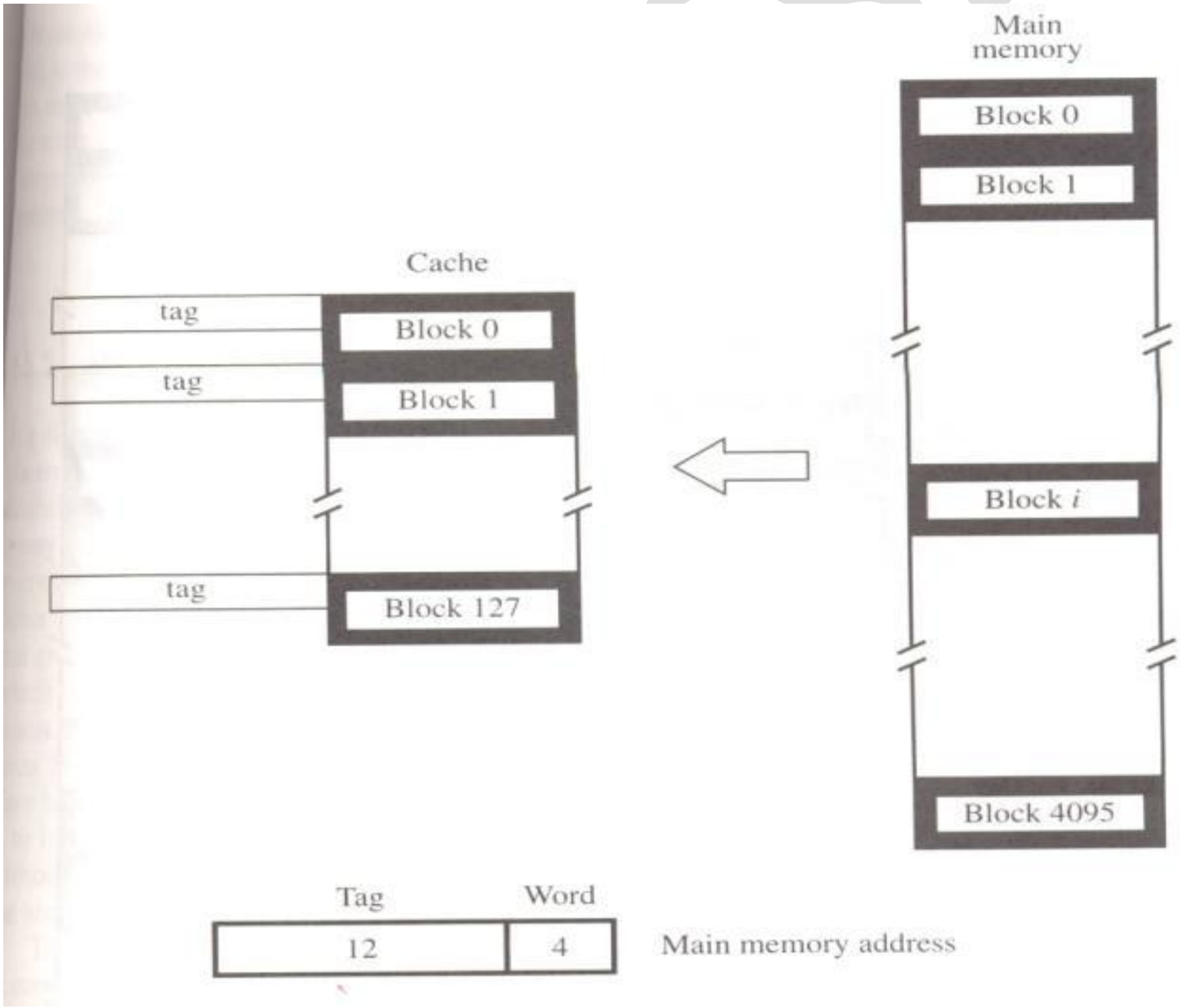
- It is easy to implement.

Demerit:

- It is not very flexible.

2. Associative Mapping

- In this method, the main memory block can be placed into any cache block position
- 12 tag bits will identify a memory block when it is resolved in the cache.
- The tag bits of an address received from the processor are compared to the tag bits of each block of the cache to see if the desired block is present. This is called **associative mapping.**
- **I**t gives complete freedom in choosing the cache location.
- A new block that has to be brought into the cache has to replace(eject)an existing block if the cache is full.
- In this method, the memory has to determine whether a given block is in the cache.
- A search of this kind is called an **associative Search.**



Merit:

- It is more flexible than direct mapping technique.

Demerit:

- Its cost is high.

3.Set-Associative Mapping

- It is the combination of direct and associative mapping.
- The blocks of the cache are grouped into sets and the mapping allows a block of the main memory to reside in any block of the specified set.
- In this case, the cache has two blocks per set, so the memory blocks 0,64,128……..4032 maps into cache set “0‟ and they can occupy either of the two block position within the set.
- 6 bit set field-> Determines which set of cache contains the desired block .
- 6 bit tag field->The tag field of the address is compared to the tags of the two blocks of
- the set to clock if the desired block is present.

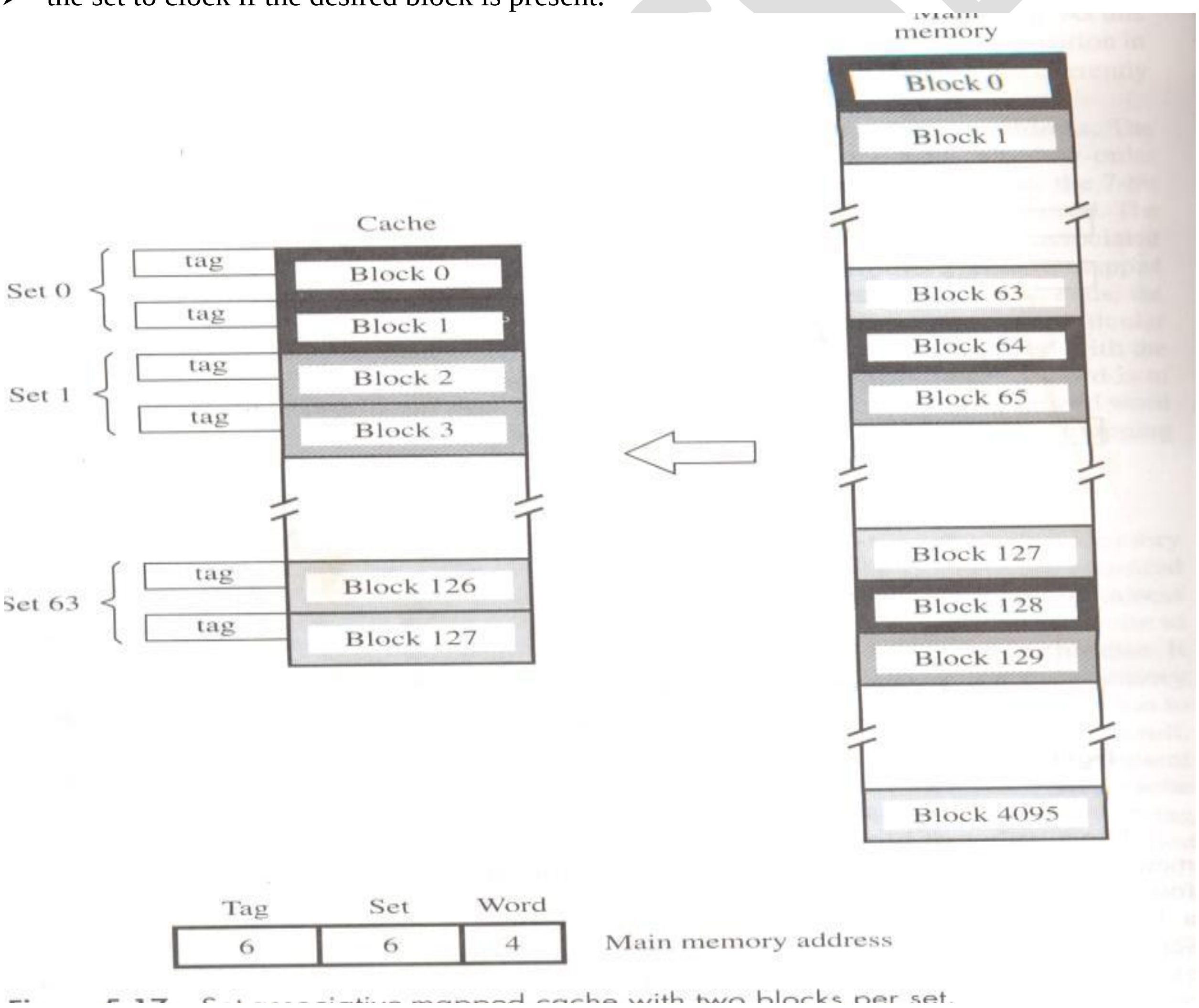


Fig: Set-Associative Mapping

No of blocks per set	no of set field
2	6
3	5
8	4
128	no set field

5.5 SECONDARY STORAGE

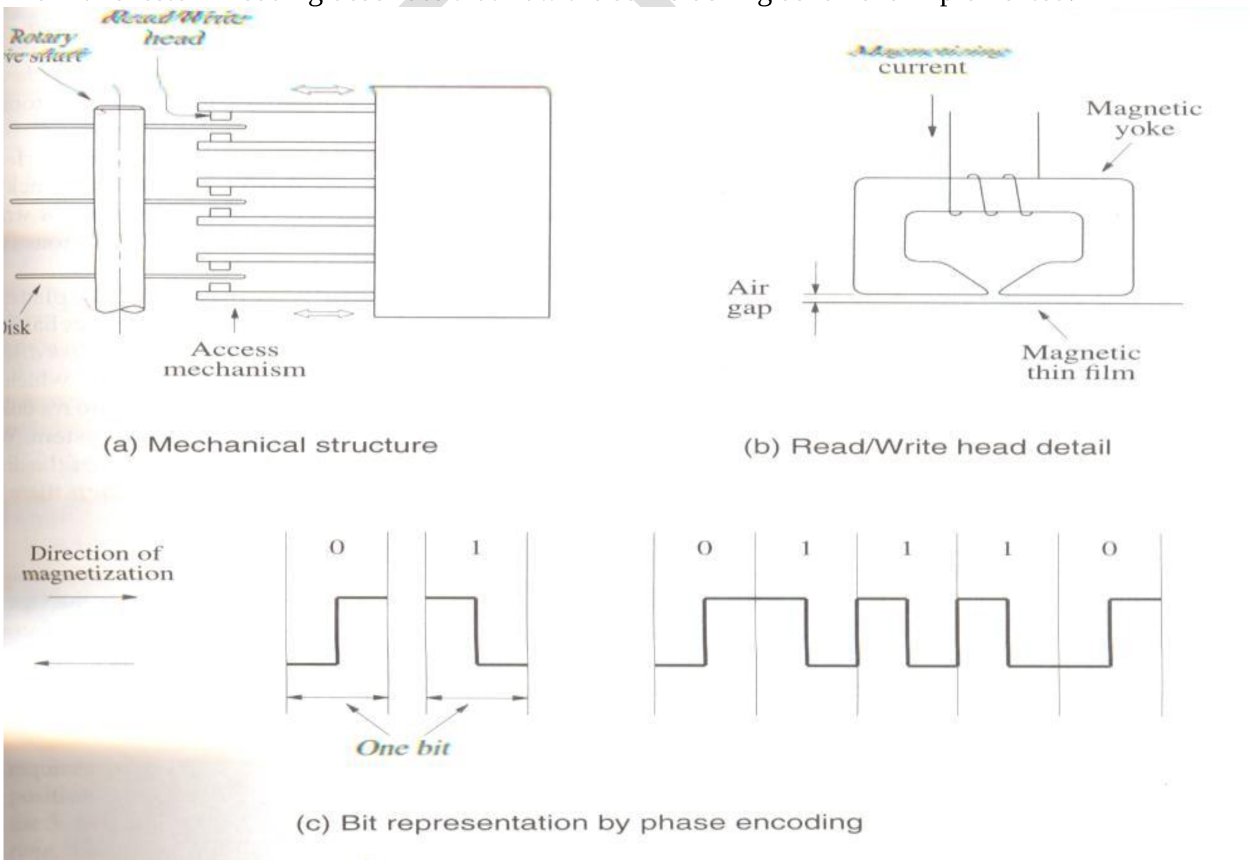
The Semi-conductor memories do not provide all the storage capability. The Secondary storage devices provide larger storage requirements. Some of the Secondary Storage devices are,

- Magnetic Disk
- Optical Disk
- Magnetic Tapes.

Magnetic Disk:

Magnetic Disk system consists o one or more disk mounted on a common spindle. A thin magnetic film is deposited on each disk, usually on both sides. The disk are placed in a rotary drive so that the magnetized surfaces move in close proximity to read /write heads. Each head consists of **magnetic yoke & magnetizing coil**. Digital information can be stored on the magnetic film by applying the current pulse of suitable polarity to the magnetizing coil. Only changes in the magnetic field under the head can be sensed during the Read operation. Therefore if the binary states 0 & 1 are represented by two opposite states of magnetization, a voltage is induced in the head only at 0-1 and at 1-0 transition in the bit stream. A consecutive (long string) of 0‟s & 1‟s are determined by using the clock which is mainly used for synchronization. Phase Encoding or Manchester Encoding is the technique to combine the clocking information with data.

The Manchester Encoding describes that how the self-clocking scheme is implemented.



(a) Mechanical structure

(b) Read/Write head detail

(c) Bit representation by phase encoding

The Read/Write heads must be maintained at a very small distance from the moving disk surfaces in order to achieve high bit densities. When the disk are moving at their steady state, the air pressure develops between the disk surfaces & the head & it forces the head away from the surface. The flexible spring connection between head and its arm mounting permits the head to fly at the desired distance away from the surface.

Winchester Technology

Read/Write heads are placed in a sealed, air –filtered enclosure called the Winchester Technology In such units, the read/write heads can operate closure to magnetic track surfaces because the dust particles which are a problem in unsealed assemblies are absent.

Merits:

It have a larger capacity for a given physical size. The data intensity is high because the storage medium is not exposed to contaminating elements. The read/write heads of a disk system are movable.

The disk system has 3 parts. They are,

- **Disk Platter**(Usually called Disk)
- **Disk Drive(**spins the disk & moves Read/write heads)
- **Disk Controller**(controls the operation of the system.)

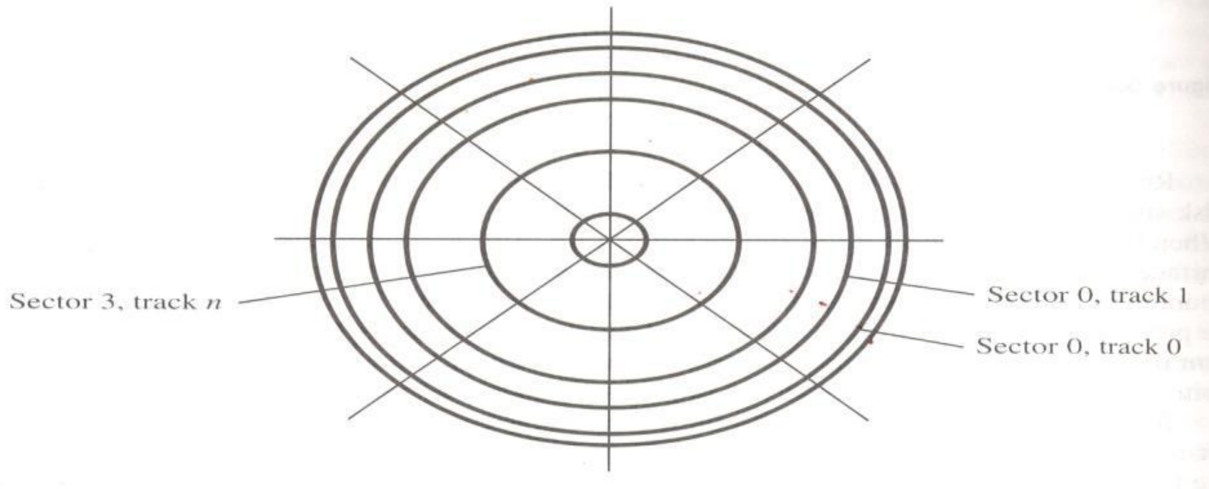


Fig:Organizing & Accessing the data on disk

Each surface is divided into concentric **tracks**. Each track is divided into **sectors**. The set of corresponding tracks on all surfaces of a stack of disk form a **logical cylinder.** The data are accessed by specifying **the surface number, track number and the sector number.** The Read/Write operation start at sector boundaries. Data bits are stored serially on each track. Each sector usually contains 512 bytes.

Sector header -> contains identification information. It helps to find the desired sector on the selected track.

ECC (Error checking code)- used to detect and correct errors. An unformatted disk has no information on its tracks. The formatting process divides the disk physically into tracks and sectors and this process may discover some defective sectors on all tracks. The disk controller keeps a record of such defects.

The disk is divided into logical partitions. They are,

- Primary partition
- Secondary partition

In the diagram, Each track has same number of sectors. So all tracks have same storage capacity. Thus the stored information is packed more densely on inner track than on outer track.

Access time

There are 2 components involved in the time delay between receiving an address and the beginning of the actual data transfer. They are,

- Seek time
- Rotational delay / Latency

Seek time – Time required to move the read/write head to the proper track.

Latency – The amount of time that elapses after the head is positioned over the correct track until the starting position of the addressed sector passes under the read/write head.

Seek time + Latency = Disk access time

Disk Controller

The disk controller acts as interface between disk drive and system bus. The disk controller uses DMA scheme to transfer data between disk and main memory. When the OS initiates the transfer by issuing Read/Write request, the controllers register will load the following information. They are, Main memory address(address of first main memory location of the block of words involved in the transfer)

Disk address(The location of the sector containing the beginning of the desired block of words) (number of words in the block to be transferred).