COMPUTER ORGANIZATION
UNIT-5: THE MEMORY SYSTEMS

Syllabus:

- Basic memory concepts
- Memory System Consideration
- Read- Only Memory
  - ROM
- PROM
- EPROM
- EEPROM
- Flash Memory
- Cache Memories
  - Mapping Functions
- INTERLEAVING
- Secondary Storage
  - Magnetic Hard Disks
  - Optical Disks

Text Book to Follow:

1. Computer Organization, Carl Hamacher, ZvonksVranesic, SafeaZaky, 5th Edition, McGraw Hill

1. BASIC MEMORY CONCEPTS:

- A memory unit is the collection of storage units or devices together.
- The memory unit stores the binary information in the form of bits.
- Memory/ Storage is classified into 2 categories:
  - i) Volatile Memory
  - ii) Non Volatile Memory

i) Volatile Memory:

- This loses its data, when power is switched off.

ii) Non-Volatile Memory:

- This is a permanent storage and does not lose any data when power is switched off.

Memory Access Methods:

i) Random Access:

- Main memories are random access memories, in which each memory location has a unique address.
- Using this unique address any memory location can be reached in the same amount of time in any order.

ii) Sequential Access:

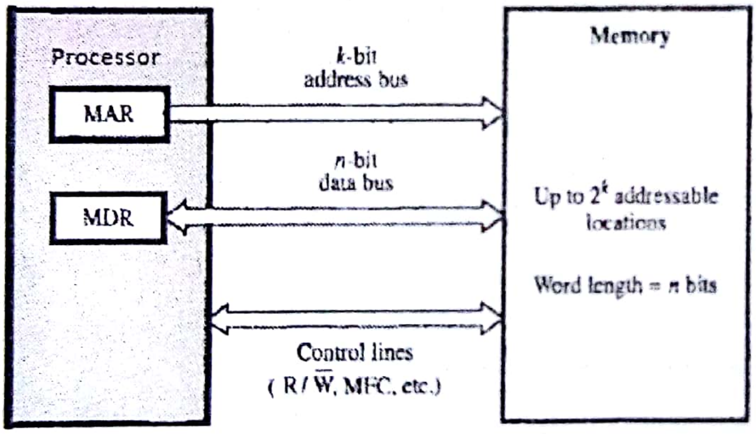
- This method allows memory access in a sequence or in order.

iii) Direct Access:

- In this mode, information is stored in tracks, with each track having a separate read/write head.

Connection of the memory to the Processor:

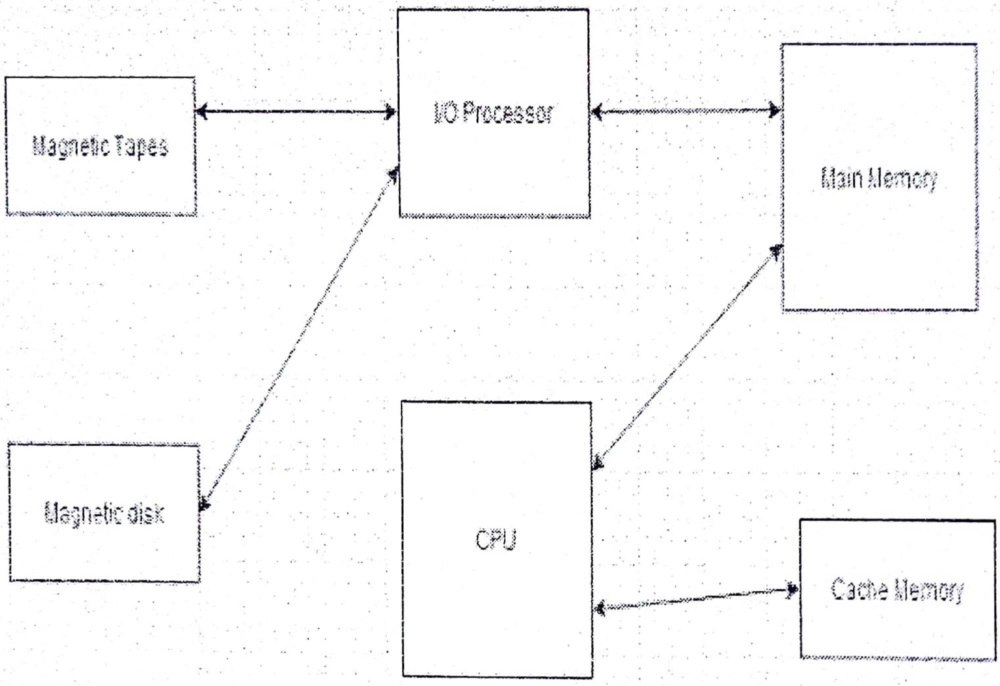
- Connection of the memory to the Processor is depicted as,



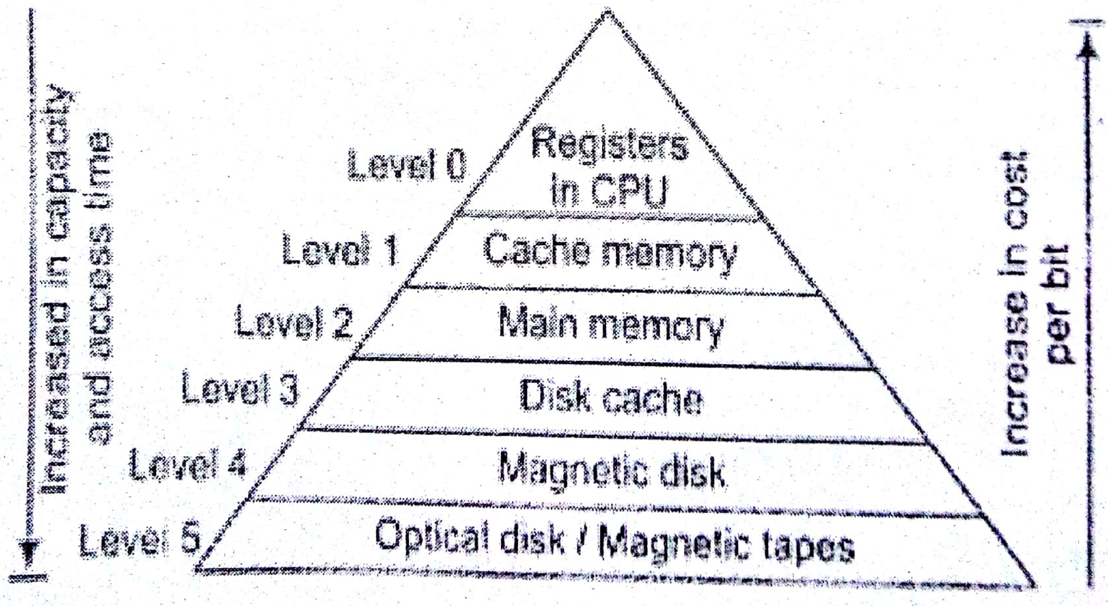
- Data transfer between the memory and the processor takes place through the use of two processor registers
- MAR (Memory Address Register)
- MDR (Memory Data Register)
- If MAR is k bits long and MDR is n bits long, then the memory unit may contain upto 2^k addressable locations.
- During a memory cycle, n bits of data are transferred between the memory and the processor
- This transfer takes place over the processor bus, which has k address lines and n data lines
- The bus also includes the control lines Read/ Write (R/W) and Memory Function Completed (MFC) for coordinating data transfers
- Other control lines may be added to indicate the number of bytes to be transferred

Memory Hierarchy:

- The Memory Hierarchy is depicted as,



- The comparisons in the Memory Hierarchy is depicted as,



2. It uses Semiconductor technology
~~hence known as~~ semiconductor memory.

1. It stores data & programs during computer operations.

2. MAIN MEMORY:

- The memory unit that communicates directly within the CPU, Auxiliary memory and Cache memory, is called main memory.
- It is the central storage unit of the computer system.
- It is a large and fast memory used to store data during computer operations.
- Main memory is made up of **RAM** and **ROM**

i) RAM: Random Access Memory:

DRAM:

- Dynamic RAM, is made of capacitors and transistors
- It is slower and cheaper than SRAM.
- Dynamic RAMs are the pre-dominant choice for implementing computer main memories.
- The high densities achievable in these chips make large memories economically feasible.

SRAM:

- Static RAM, has a six transistor circuit in each cell and retains data, until powered off.
- The choice of a RAM chip for a given application depends on several factors such as, cost, speed, power dissipation, and size of the chip.
- Static RAMs are generally used only when very fast operation is the primary requirement.
- Their cost and size are adversely affected by the complexity of the circuit that realizes the basic cell.
- They are used mostly in cache memories.

ii) ROM: Read Only Memory:

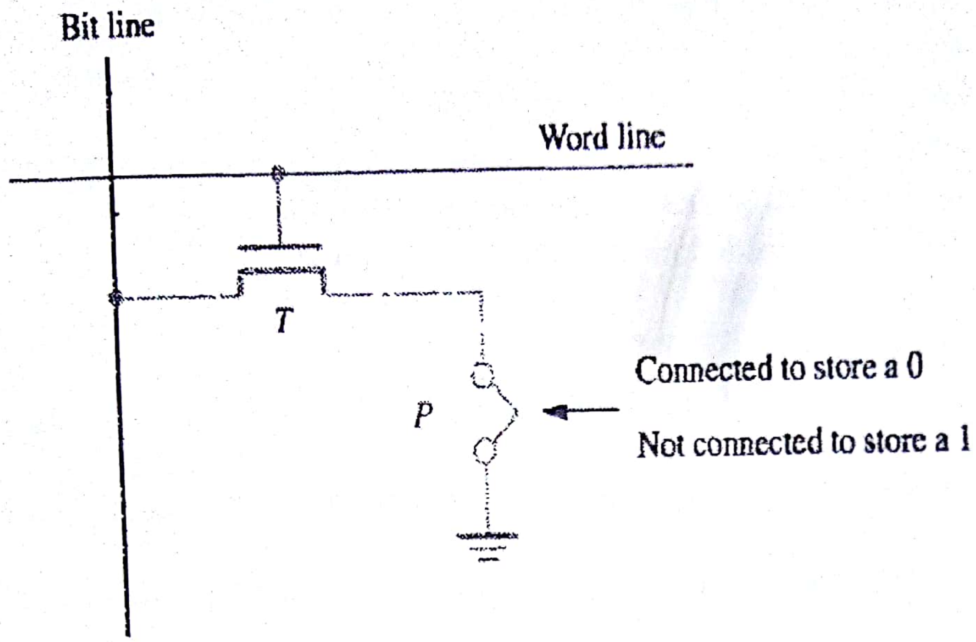
- It is non-volatile and is more like a permanent storage for information.
- It also stores the **bootstrap loader** program, to load and start the operating system when computer is turned on.

The various types of ROM are

- PROM
- EPROM
- EEPROM
- FLASH MEMORY

ROM:

- A ROM cell is depicted as,



- A logic value 0 is stored in the cell if the transistor is connected to ground at point P. otherwise a 1 is stored.
- The bit line is connected through a resistor to the power supply
-To read the state of the cell, the word line is activated
- Thus the transistor switch is closed and the voltage on the bit line drops to near zero if there is a connection between the transistor and ground.
- If there is no connection to ground, the bit line remains at the high voltage, indicating a 1.
- A sense circuit at the end of the bit line generates the proper output value.
- Data are written into a ROM when it is manufactured.

PROM:

- It stands for Programmable Read Only Memory.
- It was first developed in 70s by Texas Instruments.
- It is made as a blank memory.
- A PROM programmer or PROM burner is required in order to write data onto a PROM chip.
- The data stored in it cannot be modified and therefore it is also known as one time programmable device.

EPROM:

- It stands for Erasable Programmable ROM.
- It is different from PROM as unlike PROM the program can be written on it more than once.
- This comes as the solution to the problem faced by PROM.
- The bits of memory come back to 1, when ultra violet rays of some specific wavelength falls into its chip's glass panel.
- The fuses are reconstituted and thus new things can be written on the memory.

EEPROM:

- It stands for Electrically Erasable Read Only Memory.

These are also erasable like EPROM, but the same work of erasing is performed with electric current. Thus, it provides the ease of erasing it even if the memory is positioned in the computer.

It stores computer system's BIOS. Unlike EPROM, the entire chip does not have to be erased for changing some portion of it.

- Thus, it even gets rid of some biggest challenges faced by using EPROMs.

FLASH ROM:

It is an updated version of EEPROM.

In EEPROM, it is not possible to alter many memory locations at the same time.

However, Flash memory provides this advantage over the EEPROM by enabling this feature of altering many locations simultaneously.

It was invented by Toshiba and got its name from it capability of deleting a block of data in a flash.

Flash Cards:

- One way of constructing a larger module is to mount flash chips on a small card.
- Such flash cards have a standard interface that makes them usable in a variety of products.
- A card is simply plugged into a conveniently accessible slot.
- Flash cards come in a variety of memory sizes.
- Typical sizes are 8, 32, and 64 Mbytes.

Flash Drives:

- Larger flash memory modules have been developed to replace hard disk drives.

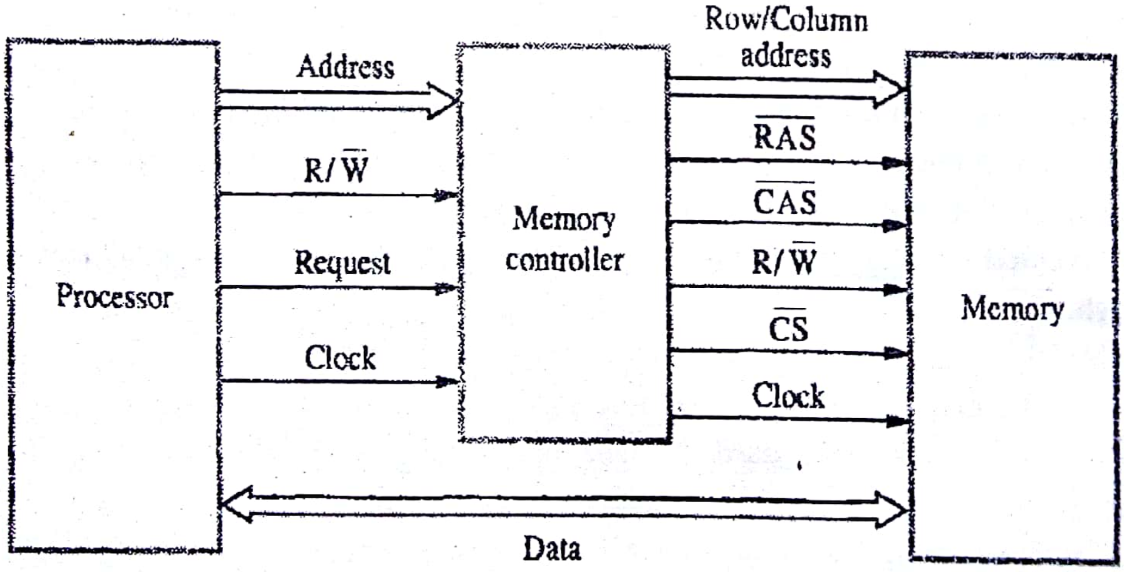
- These flash drives are designed to fully emulate the hard disks, to the point that they can be fitted into standard disk drive bays.
- However, the storage capacity of the flash drives is significantly lower.
- Currently, the capacity of flash drives is less than one Giga Byte.

3. MEMORY SYSTEM CONSIDERATIONS:

- The choice of a RAM chip for a given application depends on several factors such as, cost, speed, power dissipation, and size of the chip.

Memory controller:

- The address is divided into two parts.
- The high-order address bits, which select a row in the cell array, are provided first and latched into the memory chip under control of the RAS signal.
- Then the lower-order address bits, which select a column, are provided on the same address pins and latched using the CAS signal.
- A typical procedure issues all bits of an address at the same time.
- The required multiplexing of address bits is usually performed by a Memory Controller Circuit, which is interposed between the processor and the dynamic memory
- Use of a Memory Controller is depicted as,

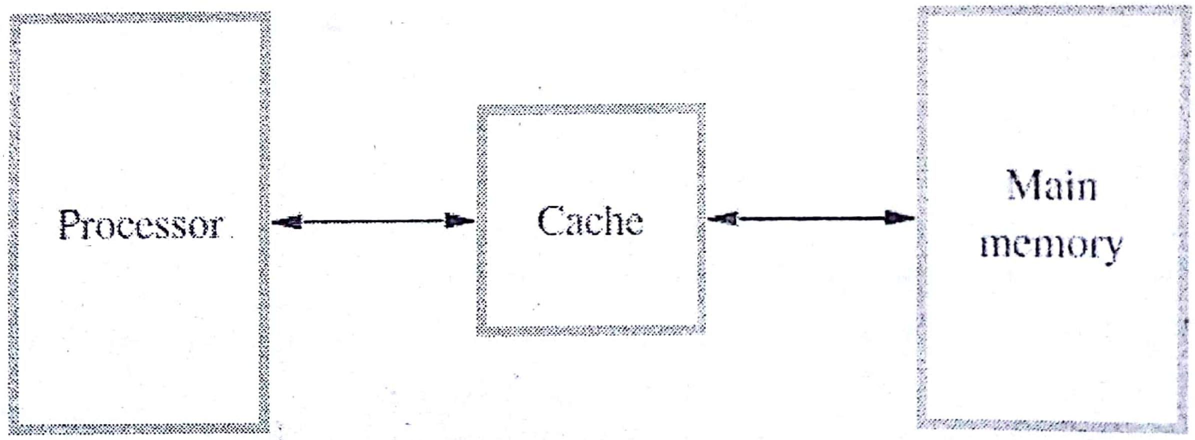


- The controller accepts a complete address and the R/W signal from the processor, under control of a request signal which indicates that a memory access operation is needed.
- The controller then forwards the row and column portions of the address to the memory and generates the RAS and CAS signals.

- Thus, the controller provides the RAS-CAS timing, in addition to its address multiplexing function.
- It also sends the R/W and CS signals to the memory.
- The CS signal is usually active low.
- Data lines are connected directly between the processor and the memory.
- Note that the clock signal is needed in SDRAM chips.

4. CACHE MEMORY:

- The cache is a small and very fast memory, interposed between the processor and the main memory.
- Its purpose is to make the main memory appear to the processor to be much faster than it actually is.
- The effectiveness of this approach is based on a property of computer programs called *locality of reference.*
- Consider the arrangement, the use of a cache memory is depicted as,



- When the processor issues a Read request, the contents of a block of memory words containing the location specified are transferred into the cache.
- Subsequently, when the program references any of the locations in this block, the desired contents are read directly from the cache.
- Usually, the cache memory can store a reasonable number of blocks at any given time, but this number is small compared to the total number of blocks in the main memory.
- The correspondence between the main memory blocks and those in the cache is specified by a *mapping function.*

Hit Ratio:

- The performance of cache memory is measured in terms of a quantity called **hit ratio.**
- 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 cache, it is in main memory then it counts as a **miss.**

- The ratio of the number of hits to the total CPU references to memory is called hit ratio.

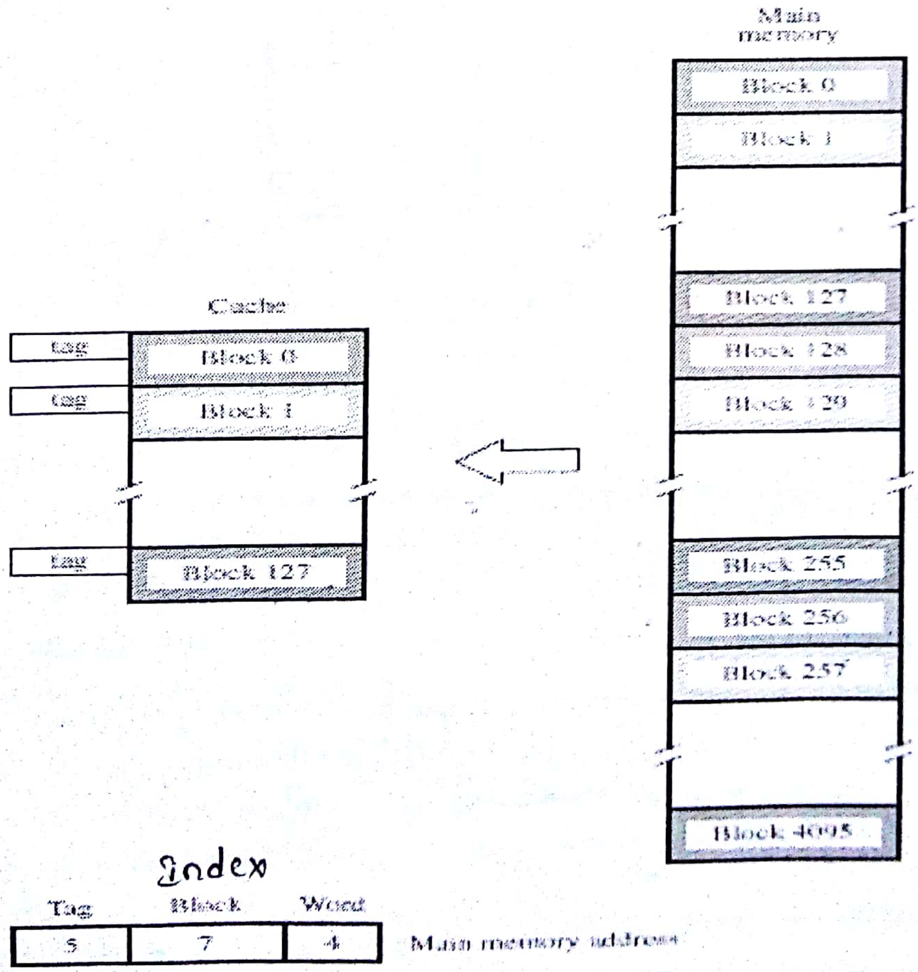
  Hit Ratio = Hit / (Hit + Miss)

***) MAPPING FUNCTIONS:**

- There are several possible methods for determining where memory blocks are placed in the cache.
- Consider a cache consisting of 128 blocks of 16 words each, for a total of 2048 (2K) ⇒ 2^{10} words, and assume that the main memory is addressable by a 16-bit address.
- The main memory has64K words, which we will view as 4K blocks of 16 words each.

i) Direct Mapping:

- The simplest way to determine cache locations in which to store memory blocks is the direct-mapping technique.
- In this technique, block j of the main memory maps onto block j modulo 128 of the cache, is depicted as,



$2^k = 2\times2\times2\times2\times2\times2\times2$

Word
Index
Tag

Word → No. of bits required to identify a particular word with in

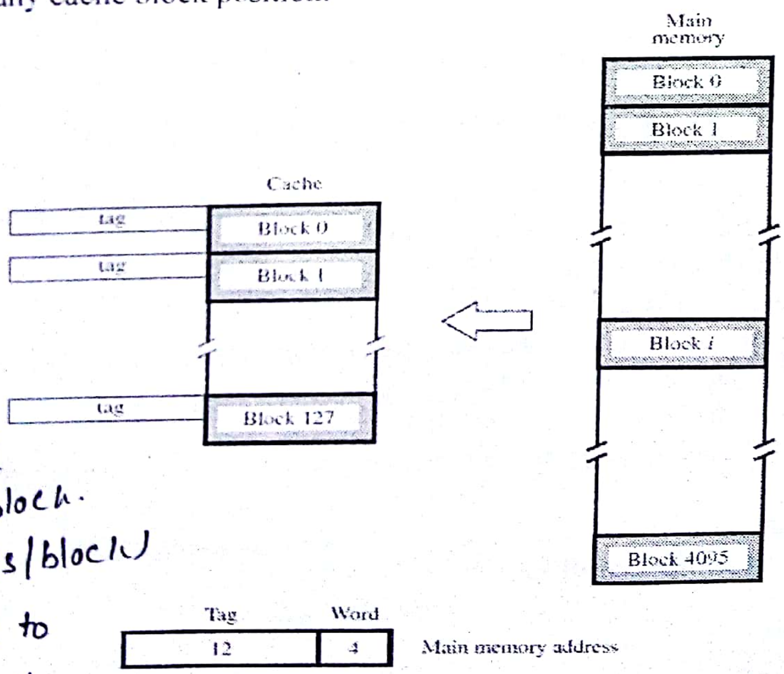
$= \log_2$ (No. of

Index → No. of bits req to identify the block no memory when a M.M

$= \log_2$ (No. of

- Thus, whenever one of the main memory blocks 0, 128, 256, . . . is loaded into the cache, it is stored in cache block0.
- Blocks 1, 129, 257, . . . are stored in cache block 1, and so on.
- The memory address can be divided into *three* fields
- The low-order 4 bits select one of 16 words in a block.
- When a new block enters the cache, the 7-bit cache block field determines the cache position in which this block must be stored.
- The high-order 5 bits of the memory address of the block are stored in 5 tag bits associated with its location in the cache.
- The tag bits identify which of the 32 main memory blocks mapped into this cache position is currently resident in the cache.
- As execution proceeds, the 7-bit cache block field of each address generated by the processor points to a particular block location in the cache.
- The high-order 5 bits of the address are compared with the 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 first be read from the main memory and loaded into the cache.

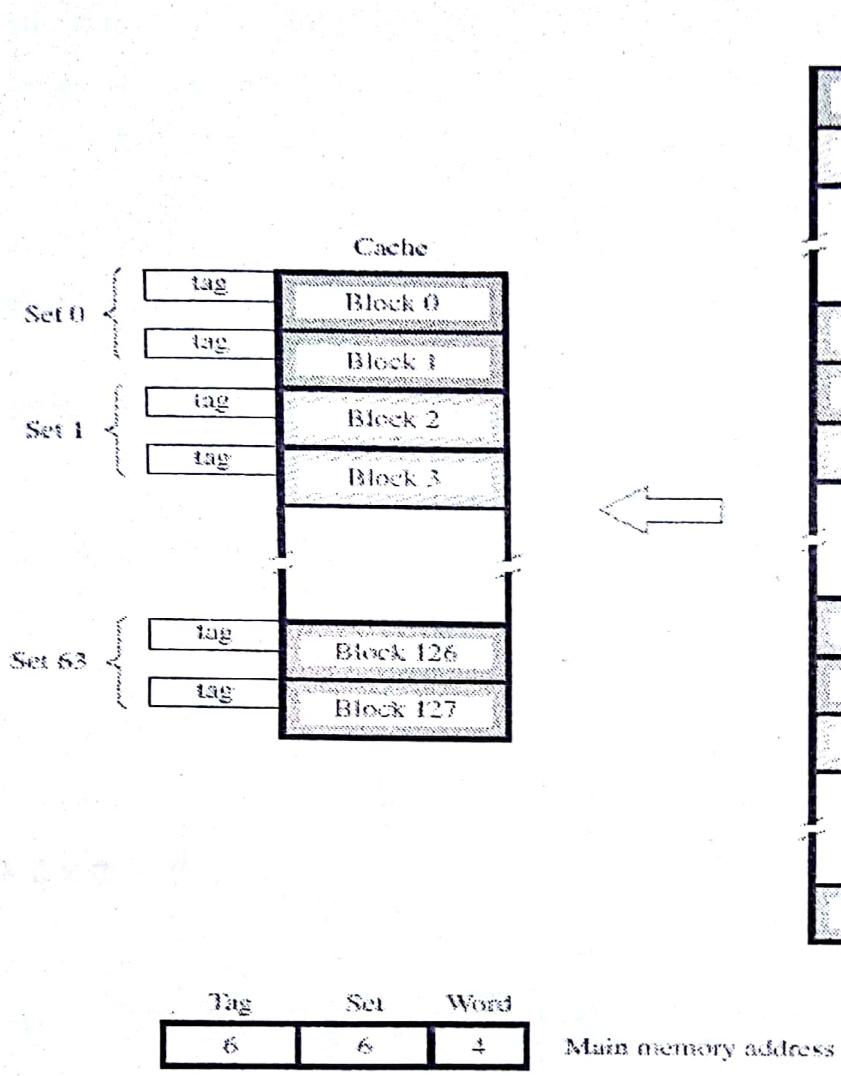
ii) Associative Mapping:
- The most flexible mapping method, in which a main memory block can be placed into any cache block position.



- In this case, 12 tag bits are required to identify a memory block when it is resident 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 the *associative-mapping technique*.
- It gives complete freedom in choosing the cache location in which to place the memory block, resulting in a more efficient use of the space in the cache.
- When a new block is brought into the cache, it replaces (ejects) an existing block only if the cache is full.
- The complexity of an associative cache is higher than that of a direct-mapped cache, because of the need to search all 128 tag patterns to determine whether a given block is in the cache.
- To avoid a long delay, the tags must be searched in parallel.
- A search of this kind is called an *associative search.*

iii) Set-Associative Mapping:

- Another approach is to use a combination of the direct- and associative-mapping techniques.
- 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 a specific set.
- Hence, the contention problem of the direct method is eased by having a few choices for block placement.
- At the same time, the hardware cost is reduced by decreasing the size of the associative search.
- An example of this set-associative-mapping technique is shown in Figure 8.18 for a cache with two blocks per set.
- In this case, memory blocks 0, 64, 128, . . . , 4032 map into cache set 0, and they can occupy either of the two block positions within this set.
- Having 64 sets means that the6-bit set field of the address determines which set of the cache might contain the desired block.
- The tag field of the address must then be associatively compared to the tags of the two blocks of the set to check if the desired block is present.
- This two-way associative search is simple to implement.
- The number of blocks per set is a parameter that can be selected to suit the requirements of a particular computer.
- For the main memory and cache sizes, four blocks per set can be accommodated by a 5-bit set field, eight blocks per set by a 4-bit set field, and so on.

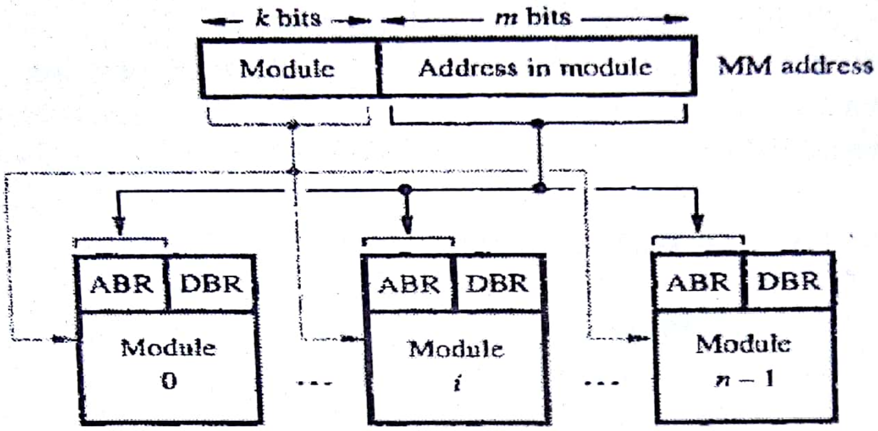


- The extreme condition of 128 blocks per set requires no set bits and corresponds to the fully-associative technique, with 12 tag bits.
- The other extreme of one block per set is the direct-mapping method.
- A cache that has k blocks per set is referred to as a k-way set-associative cache.

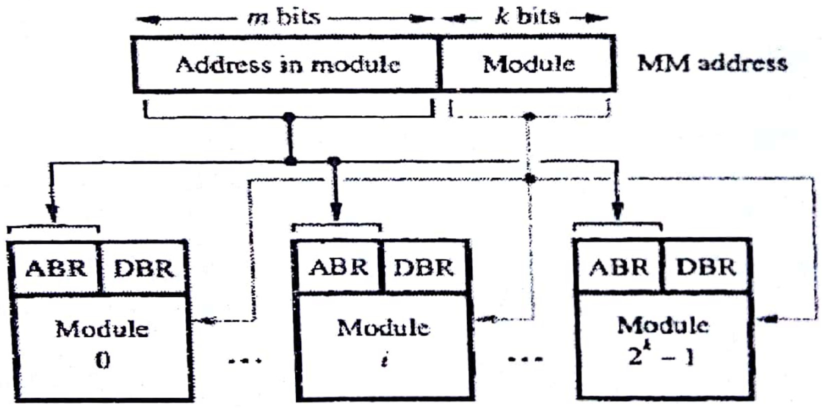
5. MEMORY INTERLEAVING:

- If the main memory of a computer is structured as a collection of physically separate modules, each with its own address buffer register (ABR) and data buffer register (DBR), memory access operations may proceed in more than one module at the same time.
- Thus the aggregate rate of transmission of words to and from the main memory system can be increased.
- How individual addresses are distributed over the modules is critical in determining the average number of modules that can be kept busy as computations proceed.

- Two methods of address layout, i.e., Addressing Multiple-module Memory Systems are depicted as,



(a) Consecutive words in a module



(b) Consecutive words in consecutive modules

- In the first case, the memory address generated by the processor is decoded.
- The high-order k-bits name one of n modules, and the low-order m-bits name a particular word in that module.
- When consecutive locations are accessed, as happens when a block of data is transferred to a cache, only one module is involved.
- At the same time, however, devices with direct memory access (DMA) ability may be accessing information in other memory modules.
- The second and more effective way to address the modules is called Memory Interleaving.

- The lower-order k-bits of the memory address select a module, and the high-order m-bits name a location within that module.
- Thus any component of the system that generates requests for access to consecutive memory locations can keep several modules busy at any one time
- This results in both faster accesses to a block of data and higher average utilization of the memory system as a whole.

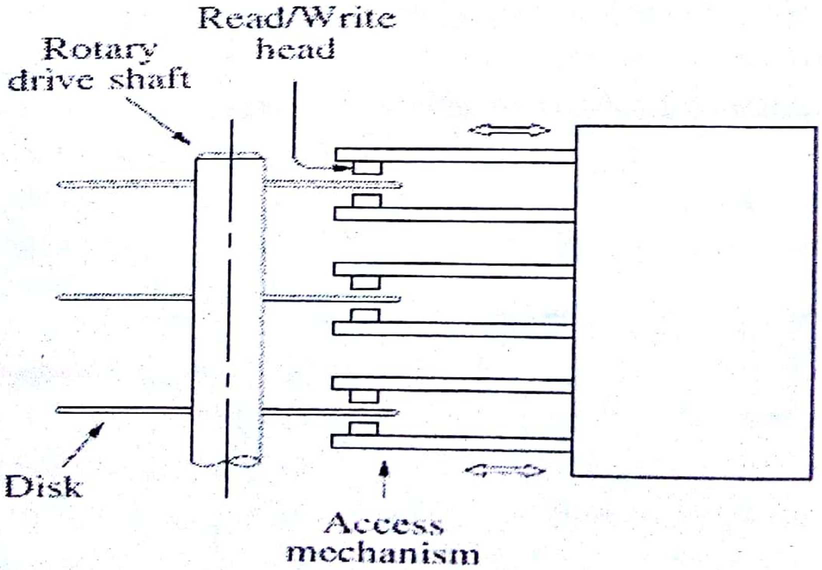
To implement the interleaved structure, there must be 2^k modules, otherwise there will be gaps of non-existing locations in the memory address space.

6. SECONDARY STORAGE:

- The *semiconductor memories* discussed in the previous sections cannot be used to provide all of the storage capability needed in computers.
- Their main limitation is the cost per bit of stored information.
- The large storage requirements of most computer systems are economically realized in the form of magnetic and optical disks, which are usually referred to as *secondary storage devices*.

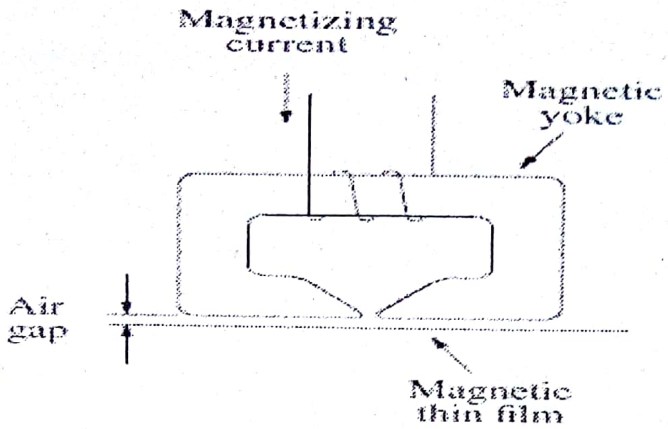
i) Magnetic Hard Disks:

- The storage medium in a magnetic-disk system consists of one or more disk platters mounted on a common spindle.
- A thin magnetic film is deposited on each platter, usually on both sides.
- The assembly is placed in a drive that causes it to rotate at a constant speed.
- The magnetized surfaces move in close proximity to read/write heads, is depicted as.



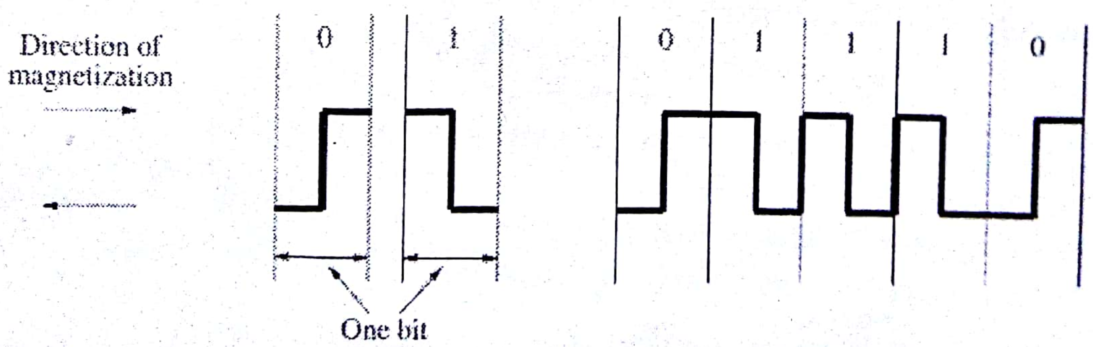
(a) Mechanical structure

- Data are stored on concentric tracks, and the read/write heads move radially to access different tracks.
- Each read/write head consists of a magnetic yoke and a magnetizing coil, is depicted as,



(b) Read/Write head detail

- Digital information can be stored on the magnetic film by applying current pulses of suitable polarity to the magnetizing coil.
- This causes the magnetization of the film in the area immediately underneath the head to switch to a direction parallel to the applied field.
- The same head can be used for reading the stored information.
- In some early designs, a clock was stored on a separate track, on which a change in magnetization is forced for each bit period.
- Using the clock signal as a reference, the data stored on other tracks can be read correctly.
- The modern approach is to combine the clocking information with the data.
- Several different techniques have been developed for such encoding.
- One simple scheme is known as *phase encoding* or *Manchester encoding*.

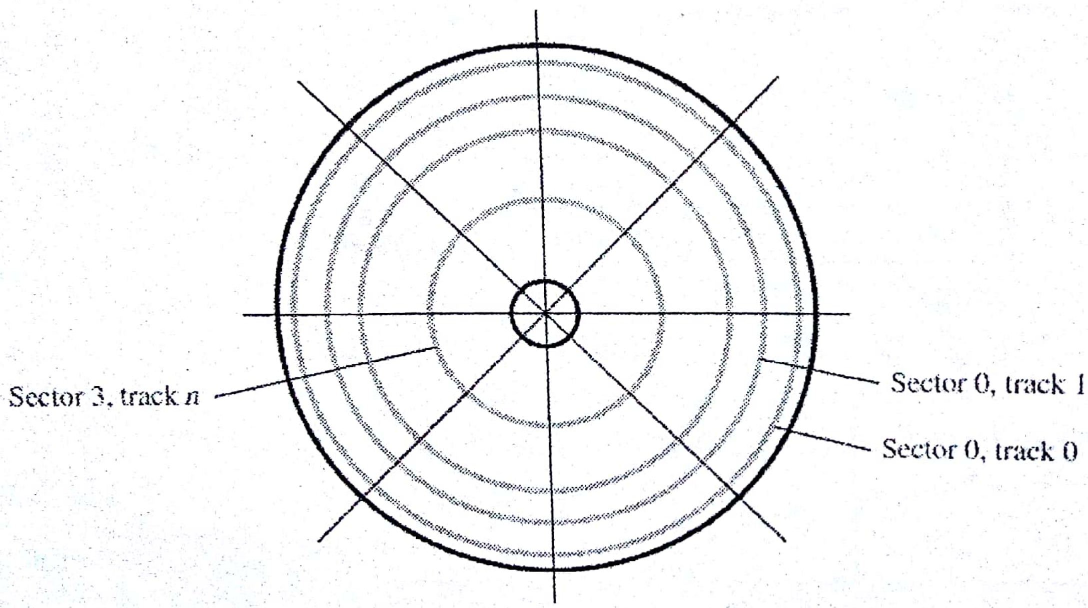


(c) Bit representation by phase encoding

- In this scheme, changes in magnetization occur for each data bit.
- Clocking information is provided by the change in magnetization at the midpoint of each bit period.
- The drawback of Manchester encoding is its poor bit-storage density.
- The space required to represent each bit must be large enough to accommodate two changes in magnetization.
- In most modern disk units, the disks and the read/write heads are placed in a sealed, air-filtered enclosure.
- This approach is known as Winchester technology.

Organization and Accessing of Data on a Disk:

- The organization of data on one surface of a disk is depicted as,



- Each surface is divided into concentric *tracks*, and each track is divided into *sectors*.
- The set of corresponding tracks on all surfaces of a stack of disks forms a logical *cylinder*.
- All tracks of a cylinder can be accessed without moving the read/write heads.
- Data are accessed by specifying the surface number, the track number, and the sector number.
- Read and Write operations always start at sector boundaries.
- Data bits are stored serially on each track.
- Each sector may contain 512 or more bytes.
- The data are preceded by a *sector header* that contains identification (addressing) information used to find the desired sector on the selected track.
- Following the data, there are additional bits that constitute an *error-correcting code* (ECC).

- The ECC bits are used to detect and correct errors that may have occurred in writing or reading the data bytes.
- There is a small *inter-sector gap* that enables the disk control circuitry to distinguish easily between two consecutive sectors.
- An unformatted disk has no information on its tracks.
- The formatting process writes markers that divide the disk into tracks and sectors.
- During this process, the disk controller may discover some sectors or even whole tracks that are defective.
- The disk controller keeps a record of such defects and excludes them from use.
- The formatting information comprises sector headers, ECC bits, and inter-sector gaps.
- The capacity of a formatted disk, after accounting for the formatting information overhead, is the proper indicator of the disk's storage capability.
- After formatting, the disk is divided into logical partitions.

Access Time:

- There are two components involved in the time delay between the disk receiving an address and the beginning of the actual data transfer.
- The first, called the *seek time*, is the time required to move the read/write head to the proper track.
- The second component is the *rotational delay*, also called *latency time*, which is the time taken to reach the addressed sector after the read/write head is positioned over the correct track.
- On average, this is the time for half a rotation of the disk.
- The sum of these two delays is called the disk *access time*.

Data Buffer/Cache:

- A disk drive is connected to the rest of a computer system using some standard interconnection scheme, such as SCSI or SATA.
- The interconnection hardware is usually capable of transferring data at much higher rates than the rate at which data can be read from disk tracks.
- An efficient way to deal with the possible differences in transfer rates is to include a *data buffer* in the disk unit.
- The buffer is a semiconductor memory, capable of storing a few megabytes of data.

Disk Controller:

- Operation of a disk drive is controlled by a *disk controller* circuit, which also provides an interface between the disk drive and the rest of the computer system.
- One disk controller may be used to control more than one drive.
- A disk controller that communicates directly with the processor contains a number of registers that can be read and written by the operating system.

- The OS initiates the transfers by issuing Read and Write requests, which entail loading the controller's registers with the necessary addressing and control information.
- Typically, this information includes:

Main memory address: The address of the 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.

Word count: The number of words in the block to be transferred.

- On the disk drive side, the controller's major functions are:

Seek: Causes the disk drive to move the read/write head from its current position to the desired track.

Read: Initiates a Read operation, starting at the address specified in the disk address register. Data read serially from the disk are assembled into words and placed into the data buffer for transfer to the main memory. The number of words is determined by the word count register.

Write: Transfers data to the disk, using a control method similar to that for Read operations.

Error checking: Computes the error correcting code (ECC) value for the data read from a given sector and compares it with the corresponding ECC value read from the disk.

Floppy Disks:

- The disks discussed above are known as hard or rigid disk units.
- *Floppy disks* are smaller, simpler, and cheaper disk units that consist of a flexible, removable, plastic *diskette* coated with magnetic material.
- The diskette is enclosed in a plastic jacket, which has an opening where the read/write head can be positioned.
- A hole in the center of the diskette allows a spindle mechanism in the disk drive to position and rotate the diskette.
- The main feature of floppy disks is their low cost and shipping convenience.
- However, they have much smaller storage capacities, longer access times, and higher failure rates than hard disks.
- In recent years, they have largely been replaced by CDs, DVDs, and flashcards as portable storage media.

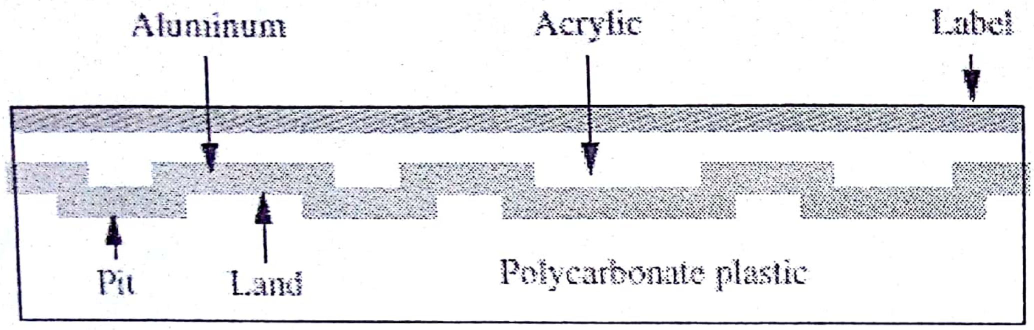
RAID Disk Arrays:

- Processor speeds have increased dramatically.
- At the same time, access times to disk drives are still on the order of milliseconds, because of the limitations of the mechanical motion involved.
- One way to reduce access time is to use multiple disks operating in parallel.

- In 1988, researchers at the University of California-Berkeley proposed such a storage system.
- They called it RAID, for Redundant Array of Inexpensive Disks.(Since all disks are now inexpensive, the acronym was later reinterpreted as Redundant Array of Independent Disks.)
- Using multiple disks also makes it possible to improve the reliability of the overall system.
- Different configurations were proposed, and many more have been developed since.
- The basic configuration, known as RAID 0, is simple. A single large file is stored in several separate disk units by dividing the file into a number of smaller pieces and storing these pieces on different disks. - This is called *data striping*.
- Various RAID configurations form a hierarchy, with each level in the hierarchy providing additional features.
- For example, RAID 1 is intended to provide better reliability by storing identical copies of the data on two disks rather than just one.
- RAID systems are available from many manufacturers for use with a variety of operating systems.

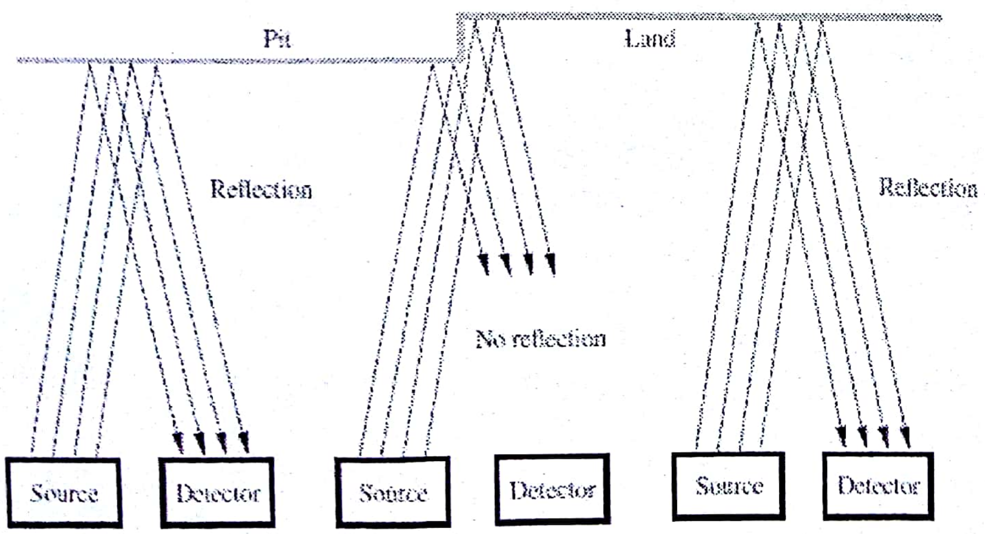
ii) Optical Disks:
- Storage devices can also be implemented using optical means.
- The familiar compact disk (CD), used in audio systems, was the first practical application of this technology.
- The first generation of CDs was developed in the mid-1980s by the Sony and Philips companies.
- Initially, CDs were designed to hold up to 75 minutes, requiring a total of about 3 × 109 bits (3 gigabits) of storage.

CD Technology:
- The optical technology that is used for CD systems makes use of the fact that laser light can be focused on a very small spot.
- A laser beam is directed onto a spinning disk, with tiny indentations arranged to form a long spiral track on its surface.
- The indentations reflect the focused beam toward a photo detector, which detects the stored binary patterns.
- The laser emits a coherent light beam that is sharply focused on the surface of the disk.
- Across-section of a small portion of a CD is depicted as,



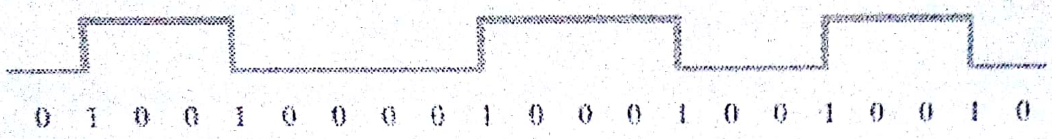
(a) Cross-section

- The bottom layer is made of transparent polycarbonate plastic, which serves as a clear glass base.
- The surface of this plastic is programmed to store data by indenting it with *pits*.
- The unintended parts are called *lands*.
- A thin layer of reflecting aluminum material is placed on top of a programmed disk.
- The aluminum is then covered by a protective acrylic.
- Finally, the topmost layer is deposited and stamped with a label.
- The total thickness of the disk is 1.2 mm, almost all of it contributed by the polycarbonate plastic.
- What happens as the laser beam scans across the disk and encounters a transition from a pit to a land is depicted as,



(b) Transition from pit to land

- Several transitions between lands and pits is depicted as,



(c) Stored binary pattern

CD-ROM:

- Since CDs store information in a binary form, they are suitable for use as a storage medium in computer systems.
- The main challenge is to ensure the integrity of stored data.
- Because the pits are very small, it is difficult to implement all of the pits perfectly.
- In audio and video applications, some errors in the data can be tolerated, because they are unlikely
to affect the reproduced sound or image in a perceptible way.
- However, such errors are not acceptable in computer applications.
- Since physical imperfections cannot be avoided, it is necessary to use additional bits to provide error detection and correction capability.
- The CDs used to store computer data are called *CD-ROMs*, because, like semiconductor ROM chips, their contents can only be read.
- Stored data are organized on CD-ROM tracks in the form of blocks called *sectors*.
- There are several different formats for a sector.
- One format, known as Mode 1, uses 2352-byte sectors.
- There is a 16-byte header that contains a synchronization field used to detect the beginning of the sector and addressing information used to identify the sector.
- This is followed by 2048 bytes of stored data.
- The basic speed, known as 1X, is 75 sectors per second.
- This provides a data rate of 153,600 bytes/s (150Kbytes/s), using the Mode 1 format.
- Higher speed CD-ROM drives are identified in relation to the basic speed.

CD-Recordable:

- A new type of CD was developed in the late 1990s on which data can be easily recorded by a computer user.
- It is known as CD-Recordable (CD-R).
- A shiny spiral track covered by an organic dye is implemented on a disk during the manufacturing process.
- Then, a laser in a CD-R drive burns pits into the organic dye.
- The burned spots become opaque.
- They reflect less light than the shiny areas when the CD is being read.
- This process is irreversible, which means that the written data are stored permanently.
- Unused portions of a disk can be used to store additional data at a later time.

CD-Rewritable:

- The most flexible CDs are those that can be written multiple times by the user.
- They are known as CD-RWs (CD-ReWritables).
- The basic structure of CD-RWs is similar to the structure of CD-Rs.
- Instead of using an organic dye in the recording layer, an alloy of silver, indium, antimony, and tellurium is used.
- This alloy has interesting and useful behavior when it is heated and cooled.

DVD Technology:

- The success of CD technology and the continuing quest for greater storage capability has led to the development of DVD (Digital Versatile Disk) technology.
- The first DVD standard was defined in 1996 by a consortium of companies, with the objective of being able to store a full-length movie on one side of a DVD disk.
- The physical size of a DVD disk is the same as that of CDs. The disk is 1.2 mm thick, and it is 120 mm in diameter.
- Using these improvements leads to a DVD capacity of 4.7 Gbytes.
- Access times for DVD drives are similar to CD drives.
- Rewritable versions of DVD devices have also been developed, providing large storage capacities.

iii) Magnetic Tape Systems:

- Magnetic tapes are suited for off-line storage of large amounts of data.
- They are typically used for backup purposes and for archival storage.
- Magnetic-tape recording uses the same principle as magnetic disks.
- The main difference is that the magnetic film is deposited on a very thin 0.5- or 0.25-inch wide plastic tape.
- Seven or nine bits (corresponding to one character) are recorded in parallel across the width of the tape, perpendicular to the direction of motion.
- A separate read/write head is provided for each bit position on the tape, so that all bits of a character can be read or written in parallel.
- One of the character bits is used as a parity bit.
- Data on the tape are organized in the form of *records* separated by gaps is depicted as.

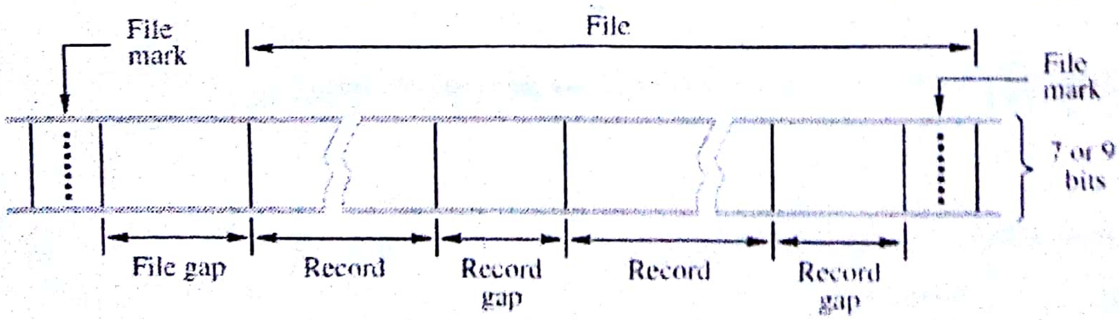


Figure 8.30 Organization of data on magnetic tape.

- The controller of a magnetic tape drive enables the execution of a number of control commands in addition to read and write commands,
  - Rewind tape
  - Rewind and Unload tape
  - Erase Tape
  - Write tape mark
  - Forward space one record
  - Backspace one record
  - Forward space one file
  - Backspace one file

SHORT ANSWER QUESTIONS:

1. Explain a) RAM b) EEPROM

2. Explain about Flash Memory or Flash ROM

3. What are the applications and roles of Flash Memory?

Roles of Flash Memory:

- Different picture images can be saved in digital cameras with the help of flash memory.
- By using flash memory in MP3 players, it is possible to store sound information.
- It contains a special type of software that helps it in operating hand-held computers and cell phones.

Applications of Flash Memory:

- Handheld Computers
- Digital Cameras
- Cell phones
- MP3 Players

4. Give the differences between EEPROM and FLASH MEMORY

5. Explain a) Read Miss b) Read Hit c) Hit Ratio

6. What is Cache Memory? Mention its advantages.

- The cache is a small and very fast memory, interposed between the processor and the main memory.

Advantages of Cache Memory:

- It is faster than Main Memory.
- It takes less time for execution.
- It stores data for less duration, i.e., for temporary storage.
- It does not require system bus for transferring data.

7. Explain about locality of reference

Locality of Reference:

- The execution time of the computer programs mainly depends on the execution of routines present in them.
- These routines are nothing but simple loops, nested loops, or branch instructions.
- Hence in any computer program, these instructions are of primary focus, which gets repeatedly executed all the time and rest of the instructions are executed rarely.
- This behavior of computer programs are called as Locality of Reference.

8. Give the advantages and disadvantages of Direct Mapping.

Advantages of Direct Cache:

- It is very simple to implement.
- It can access any block address directly

Disadvantages of Direct Cache:

- It is not effective
- It causes contention of data

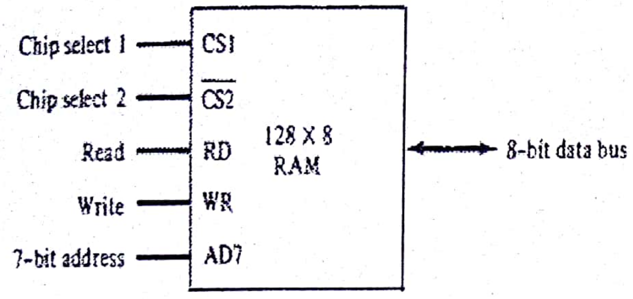
9. What is Mapping Function?

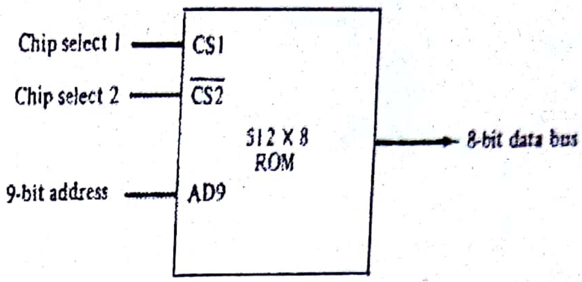
Mapping function:

- The correspondence between the main memory blocks and those in the cache is specified by a *mapping function*
- There are three mapping functions used
  - i) Associative Mapping
  - ii) Direct Mapping
  - iii) Set-associative Mapping

10. Draw the block diagrams for RAM and ROM

RAM and ROM block diagrams:





11. Discuss two update policies that a cache can use.

Update policies in Cache Memory:

- The two update policies of cache memory are,
  - i) Write-through
  - ii) Write-back

i) Write-through:

- In write-through cache, whenever a write is made to cache, immediately the write is also made to main memory.

ii) Write-back:

- In write-back cache, when the written block is about to be replaced, then the cache block is written entirely to main memory.

12. Explain about Optical Disks.

13. Explain about Magnetic Disks.

14. Explain about Magnetic Tapes.

15. Explain about Associative Memory or Content Addressable Memory.

Associative Memory:

- It is also known as *content addressable memory (CAM).*
- It is a memory chip in which each bit position can be compared.
- In this the content is compared in each bit cell which allows very fast table lookup. Since the entire chip can be compared, contents are randomly stored without considering addressing scheme.
- These chips have less storage capacity than regular memory chips.

16. Explain about Virtual Memory.

Virtual Memory:

- An imaginary memory area supported by some operating systems (for example, Windows) in conjunction with the hardware.
- You can think of virtual memory as an alternate set of memory addresses.
- Programs use these *virtual addresses* rather than real addresses to store instructions and data.
- When the program is actually executed, the virtual addresses are converted into real memory addresses.
- The purpose of virtual memory is to enlarge the *address space*, the set of addresses a program can utilize.
- For example, virtual memory might contain twice as many addresses as main memory.
- A program using all of virtual memory, therefore, would not be able to fit in main memory all at once.
- Nevertheless, the computer could execute such a program by copying into main memory those portions of the program needed at any given point during execution.
- To facilitate copying virtual memory into real memory, the operating system divides virtual memory into *pages*, each of which contains a fixed number of addresses.
- Each page is stored on a disk until it is needed.
- When the page is needed, the operating system copies it from disk to main memory, translating the virtual addresses into real addresses.

LONG ANSWER QUESTIONS:

1. What is meant by Memory? Explain how data transfer takes place between memory and processor.

2. Explain about a) Memory Access Methods b) SRAM c) DRAM

3. What is ROM? Explain different types of ROM

OR

Explain about a) PROM b) EPROM c) EEPROM d) FLASH ROM