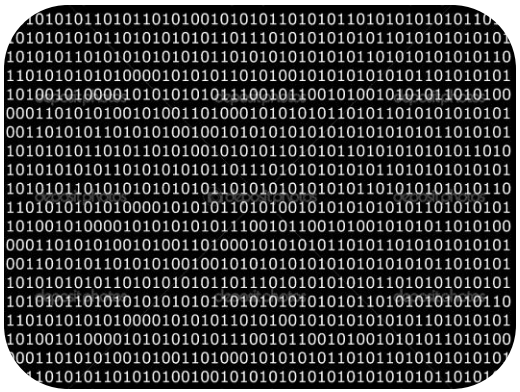
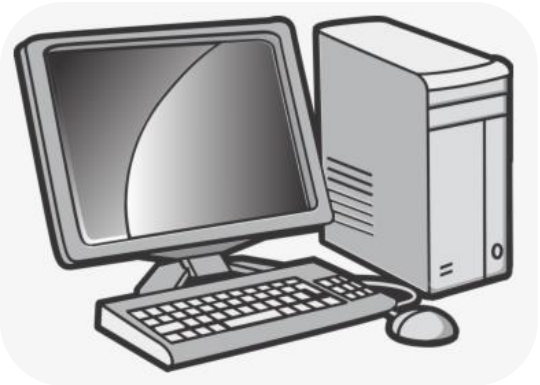
AMIRAJ

COLLEGE OF ENGINEERING & TECHNOLOGY

Unit-9
Memory Organization





Subject:- COA **Code:-3140707**	**Prepared by:** **Asst. Prof. S.Y.JOSHI** **(CSE Department,ACET)**	AMIRAJ COLLEGE OF ENGINEERING & TECHNOLOGY

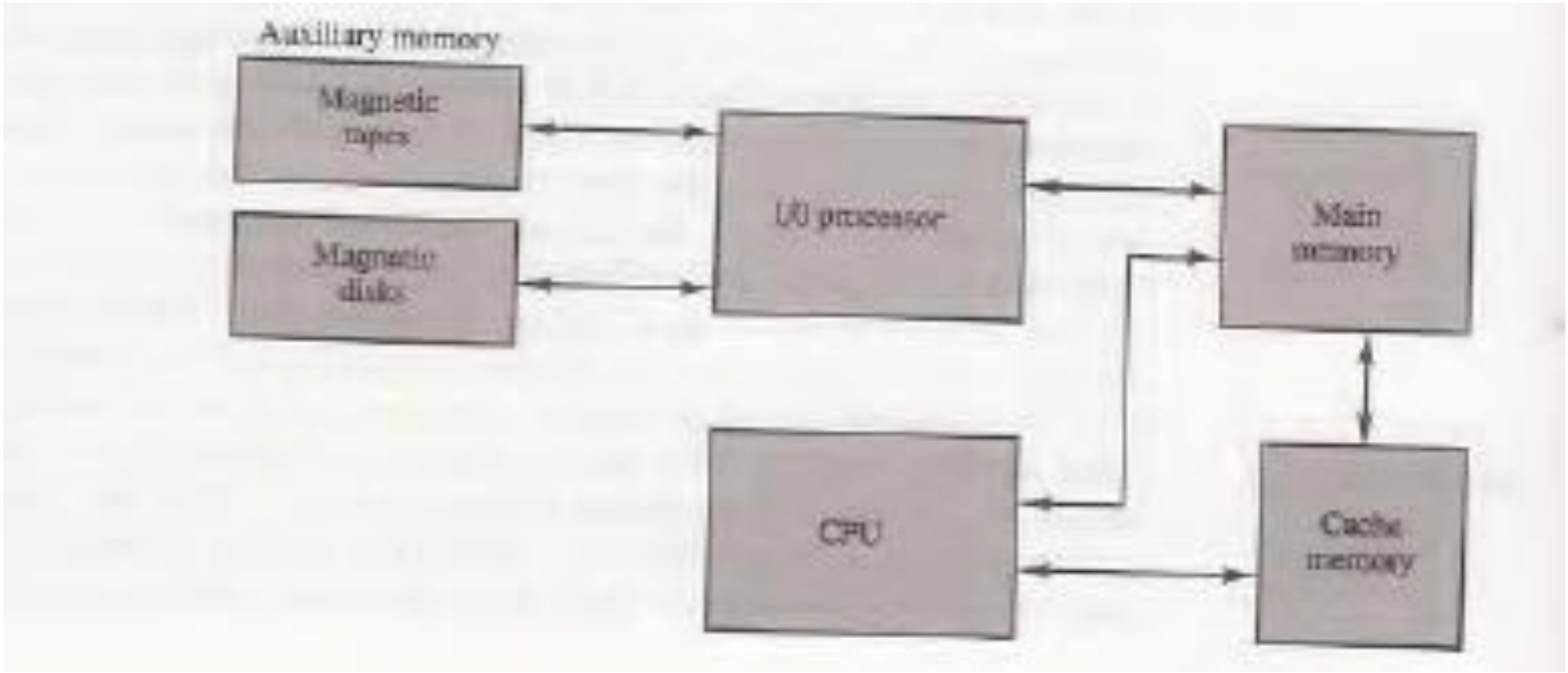
Topics to be covered

- Memory Hierarchy,
- Main Memory,
- Auxiliary Memory,
- Associative Memory,
- Cache Memory,
- Virtual Memory
- Memory Management

Memory Heirarchy

- The memory unit that communicates directly with the CPU is called the main memory.
- Devices that provide backup storage are called auxiliary memory. The most common auxiliary memory devices used in computer systems are magnetic disks and tapes.
- The total memory capacity of a computer can be visualized as being a hierarchy of components. The memory hierarchy system consists of all storage devices employed in a computer system from the slow but high-capacity auxiliary memory to a relatively faster main memory to an even smaller Cache memory.
- Cache is high speed memory used to increase the speed of processing by making current programs and data available to CPU at a rapid rate.

Memory Heirarchy

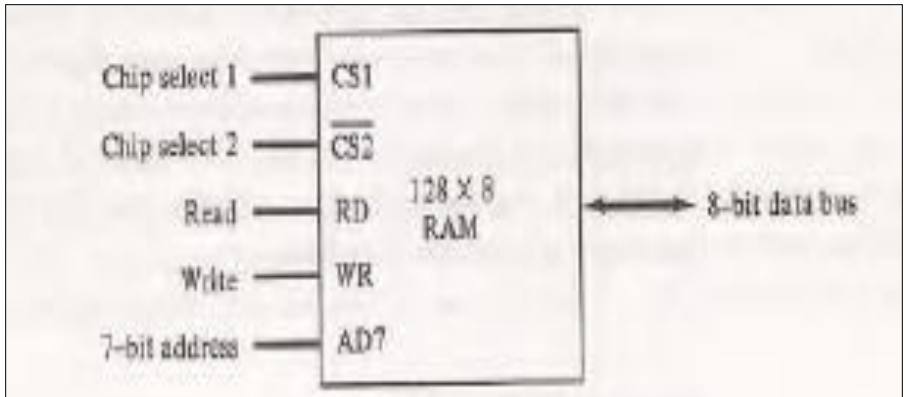


Main Memory

- The main memory is the central storage unit in a computer system. It is a relatively large and fast memory used to store programs and data during the computer operation.
- The principal technology used for the main memory is based on semiconductor integrated circuits.
- Integrated circuit RAM chips are available in two possible operating modes, static and dynamic.
- A portion of the memory may be constructed with ROM chips.
- the ROM portion of main memory is needed for storing an initial program called a bootstrap loader.

Main memory

Typical RAM chip.

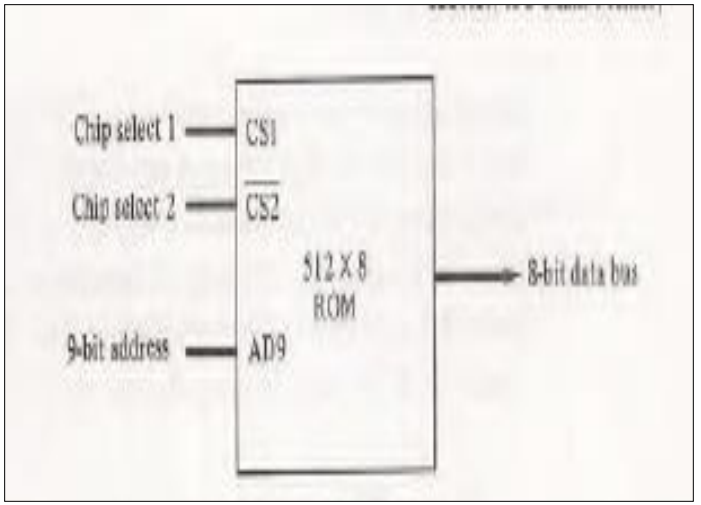


(a) Block diagram

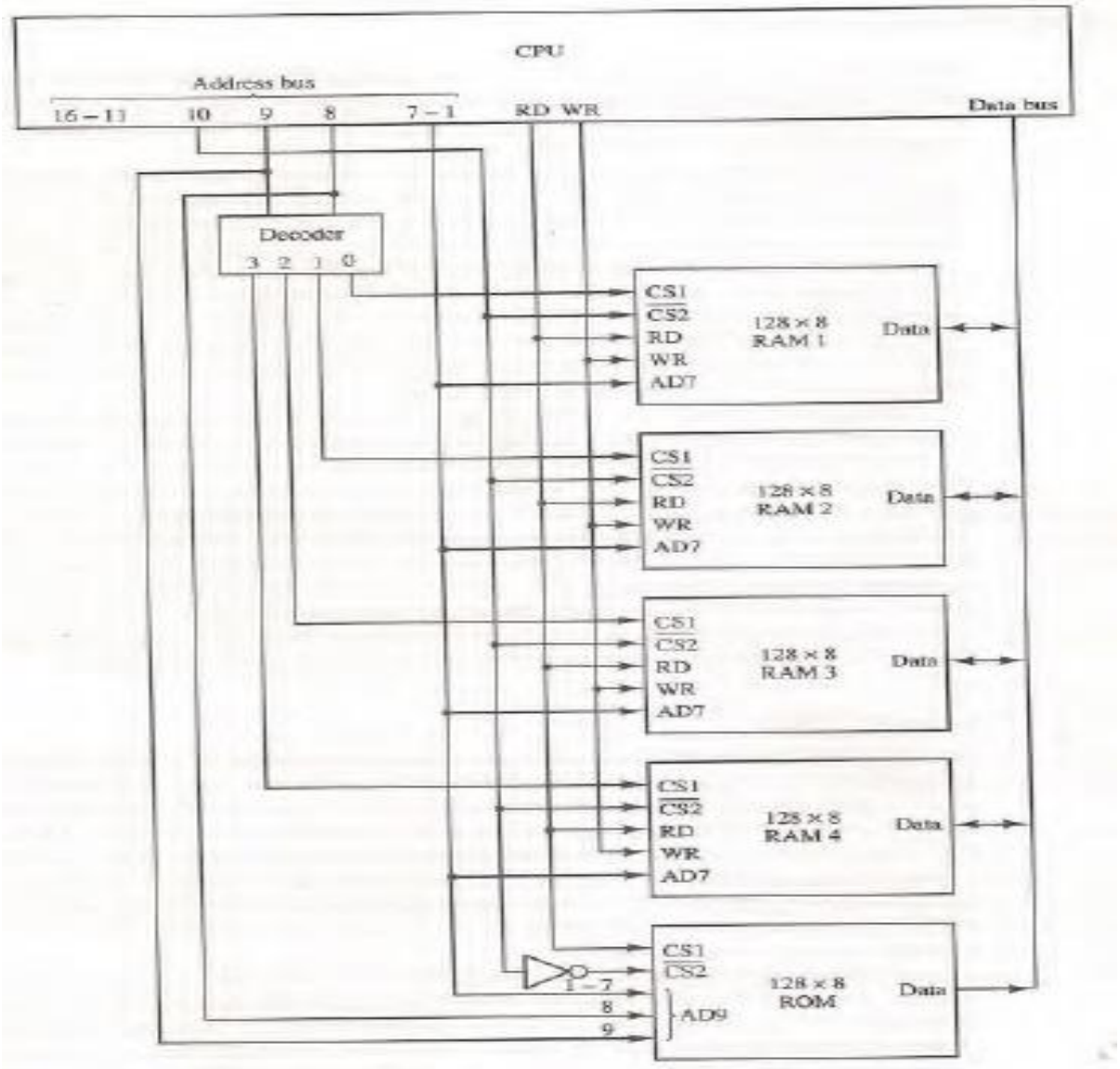
CS1	CS2	RD	WR	Memory function	State of data bus
0	0	×	×	Inhibit	High-impedance
0	1	×	×	Inhibit	High-impedance
1	0	0	0	Inhibit	High-impedance
1	0	0	1	Write	Input data to RAM
1	0	1	×	Read	Output data from RAM
1	1	×	×	Inhibit	High-impedance

(b) Function table

Typical ROM chip.



Main memory

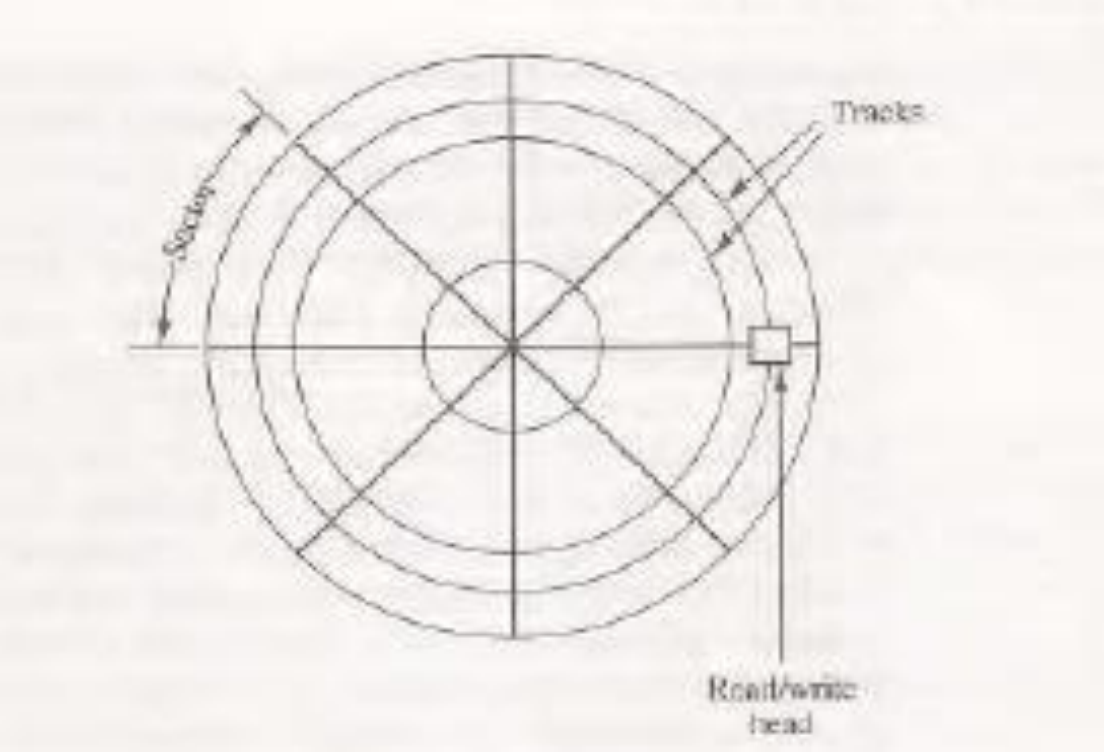


Memory connection to CPU

Auxiliary Memory

- Auxiliary memory devices used in computer systems are magnetic disks and tapes, magnetic drums, magnetic bubble memory, and optical disks.
- A magnetic disk is a circular plate constructed of metal or plastic coated with magnetized material. Often both sides of the disk are used and several disks may be stacked on one spindle with read/write heads available on each surface. All disks rotate together at high speed and are not stopped or started for access purposes. Bits are stored in the magnetized surface in spots along concentric circles called tracks. The tracks are commonly divided into sections called sectors. In most systems, the minimum quantity of information which can be transferred is a sector. The subdivision of one disk surface into tracks and sectors is shown in Fig
- Magnetic tape transport consists of the electrical, mechanical, and electronic components to provide the parts and control mechanism for a magnetic-tape unit. The tape itself is a strip of plastic coated with a magnetic recording medium. Magnetic tape units can be stopped, started to move forward or in reverse, or can be rewound.

Auxiliary Memory
Magnetic Disk

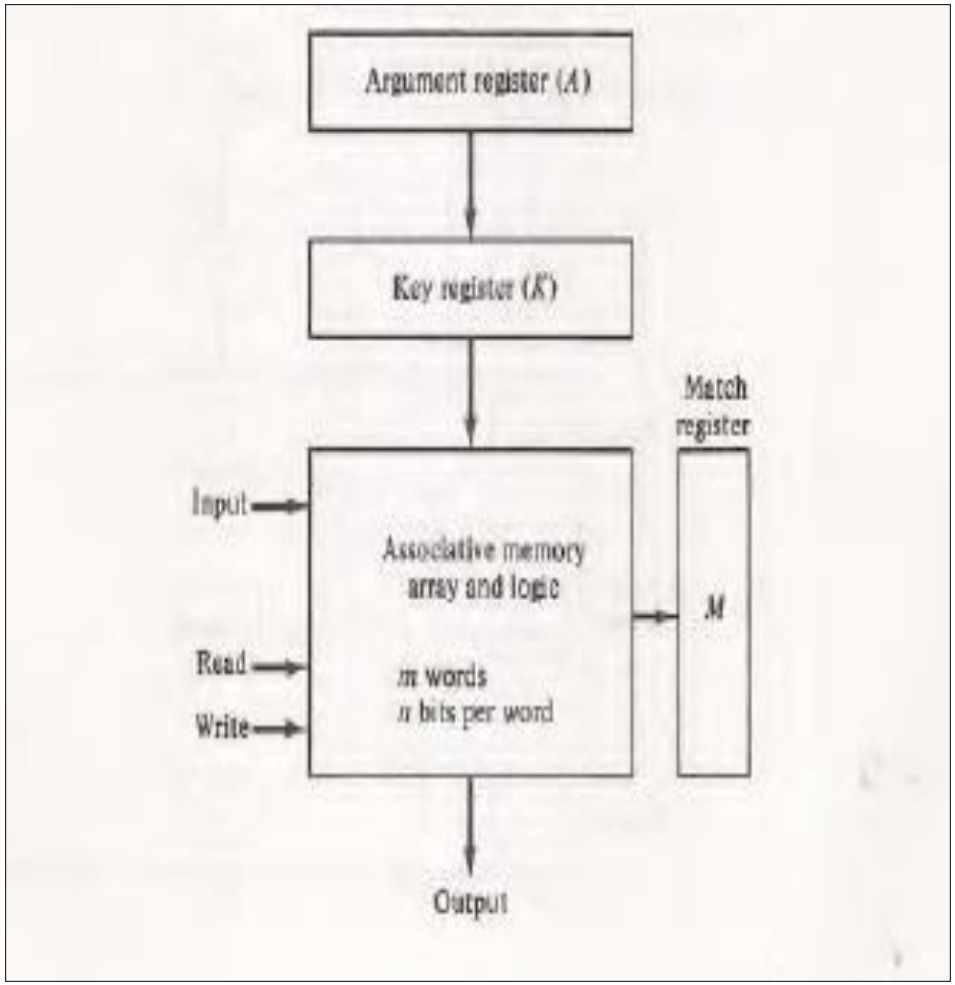


AMIRAJ
COLLEGE OF ENGINEERING & TECHNOLOGY

Associative Memory

- Many data-processing applications require the search of items in a table stored in memory.
- An assembler program searches the symbol address table in order to extract the symbol's binary equivalent. An account number may be searched in a file to determine the holder's name and account status. The established way to search a table is to store all items where they can be addressed in sequence. The search procedure is a strategy for choosing a sequence of addresses, reading the content of memory at each address, and comparing the information read with the item being searched until a match occurs. The number of accesses to memory depends on the location of the item and the efficiency of the search algorithm. Many search algorithms have been developed to minimize the number of accesses while searching for an item in a random or sequential access memory. The time required to find an item stored in memory can be reduced considerably if stored data can be identified for access by the content of the data itself rather than by an address. A memory unit accessed by content is called content addressable an associative memory or content addressable memory (CAM). This type of memory is accessed simultaneously and in parallel on the basis of data content rather than by specific address or location. When a word is written in an associative memory, no address is given.

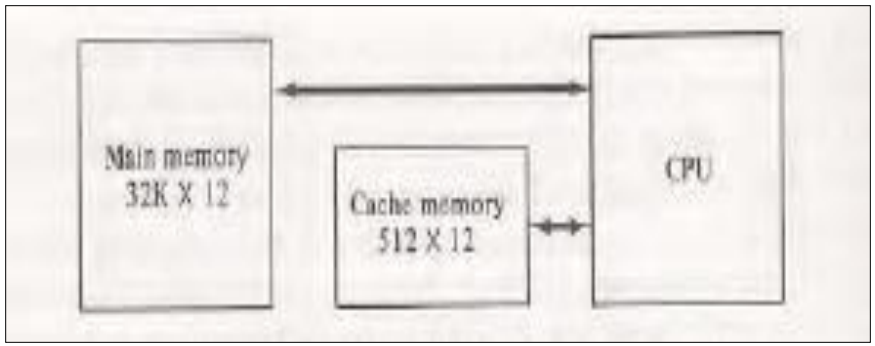
Associative Memory



- The block diagram of an associative memory is shown in Fig.
- It consists of 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 compared in parallel with the content of the argument register. The words that match the bits of the argument Register set a corresponding bit in the match register.

CACHE MEMORY

- Analysis of a large number of typical programs has shown that the references to memory at any given interval of time tend to be confined within a few localized areas in memory. This phenomenon is known as the property of locality of reference.
- The fundamental idea of cache organization is that by keeping the most frequently accessed instructions and data in the fast cache memory, age memory access time will approach the access time of the cache. in the fast cache memory.



CACHE MEMORY

- The performance of cache memory is frequently 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 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
- The average memory access time of a computer system can be improved considerably by use of a cache.
- The transformation of data from main memory to cache memory is referred to as a mapping process.
- Three types of mapping procedures are of practical interest when considering the organization of cache memory:

1. Associative mapping
2. Direct mapping
3. Set-associative mapping

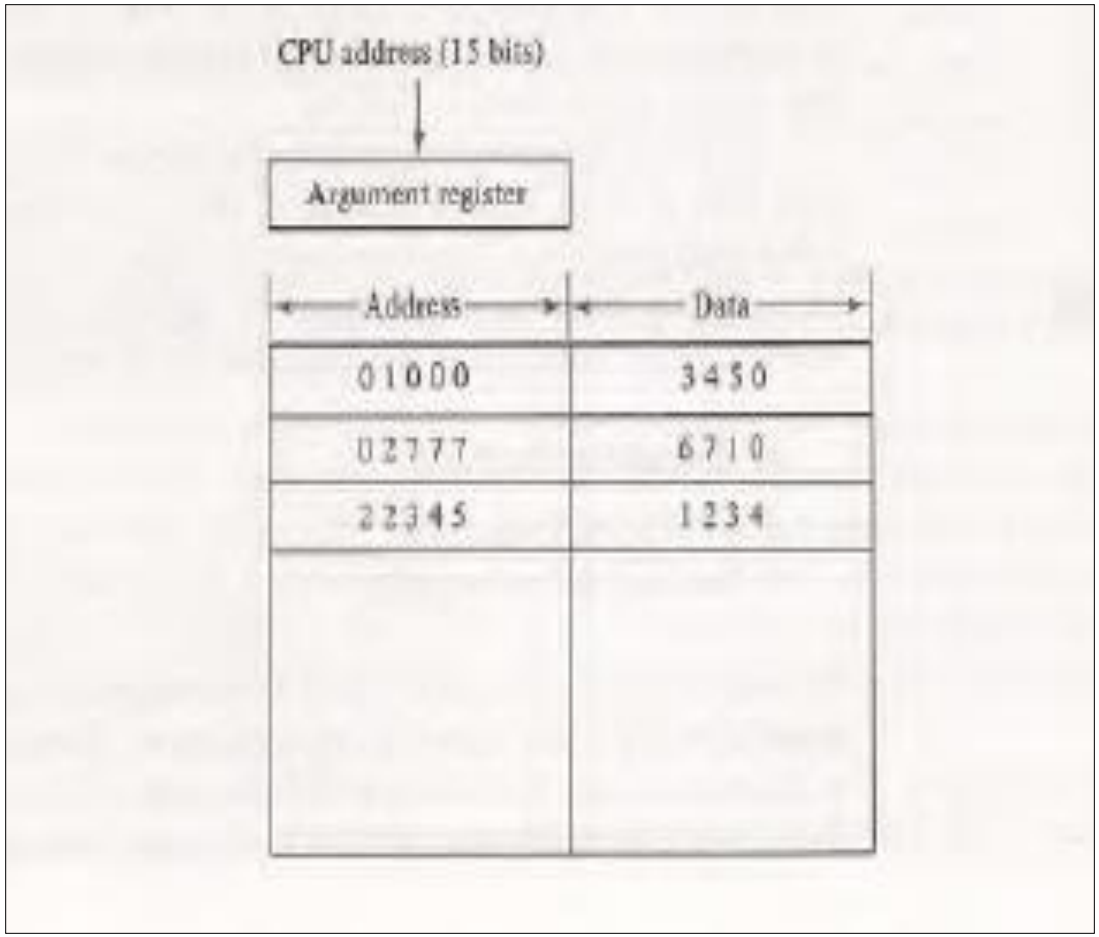
CACHE MEMORY

Associative mapping

- The fastest and most flexible cache organization uses an associative memory. The associative memory stores
- Both the address and content (data) of the memory word. This permits any location in cache to store any word from main memory.
- The diagram shows three words presently stored in the cache. The address value of 15 bits is shown as a five-digit octal number and its corresponding 12-bit word is shown as a four-digit octal number.
- A CPU address of 15 bits is placed in the argument register and the associative memory is searched for a matching address.
- If the address is found, the corresponding 12-bit data is read and sent to the CPU. If no match occurs, the main memory is accessed for the word
- first-in first-out (FIFO) replacement policy

CACHE MEMORY

- Associative mapping in cache



AMIRAJ
COLLEGE OF ENGINEERING & TECHNOLOGY

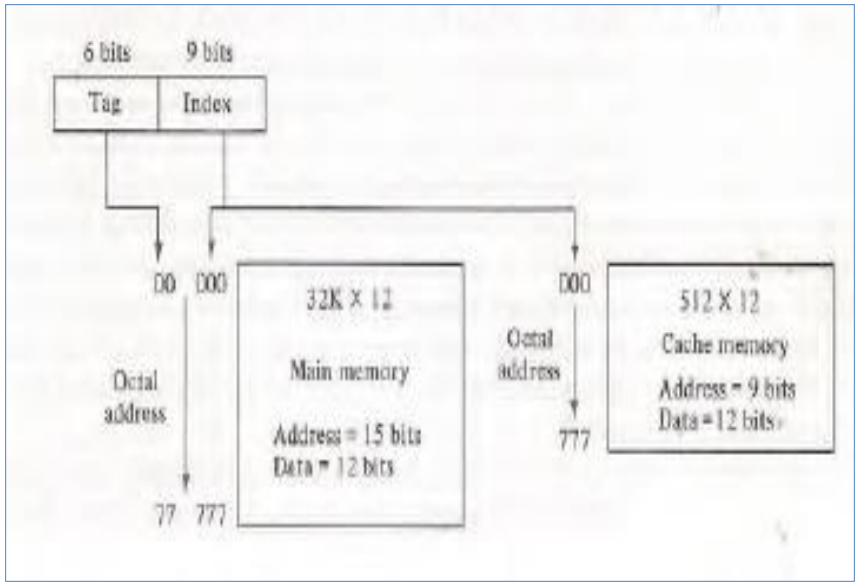
CACHE MEMORY

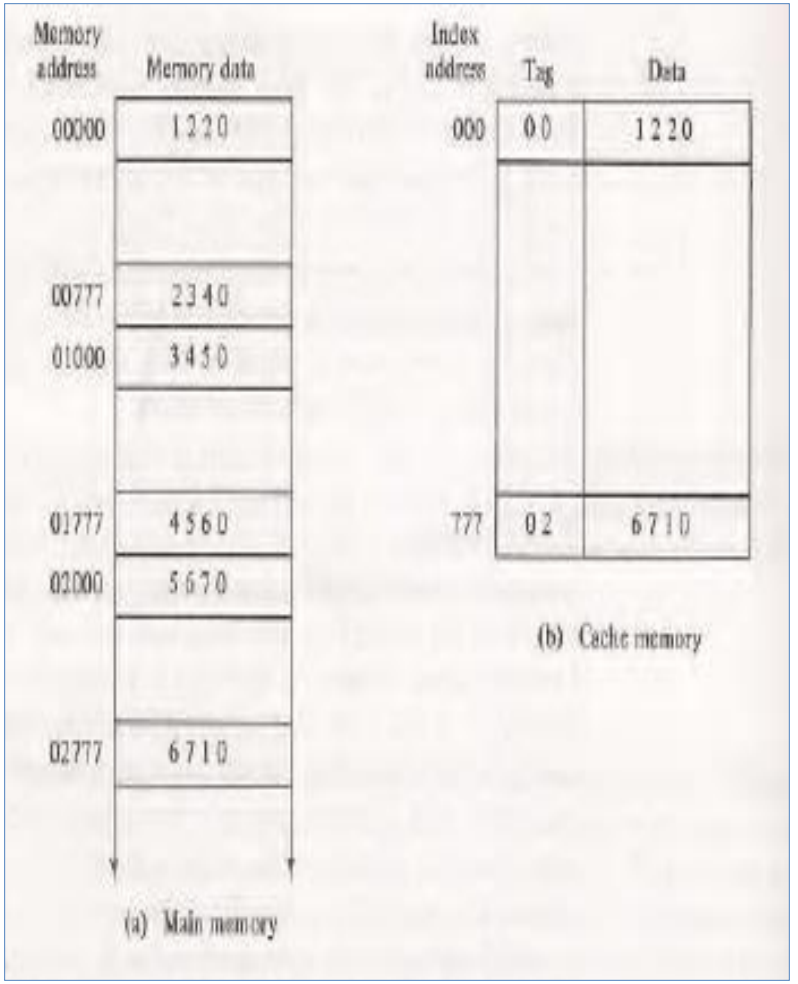
Direct mapping

- Associative memories are expensive compared to random-access memories because of the added logic associated with each cell. The possibility of using a random-access memory for the cache is investigated in Fig.
- The CPU address of 15 bits is divided into two fields. The nine least significant bits constitute the index field and the remaining six bits form the tag field.
- The figure shows that main memory needs an address that includes both the tag and the index bits.
- The number of bits in the index field is equal to the number of address bits required to access the cache memory.

CACHE MEMORY

Direct mapping

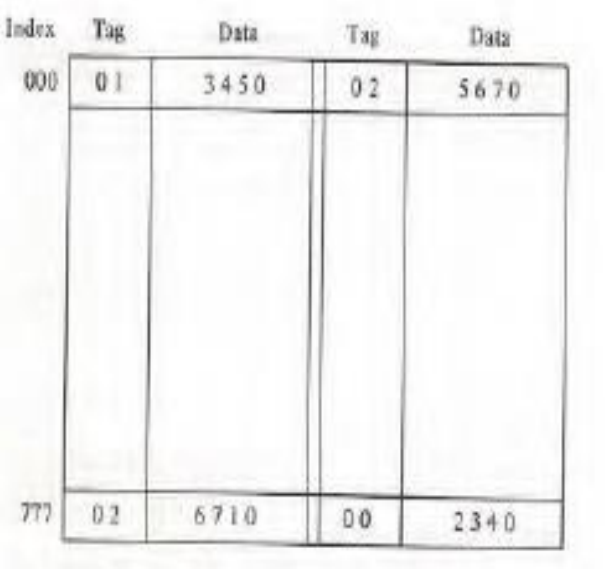




AMIRAJ COLLEGE OF ENGINEERING & TECHNOLOGY

CACHE MEMORY

- Set-associative mapping
- Set-associative mapping, is an improvement over the direct mapping organization in that each word of cache can store two or more words of memory under the same index address.
- Each data word is stored together with its tag and the number of tag-data items in one word of cache is said to form a set.
- An example of a set-associative cache organization for a set size of two is shown in Fig.



Index	Tag	Data	Tag	Data
000	01	3450	02	5670
777	02	6710	00	2340

a set-associative cache of set size k will accommodate k words of main memory in each word of cache.

CACHE MEMORY

Cache Initialization

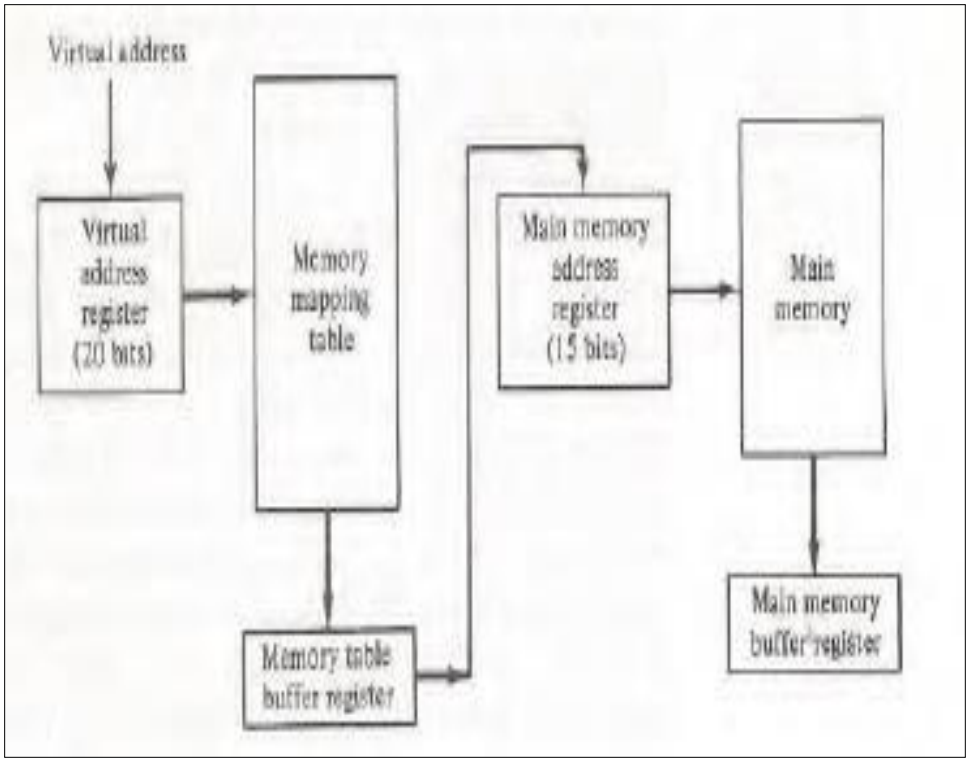
- The cache is initialized when power is applied to the computer or when the main memory is loaded with a complete set of programs from auxiliary memory.
- After initialization the cache is considered to be empty, but in effect it contains some non valid data.
- It is customary to include with each word in cache a valid bit to indicate whether or not the word contains valid data.
- The cache is initialized by clearing all the valid bits to 0. The valid bit of a particular cache word is set to 1 the first time this word is loaded from main memory and stays set unless the cache has to be initialized again.

VIRTUAL MEMORY

- Virtual memory is a concept used in some large computer systems that permit the user to construct programs as though a large memory space were available, equal to the totality of auxiliary memory.
- Each address that is referenced by the CPU goes through an address mapping from the so-called virtual address to a physical address in main memory.
- Virtual memory system provides a mechanism for translating program-generated addresses into correct main memory locations. This is done dynamically, while programs are being executed in the CPU.
- The translation or mapping is handled automatically by the hardware by means of a mapping table.

VIRTUAL MEMORY

- An address used by a programmer will be called a virtual address,
- And the set of such addresses the address space.
- An address in main memory is called a location or physical address.
- The set of such locations is called the memory space.



The mapping table may be stored in a separate memory as shown in fig

MEMORY MANAGEMENT

- In a multiprogramming environment where many programs reside in memory it becomes necessary to move programs and data around the memory, to vary the amount of memory in use by a given program, and to prevent a program from changing other programs. The demands on computer memory brought about by multiprogramming have created the need for a memory management system.
- A **memory management system** is a collection of hardware and software procedures for managing the various programs residing in memory. The memory management software is part of an overall operating system available in many computers.

MEMORY MANAGEMENT

The basic components of a memory management unit are:

1. A facility for dynamic storage relocation that maps logical memory references into physical memory addresses
2. A provision for sharing common programs stored in memory by different users
3. Protection of information against unauthorized access between users and preventing users from changing operating system functions

References

- **Images , descriptive Tables , from Computer System Architecture, Morris Mano, 3rd edition Prentice Hall**

- **Note: These pdf/ppt notes are for purpose of teaching aids to classroom/online sessions study, and in no case imply for GTU syllabus or GTU exam. For GTU syllabus or exam related preparation, one may, however will need to attend college/online lectures and refer books given by GTU in their official syllabus.**