Chapter 4: Memory Organization

4.1. Memory Hierarchy

The hierarchical arrangement of storage in current computer architectures is called the memory hierarchy. The memory unit is an essential component in any digital computer since it is needed for storing programs and data.

- The memory unit that communicates directly with the CPU is called the Main Memory (or Primary memory).
- Devices that provide backup storage are called auxiliary Memory (or Secondary).

Only programs and data currently needed by the processor reside in Main memory. All other information is stored in Auxiliary memory and transferred to main memory when needed.

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 and faster cache memory accessible to the high speed processing logic.

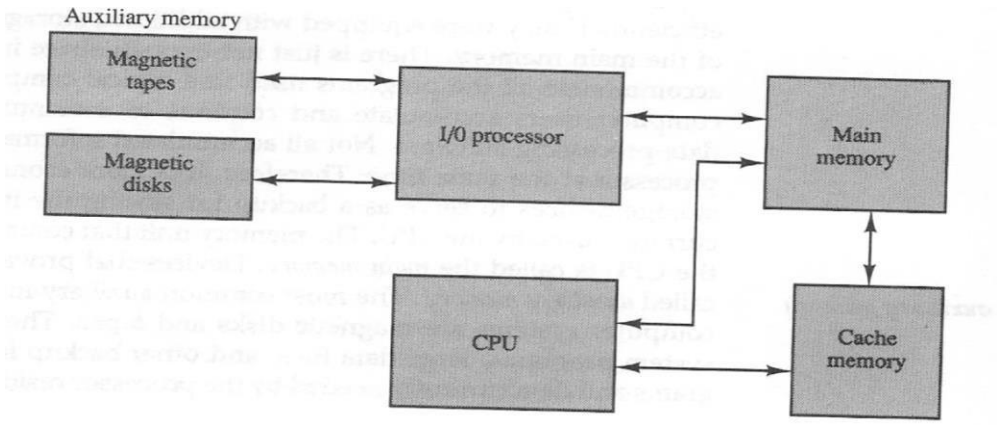


Figure: 4.1: Memory Hierarchy in a computer system

When programs not residing in main memory are 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 for currently used programs and data.

The cache memory is employed in computer systems to compensate for the speed differential between main memory access time and processor logic.

The cache is used for storing segments of programs currently being executed in the CPU and temporary data frequently needed in the present calculations.

By making programs and data available at a rapid rate, it is possible to increase the performance rate of the computer using cache memory.
While the I/O processor manages data transfers between auxiliary memory and main memory, the cache organization is concerned with the transfer of information between main memory and CPU. Thus each is involved with a different level in the memory hierarchy is economics.

4.2 Main Memory

It 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.

The static RAM (SRAM) consists essentially of internal flip flops that store the binary information. The stored information remains valid as long as power is applied to the unit.

The dynamic RAM (DRAM) stores the binary information in the form of electric charges that are applied to capacitors. The capacitors are provided inside the chip by MOS transistors. The stored charge on the capacitors tends to discharge with time and the capacitors must be periodically recharged by refreshing the dynamic memory. Refreshing is done by cycling through the words every few milliseconds to restore the decaying charge. The dynamic RAM offers reduced power consumption and larger storage capacity in a single memory chip. The static RAM is easier to use and has shorter read and write cycles.

The ROM portion of main memory is needed for storing an initial program called a **Bootstrap Loader**. The bootstrap loader is a program whose function is to start the computer software operating when power is turned on. The contents of ROM remain unchanged after power is turned off and on again. The start-up of a computer consists of turning the power on and starting the execution of an initial program. Thus when power is turned on, the hardware of the computer sets the program counter to the first address of the bootstrap loader. The bootstrap program loads a portion of the operating system from disk to main memory and control then transferred to the operating system, which prepares the computer for general use.

RAM and ROM chips:

RAM and ROM chips are available in a variety of sizes. If the memory needed for the computer is larger than the capacity of one chip, it is necessary to combine a number of chips to form the required memory size.
A RAM chip is better suited for communication with the CPU if it has one or more control inputs that select the chip only when needed. Another common feature is a bidirectional data bus that allows the transfer of data either from memory to CPU during a read operation or from CPU to memory during a write operation.

Memory address map:

The designer of a computer system must calculate the amount of memory required for the particular application and assign it to either RAM or ROM.

The interconnection between memory and processor is then established from knowledge of the size of memory needed and the type of RAM and ROM chips available.

The addressing of memory can be established by means of a table that specifies the memory address assigned to each chip. The table, called a **memory address map**, is a pictorial representation of assigned address space for each chip in the system.

4.3 Cache Memory

If the active portions of the program and data are placed in a fast small memory, the average memory access time can be reduced, thus reducing the total execution time of the program. Such a fast small memory is referred to as cache memory.

- The cache is the fastest component in the memory hierarchy and approaches the speedof CPU component.
- It is placed between the CPU and main memory.
- When 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.
- A block of words containing the one just accessed is then transferred from main memory to 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.** Otherwise, it is a **miss.**

Hit ratio = hit / (hit+miss)

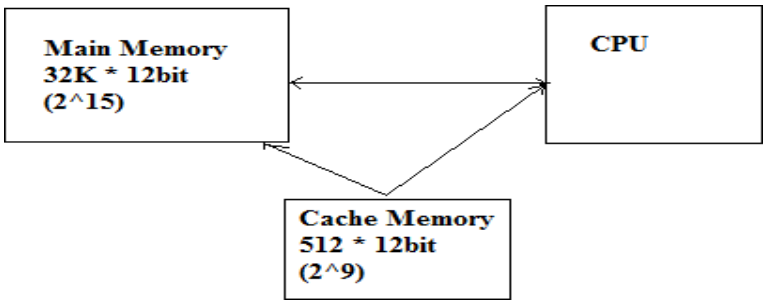
- The basic characteristic of cache memory is its fast access time,
- Therefore, very little or no time must be wasted when searching the words in the cache

The transformation of data from main memory to cache memory is referred to as a **mapping process**.

There are three types of mapping:

- Associative mapping
- Direct mapping
- Set-associative mapping

- To help understand the mapping procedure, we have the following example:



I. Associative mapping

- The fastest and most flexible cache organization uses an associative memory.
- The associative memory stores both the address and data of the memory word.
- This permits any location in cache to store ant word from main memory.

II. Direct Mapping

- Associative memory is expensive compared to RAM.
- In general case, there are 2^k words in cache memory and 2^n words in main memory (in our case, k=9, n=15)
- The n bit memory address is divided into two fields: k-bits for the index and n-k bits for the tag field

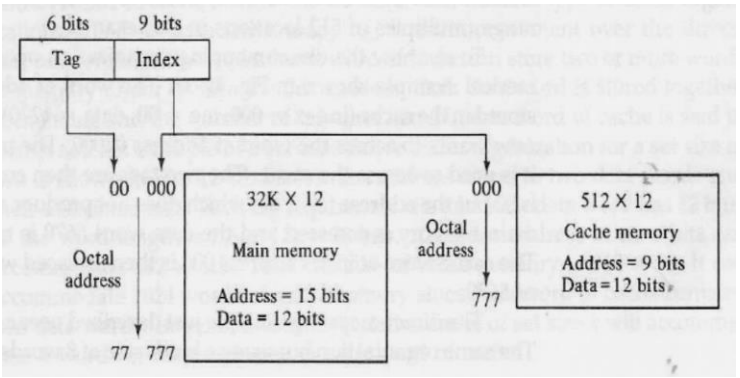
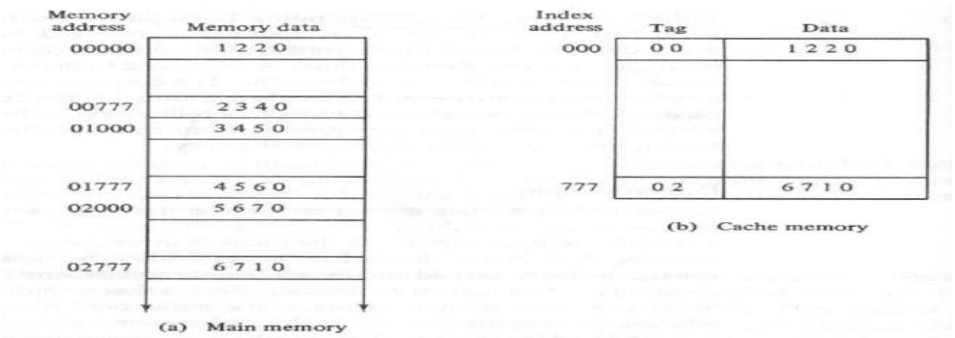


Figure 4.3: Direct memory mapping

Addressing relationships between main and cache memories



The internal organization of the words in the cache memory is as shown in the above figure. Each word in cache consists of the data word and its associated tag. When a new word is first brought into the cache, the tag bits are stored alongside the data bits. When the CPU generates a memory request, the index field is used for the address to access the cache.

The tag field of the CPU address is compared with the tag in the word read from the cache. If the two tags match, there is a hit and the desired data word is in cache. If there is no match, there is a miss and the required word is read from main memory. It is then the cache together with the new tag, replacing the previous value.

The disadvantage of direct mapping is that the hit ratio can drop considerably if two or more words whose addresses have the same index but different tags are accessed repeatedly.

III. Set-Associative Mapping

- The disadvantage of direct mapping is that two words with the same index in their address but with different tag values cannot reside in cache memory at the same time.
- Set-Associative Mapping is an improvement over the direct-mapping in that each word of cache can store two or more word of memory under the same index address.
- Each index address refers to two data words and their associated tags
- Each tag requires six bits and each data word has 12 bits, so the word length is 2*(6+12) = 36 bits

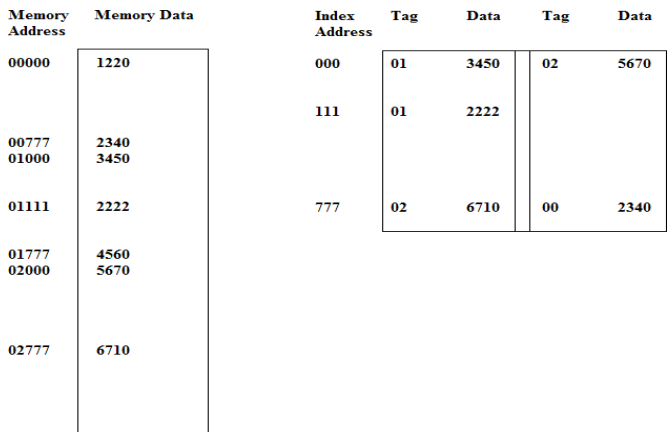


Figure 4.5: Set-Associative Mapping

The words stored at address 01000 and 02000 of main memory are stored in cache memory inindex address 000. Similarly, the words at addresses 02777 and 00777 are stored in cache at index address 777. When the CPU generates a memory request, the index value of the address is used to access the cache. The tag field of the CPU address is then compared withboth tags in the cache to determine if a match occurs.

The hit ratio will improve as the set size increase because more words with the same index but different tags can reside in cache. However, an increase in the set size increases the number of bits in words of cache and requires more complex comparison logic.

4.4 Associate 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 **an associative memory or content address 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. The memory is capable of finding an empty unused location to store the word. When a word is to be read from an associative memory, the content of the word, or partof the word, is specified. The memory locates all words which match the specified contentand marks them for reading.

Hardware Organization:
The block diagram of an associative memory is shown in the Figure. It consists of a memory array and logic form words with n bits per word.

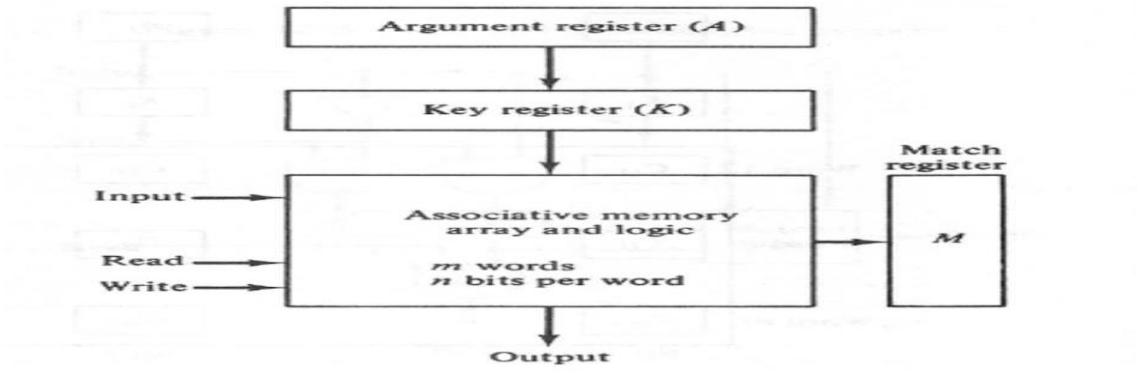


Figure 4.6: Block diagram of Associative memory

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.

The key register provides a mask for choosing a particular field or key in the argument word. The entire argument is compared with each memory word if the key register contains all 1's. Otherwise, only those bits in the argument that have 1's in their corresponding position of thekey register are compared.

4.5 Auxiliary Memory

The most common auxiliary memory devices used in computer systems are **magnetic disks** and **tapes**. The important characteristics of any device are its access mode, access time, transfer rate, capacity and cost.

The average time required to reach a storage location in memory and obtain its contents is called the **access time**.

The access time = seek time + transfer time

- Seek time : required to position the read-write head to a location
- Transfer time : required to transfer data to or from the device

Auxiliary storage is organized in records or blocks. A record is a specified number of characters or words. Reading or writing is always done on entire records. The transfer rate is the number of characters or words that the device can transfer per second.

Magnetic drums and disks are consisting of high speed rotating surfaces coated with a magnetic recording medium. Therecording surface rates at uniform speed. Bits are recorded as magnetic spots on the surface as it passes a stationary mechanism called a **write head.**

Stored bits are detected by a change in magnetic field produced by a recorded spot on the surface as it passes through a **read head.**

Magnetic 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.

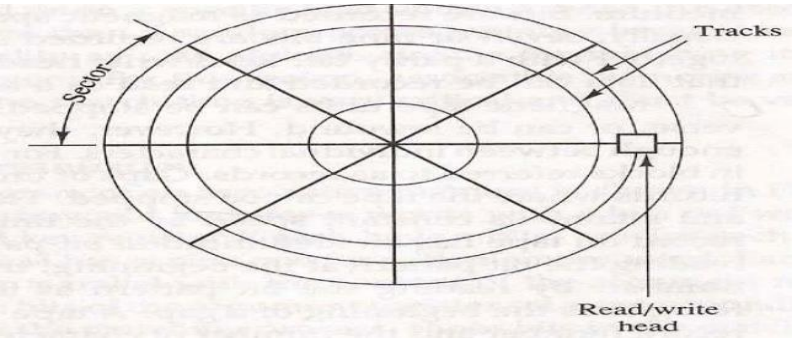


Figure 4.8: Magnetic Disk

Tracks and sectors

a) All tracks have same number of sectors
b) Outer tracks have more sectors than inner tracks.

The number of bytes stored in each sector is kept same. All tracks store the same amount ofdata. This results higher bit density in inner tracks than that of the outer tracks.

Since the same number of bytes is stored in each sector, the size of the inner sectors decidesthe storage capacity for all other sectors on the disk. Ex. Hard disk, floppy disk...

Magnetic Tape:

Magnetic tapes are used for backup memory. A magnetic tapes 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.

- Bits are recorded as magnetic sports on the tape along several tracks. Read/write heads are mounted one in each track so that data can be recorded and read as a sequence of characters. Magnetic tape units can be stopped, started to move forward or in reverse, or can be re-wound.
- However, they cannot be started or stopped fast enough between individual characters. For this reason, information is recorded in blocks referred to as records.

Optical disks:

Optical disks are used for backup memory. Information is written to or read from an optical disk using laser beam. It has very high storing capacity as compared to magnetic floppy disks. It has very long life. The capacity of optical disks varies from 650 MB to 17 GB. Dvds of 15, 25, 30 and 50 GB capacity etc., an optical disk is a direct access device. As its read /write head does not touch the disk surface, there is no disk wear and problem of head crash. Elaborate error checking codes can be used as there is no problem of space because of its high storage capacity.

The greatest drawback of an optical disk drive system is its large access time as compared to magnetic hard disk drive. An optical disk system the drive has to move on a sizable optical assembly across the disk surface. This results in an increased access time.

Types of optical disks: CD(compact disk), CD-R(recordable), CD-RW(read/write), DVD(digital versatile disk), DVD-R, DVD-RW.

4.6 Virtual Memory

Virtual memory is used to give programmers the illusion that they have a very large memory at their disposal, even though the computer actually has a relatively small main memory.

A 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 mappingtable.

- The address used by a programmer will be called a virtual address or logical address.
- An address in main memory is called a physical address

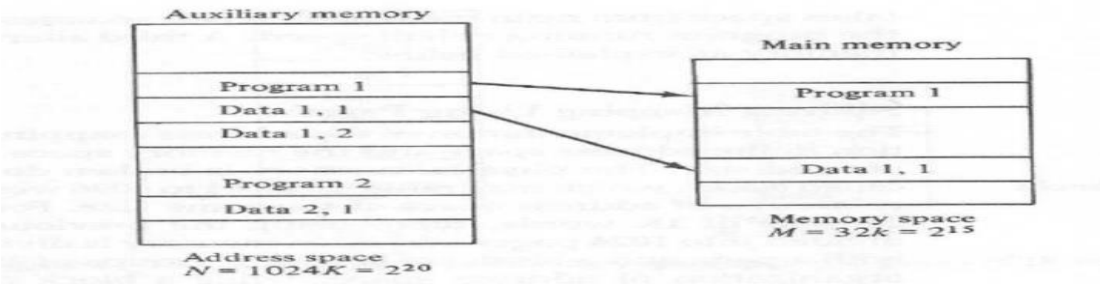


Figure 4.9: Relation between address and memory space in a virtual memory system

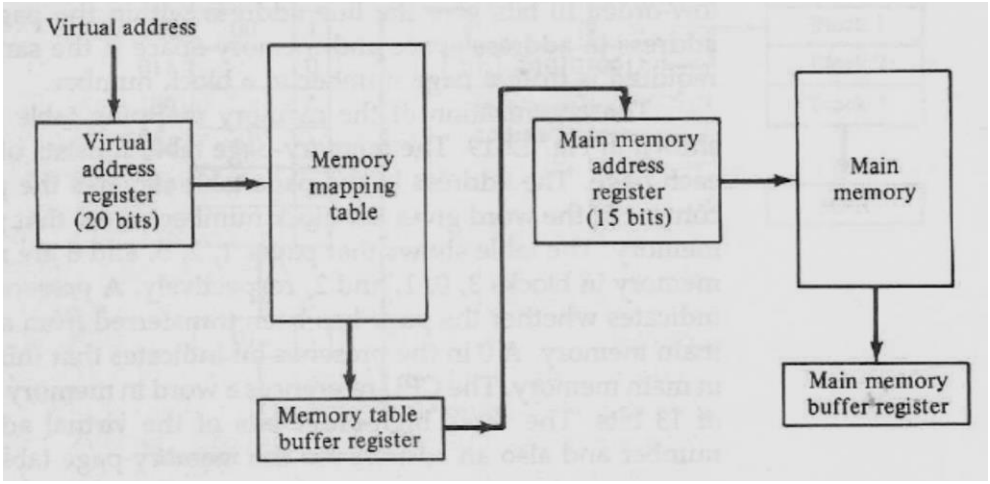


Figure 4.10: Memory table for mapping a virtual address