Unit -5

Syllabus: Memory Organization: Main memory- RAM, ROM, Secondary Memory – Magnetic Tape, Disk, Optical Storage, Cache Memory: Cache Structure and Design, Mapping Scheme, Replacement Algorithm, Improving Cache Performance, Virtual Memory, memory management hardware.

Memory Organization:

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 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. They are used for storing system programs, large data files, and other backup information. Only programs and data currently needed by the processor reside in main memory. All the other information is stored in auxiliary memory and transferred to main memory when needed.

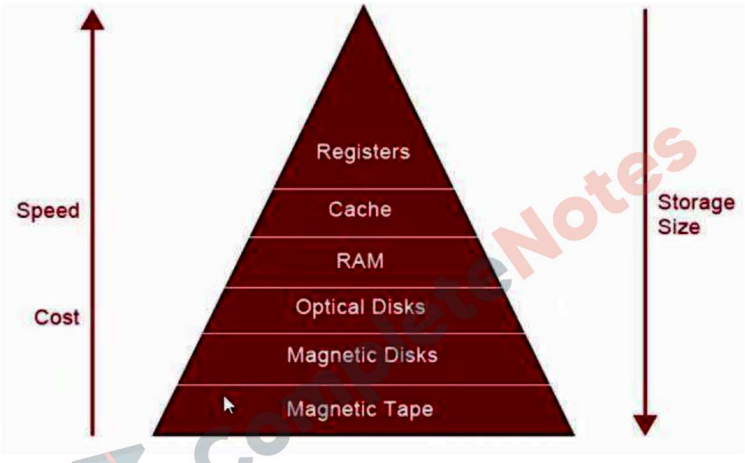


Figure 1: Memory hierarchy

Figure 1 illustrates the components in a typical memory hierarchy.

- At the bottom of the hierarchy are the relatively slow magnetic tapes used to store removable files.
- Next are the magnetic disks used as backup storage.
- The main memory occupies a central position by being able to communicate directly with the CPU and with auxiliary memory devices through an I/O processor.
- A special very-high-speed memory called a cache memory is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate.
- The cache memory is employed in computer systems to compensate for the speed differential between main memory access time and processor logic.

MAIN MEMORY:

- The main memory is the central storage unit in a computer system.
- It is a relatively 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 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 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.
- 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.

RAM CHIPS:

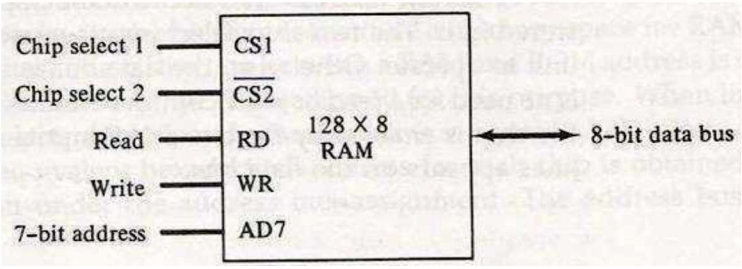


Figure 2:Block diagram of a RAM chip

The block diagram of a RAM chip is shown in Figure 2. The capacity of the memory is 128 words of eight bits (one byte) per word. This requires a 7-bit address and an 8- bit bidirectional data bus.

The read and write inputs specify the memory operation and the two chips select (CS) control inputs are enabling the chip only when it is selected by the microprocessor. The availability of more than one control input to select the chip facilitates the decoding of the address lines when multiplechips are used in the microcomputer. The read and write inputs are sometimes combined into one line labeled R/W. When the chip is selected, the two binary states in this line specify the two operations of read or write. The function table listed in Figure 3 specifies the operation of the RAM chip. When the WR input is enabled, the memory stores a byte from the data bus into a location specified by the address input lines. When the RDinput is enabled, the content of the selected byte is placed into the data bus. The RD and WR signals control the memory operation as well as the bus buffers associated with the bidirectional data bus.

CS1	$\overline{CS2}$	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

Figure 3: Function Table of RAM

ROM CHIPS:

ROM chip is organized externally in a similar manner. However, since a ROM can only read, the data bus can only be in an output mode. The block diagram of a ROM chip is shown in Figure 4. For the same-size chip, it is possible to have more bits of ROM than of RAM, because the internal binary cells in ROM occupy less space than in RAM. For this reason, the diagram specifies a 512-byte ROM, while the RAM has only 128 bytes. The nine address lines in the ROM chip specify any one of the 512 bytes stored in it. The two chip select inputs must be CS1 = 1 and CS2 = 0 for the unit to operate. Otherwise, the data bus is in a high-impedance state. There is no need for a read or write control because the unit can only read. Thus when the chip is enabled by the two select inputs, the byte selected by the address lines appears on the data bus.

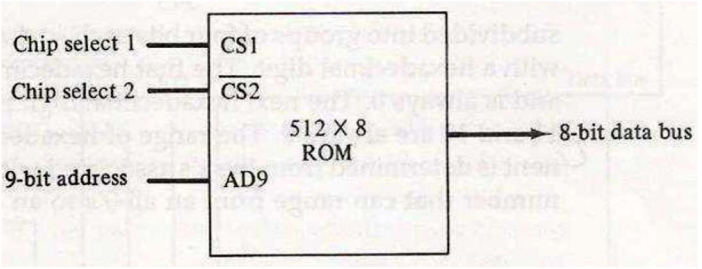


Figure 4: Rom Chip

Secondary Memory –The most common auxiliary memory devices used in computer systems are magnetic disks and tapes. Other components used, but not as frequently, are magnetic drums, magnetic bubble memory, and optical disks.

The recording surface rotates at uniform speed and is not startedor stopped during access operations. 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. The amount of surface available for recording in a disk is greater than in a drum of equal physical size. Therefore, more information can be stored on a disk than on a drum of comparable size. For this reason, disks have replaced drums in more recent computers.

Magnetic Disks:

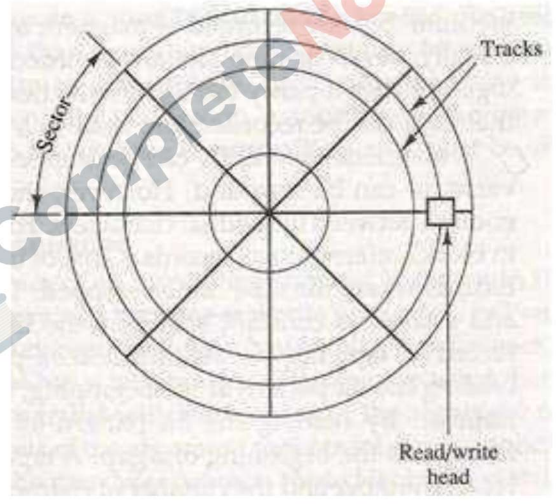


Figure 5: Magnetic Disk

- 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 dividedinto 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 Figure5.
- There are two sizes commonly used, with diameters of 5.25 and 3.5 inches. The 3.5-inch disks are smaller and can store more data than can the 5.25-inch disks.

Magnetic Tape:

- A 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.

- Bits are recorded as magnetic spots on the tape along several tracks.
- Usually, seven or nine bits are recorded simultaneously to form a character together with a parity bit.
- 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 rewound.
- Gaps of unrecorded tape are inserted between records where the tape can be stopped.
- The tape starts moving while in a gap and attains its constant speed by the time it reaches the next record.
- Each record on tape has an identification bit pattern at the beginning and end.
- A tape unit is addressed by specifying the record number and the number of characters in the record. Records may be of fixed or variable length.

Optical Storage:

Optical storage devices use optical technology to save and retrieve data on discs, like a Blu-ray, CD, DVD. The device uses a laser light to read information on the disc and to "write" new information to the disc for future retrieval.

CACHE MEMORY: CACHE STRUCTURE AND DESIGN

A special very-high-speed memory called a cache memory is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate.

The basic operation of the cache is as follows:

- When the CPU needs to access memory, the cache is examined. If the word is found in the cache, it is called cache hit. If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word and it is called cache miss.
- Some data are transferred to cache so that future references to memory find the required words in the fast cache memory.
- The performance of cache memory is frequently measured in terms of a quantitycalled *hit ratio.*
- The ratio of the number of hits divided by the total CPU references to memory (hits plus misses) is the hit ratio.

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:

- Associative mapping
- Direct mapping
- Set- associative mapping

To help in the discussion of these three mapping procedures we will use a specific example of a memory organization as shown in Figure 6. The main memory can store 32K words of 12 bits each. The cache is capable of storing 512 of these words at any given time. For every word stored in cache, there is a duplicate copy in main memory. The CPU communicates with both memories. It first sends a 15-bit address to cache. If there is a hit, the CPU accepts the 12-bit data from cache. If there is a miss, the CPU reads the word from main memory and the word is then transferred to cache.

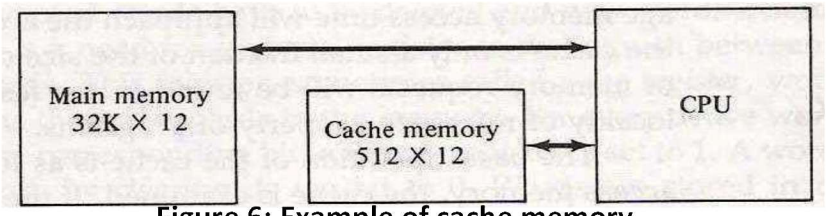


Figure 6: Example of cache memory.

MAPPING SCHEME:

1. 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.

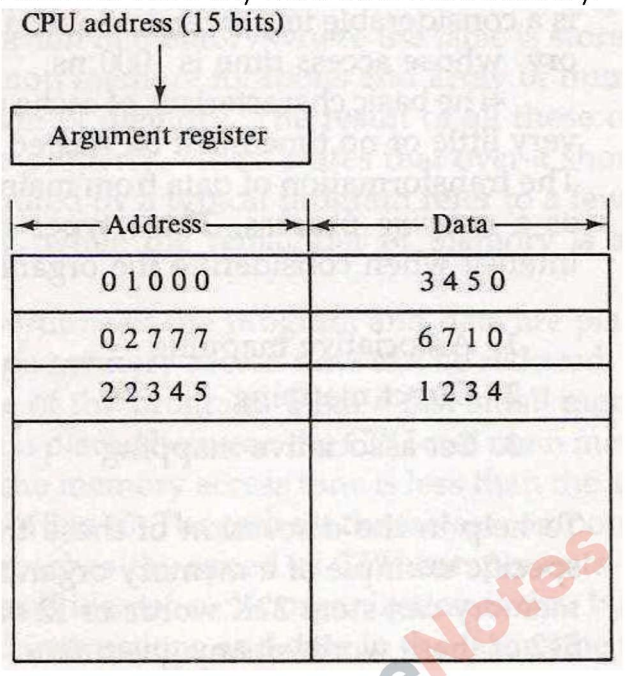


Figure 7: Associative mapping cache (all numbers in octal).

- 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.
- The address data pair is then transferred to the associative cache memory.
- If the cache is full, an address data pair must be displaced to make room for a pair that is needed and not present in the cache.
- The decision as to what pair is replaced is determined from the replacement algorithm that the designer chooses for the cache.

2. Direct Mapping:

- Associative memories are expensive compared to random-access memories because of the added logic associated with each cell.
- The CPU address of 15 bits is divided into two fields. The nine least significant bits constitute the indexfield and the remaining six bits forms the tagfield.
- 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.

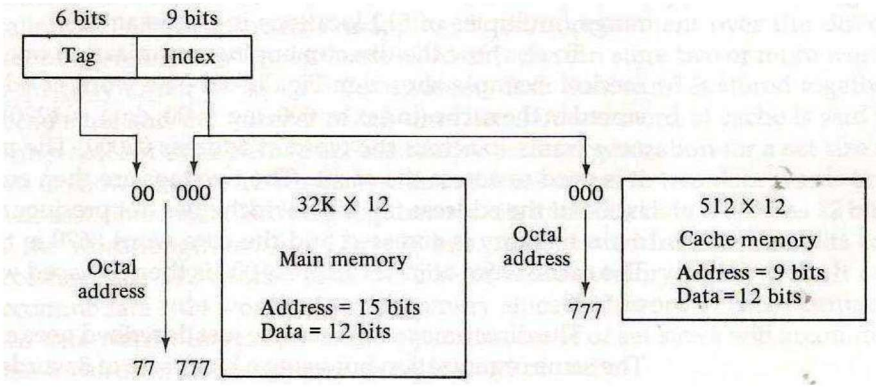
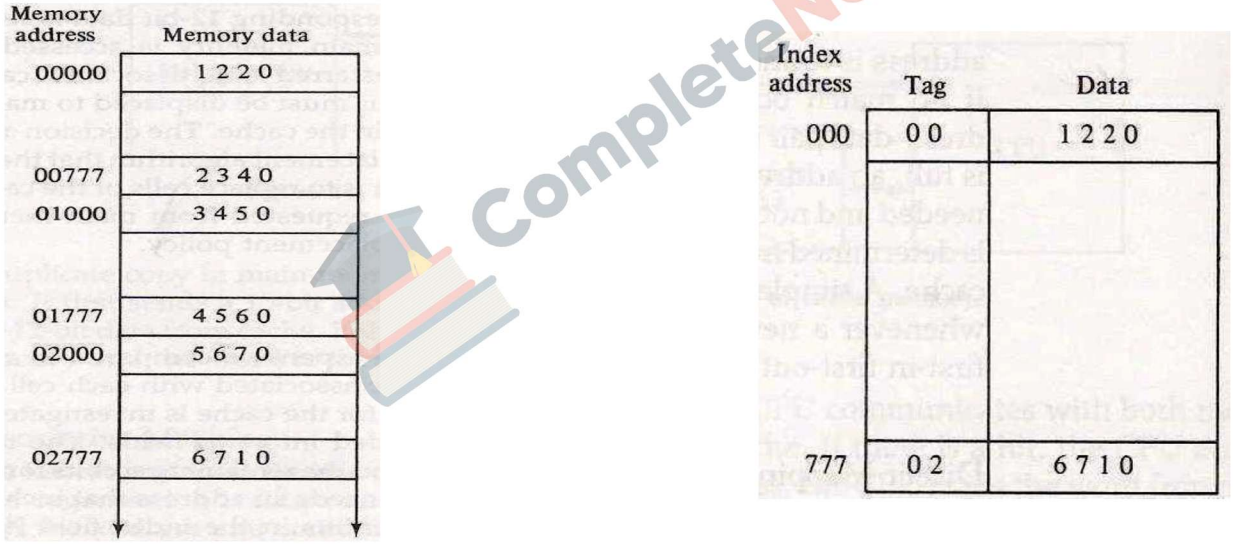


Figure 8: Addressing relationships between main and cache memories.

Consider a numerical example.

- The word at address zero is presently stored in the cache (index = 000, tag = 00, data = 1220).
- Suppose that the CPU now wants to access the word at address 02000.
- The index address is 000, so it is used to access the cache. The two tags are then compared.
- The cache tag is 00 but the address tag is 02, which does not produce a match.
- Therefore, the main memory is accessed and the data word 5670 is transferred to the CPU.
- The cache word at index address 000 is then replaced with a tag of 02 and data of 5670.



(a) Main memory **(b) Cache memory**

Figure 9: Direct mapping cache organization.

3. Set-Associative Mapping:

- It was mentioned previously that the disadvantage of direct mapping is that, two words with the same index but different tag values cannot reside in cache memory at the same time.
- A third type of cache organization, called 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.

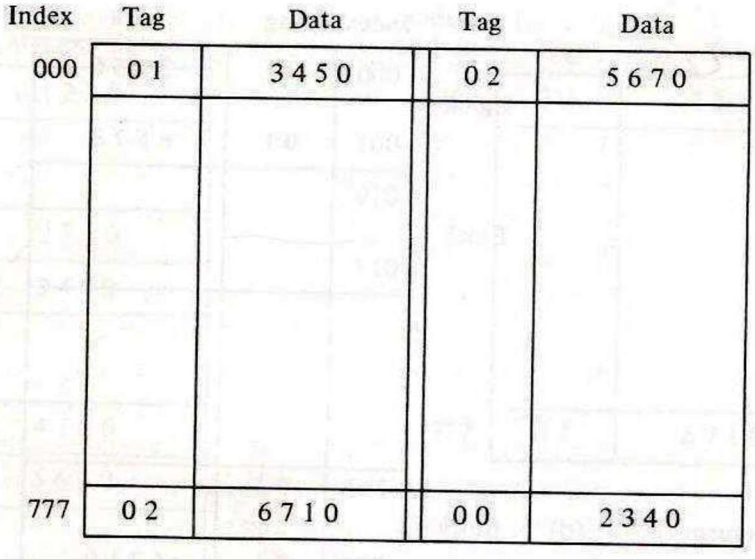


Figure 10: Two-way set -associative mapping cache.

The octal numbers listed in Figure 10 are with reference to the main memorycontents.

- The words stored at addresses 01000 and 02000 of main memory are stored in cache memory at index 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 with both tags in the cache to determine if a match occurs.
- The comparison logic is done by an associative search of the tags in the set similar to an associative memory search: thus the name "set-associative."

REPLACEMENT ALGORITHM:

Avirtual memory system is a combination of hardware and software techniques. The memory management software system handles all the software operations for the efficient utilization of memory space.

It must decide:

- Which page in main memory ought to be removed to make room for a new page.
- When a new page is to be transferred from auxiliary memory to main memory.
- Where the page is to be placed in main memory.

The hardware mapping mechanism and the memory management software together constitute the architecture of a virtual memory. When a program starts execution, one or more pages are transferred into main memory and the page table is set to indicate their position. The program is executed from main memory until it attempts to reference a page that is still in auxiliary memory. This condition is called*page fault.* When page fault occurs, the execution of the present program is suspended until the required page is brought into main memory.

Three of the most common replacement algorithms used is the **First-In-First – Out (FIFO), Least Recently Used(LRU)** and the **Optimal Page Replacement Algorithms.**

FIFO:

- The FIFO algorithm selects for replacement the page that has been in memory the longest time.
- Each time a page is loaded into memory, its identification number is pushed into a FIFO stack.

- FIFO will be full whenever memory has no more empty blocks.
- When a new page must be loaded, the page least recently brought in is removed.
- The FIFO replacement policy has the advantage of being easy to implement. It has the disadvantage that under certain circumstances pages are removed and loaded from memory too frequently.

LRU:

- The LRU policy is more difficult to implement but has been more attractive on the assumption that the least recently used page is a better candidate for removal than the least recently loaded page as in FIFO.
- The LRU algorithm can be implemented by associating a counter with every page that isin main memory.
- When a page is referenced, its associated counter is set to zero. At fixed intervals of time, the counters associated with all pages presently in memory are incremented by 1.
- The least recently used page is the page with the highest count. The counters are often called *aging registers,* as their count indicates their age, that is, how long ago their associated pages have been referenced.

Optimal Page Replacement Algorithms

- The optimal page algorithm simply says that the page with the highest label should be removed.
- If one page will not be used for 8 million instructions and another page will not be used for 6 million instructions, removing the former pushes the page fault that will fetch it back as far into the future as possible.

IMPROVING CACHE PERFORMANCE:

There are three ways to improve cache performance:

1. Reduce the miss rate.
2. Reduce the miss penalty.
3. Reduce the time to hit in the cache.

VIRTUAL MEMORY

- In a memory hierarchy system, programs and data are first stored in auxiliary memory.
- Portions of a program or data are brought into main memory as they are needed by the CPU.
- *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 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.

Address Space and Memory Space

- An address used by a programmer will be called a virtual address, and the set of such addresses is 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.
- Thus the address space is the set of addresses generated by programs as they reference instructions and data, the memory space consists of the actual main memory locations directly addressable for processing.
- In most computers the address and memory spaces are identical.
- The address space is allowed to be larger than the memory space in computers with virtual memory.

- In a multi-program computer system, programs and data are transferred to and from auxiliary memory and main memory based on demands imposed by the CPU.

Suppose that program 1 is currently being executed in the CPU. Program 1 and a portion of its associated data are moved from auxiliary memory into main memory as shown in Figure 11. Portions of programs and data neednot be in contiguous locations in memory since information is being moved in and out, and empty spaces may be available in scattered locations in memory.

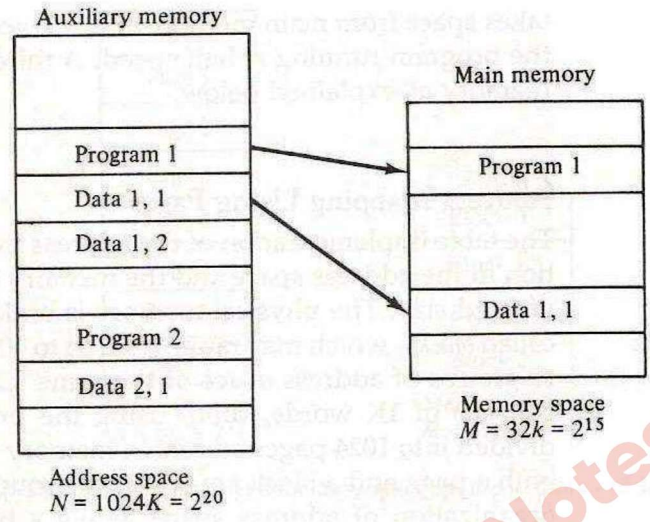


Figure 11: Relation between address and memory space in a virtual memory system.

Address Mapping Using Pages:

- The table implementation of the address mapping is simplified if the information in the address space and the memory space are each divided into groups of fixed size.
- The physical memory is broken down into groups of equal size called *blocks,* which may range from 64 to 4096 words each. The term pagerefers to groups of address space of the same size.

For example, if a page or block consists of 1K words, then, address space is divided into 1024 pages and main memory is divided into 32 blocks. Although both a page and a block are split into groups of 1K words, a page refers to the organization of address space, while a block refers to the organization of memory space. The programs are also considered to be split into pages. Portions of programs are moved from auxiliary memory to main memory in records equal to the size of a page. The term "page frame" is sometimes used to denote a block. Consider a computer with an address space of 8K and a memory space of 4K.

- The mapping from address space to memory space is facilitated if each virtual address is considered to be represented by two numbers: a page number address and a line within the page.
- In a computer with 2 words per page, *p* bits are used to specify a line address and the remaining high-order bits of the virtual address specify the page number.

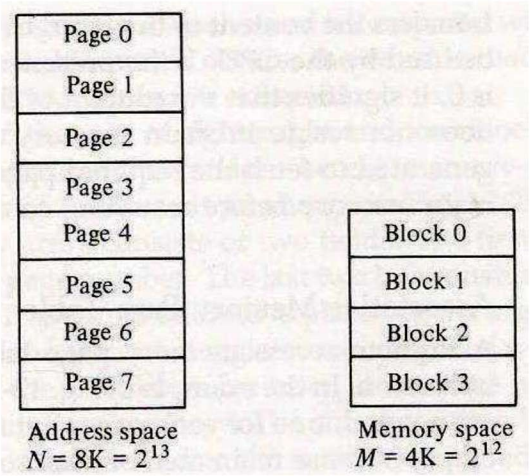


Figure 12:Address space and memory space split into groups of 1K words.

MEMORY MANAGEMENT HARDWARE:

- 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.

The basic components of a memory management unit are:

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

The dynamic storage relocation hardware is a mapping process similar to the paging system. The fixed page size used in the virtual memory system causes certain difficulties with respect to program size and the logical structure of programs.

- It is more convenient to divide programs and data into logical parts called segments.
- A segment is a set of logically related instructions or data elements associated with a given name.
- Segments may be generated by the programmer or by the operating system.
- Examples of segments are a subroutine, an array of data, a table of symbols, or a user's program.
- The address generated by a segmented program is called a logical address.
- This is similar to a virtual address except that logical address space is associated with variable-length segments rather than fixed-length pages.

Segmented-Page Mapping

- The property of logical space is that it uses variable-length segments.
- The length of each segment is allowed to grow and contractaccording to the needs of the program being executed.
- One way of specifying the length of a segment is by associating with it a number of equal-size pages.

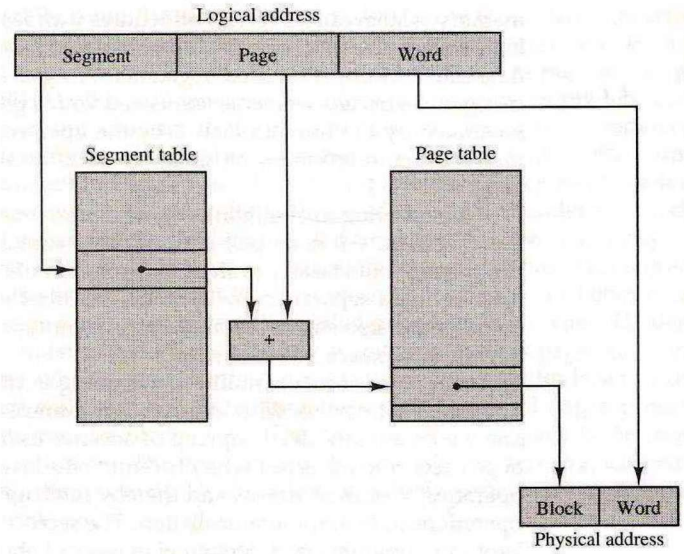


Figure 13: Logical to physical address mapping

- The logical address is partitioned into three fields: the segment field specifies a segment number, the page field specifies the page within the segment and the word field gives the specific word within the page.
- A page field of k bits can specify up to 2^k pages.
- A segment number may be associated with just one page or with as many as2^k pages.
- Thus the length of a segment would vary according to the number of pages that are assigned to it.
- The mapping of the logical address into a physical address is done by means of two tables, as shown in figure 13.

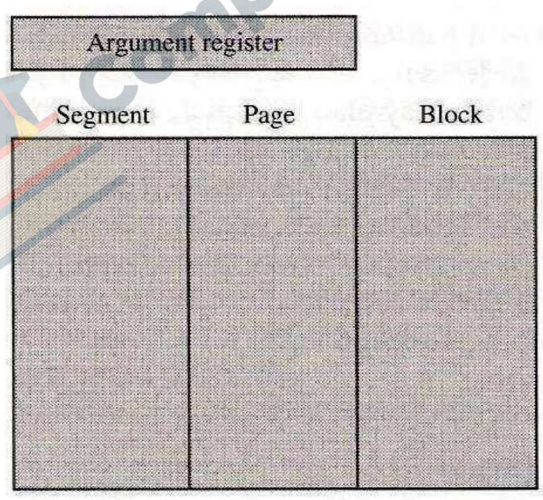


Figure 14: Associate memory translation look aside buffer TLB

The mapping of the logical address into a physical address is done by means of two tables, as shown in figure 13. The segment number of the logical address specifies the address for the segment table.

- The entry in the segment table is a pointer address for a page table base.
- The page table base is added to the page number given in the logical address.
- The sum produces a pointer address to an entry in the page table.
- The value, found in the page table pro vides the block number in physical memory.
- Theconcatenation of the block field with the word field produces the final physical mapped address.

The two mapping tables may be stored in two separate small memories or in main memory. In either case, a memory reference from the CPU will require three accesses to main memory: one from the segment table, one from the page table, and the third front main memory.