Memory Organization

- Memory Hierarchy
- Main Memory
- Associative Memory
- Cache Memory: Cache Mapping techniques
- Virtual Memory

Memory Hierarchy

- Memory unit is essential component of digital computer since it is needed for storing programs and data.
- Memory unit that communicates directly with CPU is called Main memory.
- Devices that provide backup storage is called auxiliary memory.
- Only programs and data currently needed by processor reside in the main memory.
- All other information is stored in auxiliary memory and transferred to main memory when needed.

Table 4.1 Key Characteristics of Computer Memory Systems

Location

- Internal (e.g. processor registers, main memory, cache)
- External (e.g. optical disks, magnetic disks, tapes)

Capacity

- Number of words
- Number of bytes

Unit of Transfer

- Word
- Block

Access Method

- Sequential
- Direct
- Random
- Associative

Performance

- Access time
- Cycle time
- Transfer rate

Physical Type

- Semiconductor
- Magnetic
- Optical
- Magneto-optical

Physical Characteristics

- Volatile/nonvolatile
- Erasable/nonerasable

Organization

- Memory modules

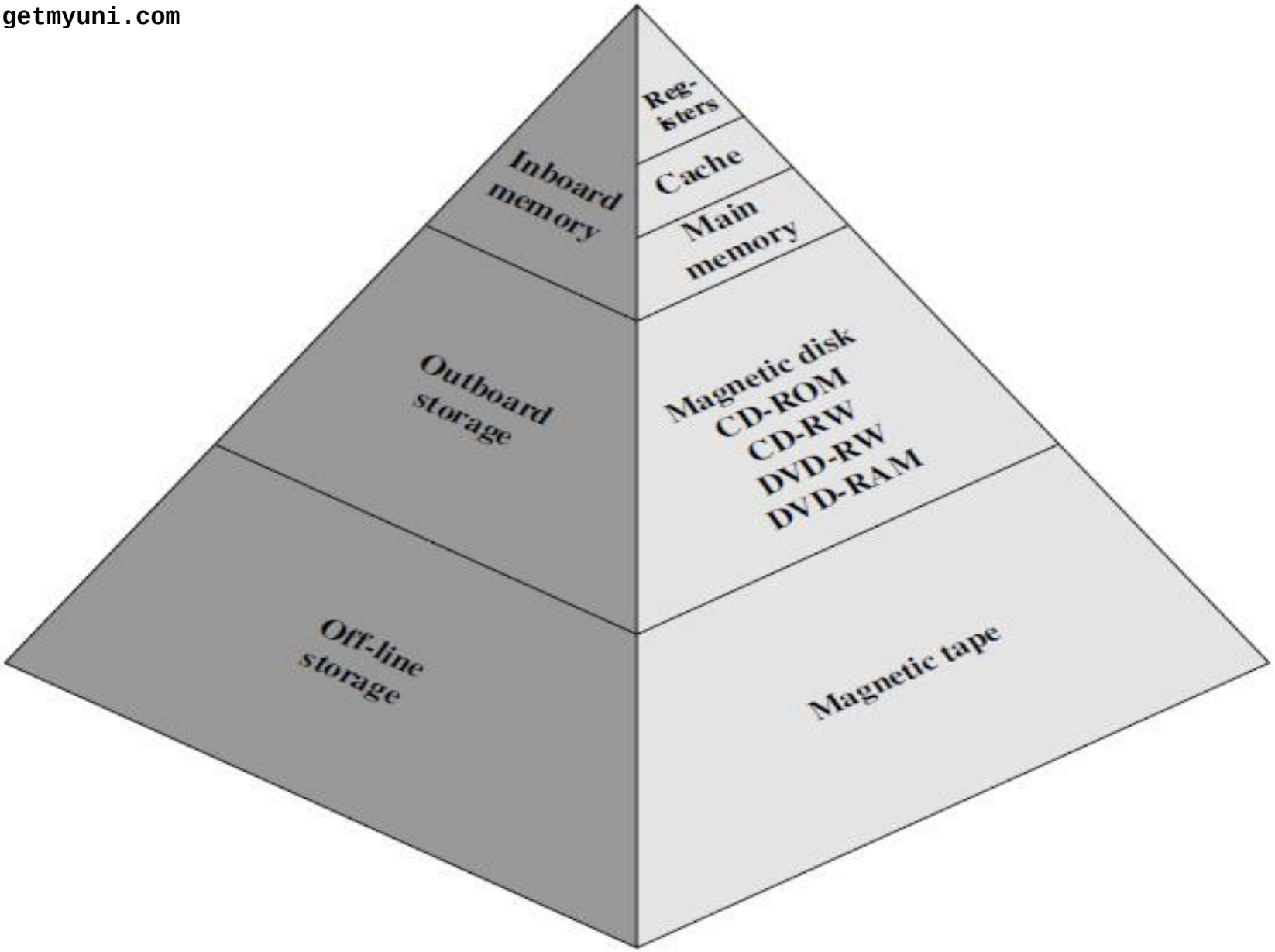


Figure 4.1 The Memory Hierarchy

- Memory hierarchy system consist of all storage devices from auxiliary memory to main memory to cache memory
- As one goes down the hierarchy :
  - Cost per bit decreases.
  - Capacity increases.
  - Access time increases.
  - Frequency of access by the processor decreases.

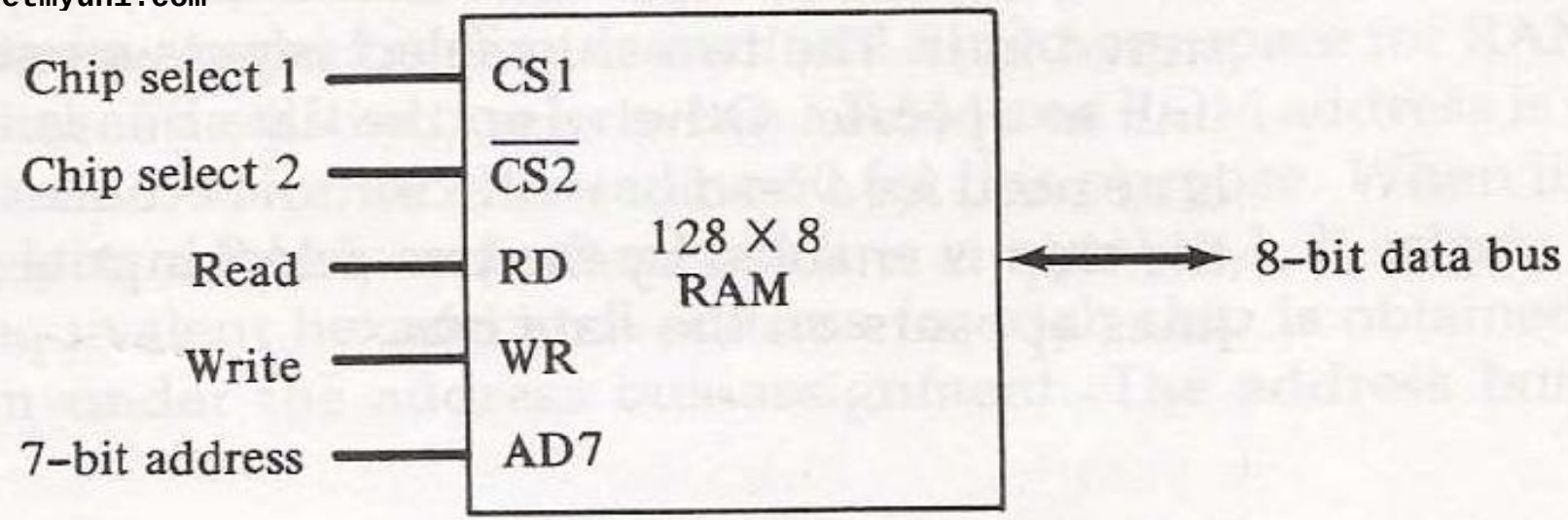
Main Memory

- It is the memory used to store programs and data during the computer operation.
- The principal technology is based on semiconductor integrated circuits.
- It consists of RAM and ROM chips.
- RAM chips are available in two form static and dynamic.

SRAM	DRAM
Uses capacitor for storing information	Uses Flip flop
More cells per unit area due to smaller cell size.	Needs more space for same capacity
Cheap and smaller in size	Expensive and bigger in size
Slower and analog device	Faster and digital device
Requires refresh circuit	No need
Used in main memory	Used in cache

- ROM is uses random access method.
- It is used for storing programs that are permanent and the tables of constants that do not change.
- ROM store program called bootstrap loader whose function is to start the computer software when the power is turned on.
- When the power is turned on, the hardware of the computer sets the program counter to the first address of the bootstrap loader.

Figure Typical RAM chip.



(a) Block diagram

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

(b) Function table

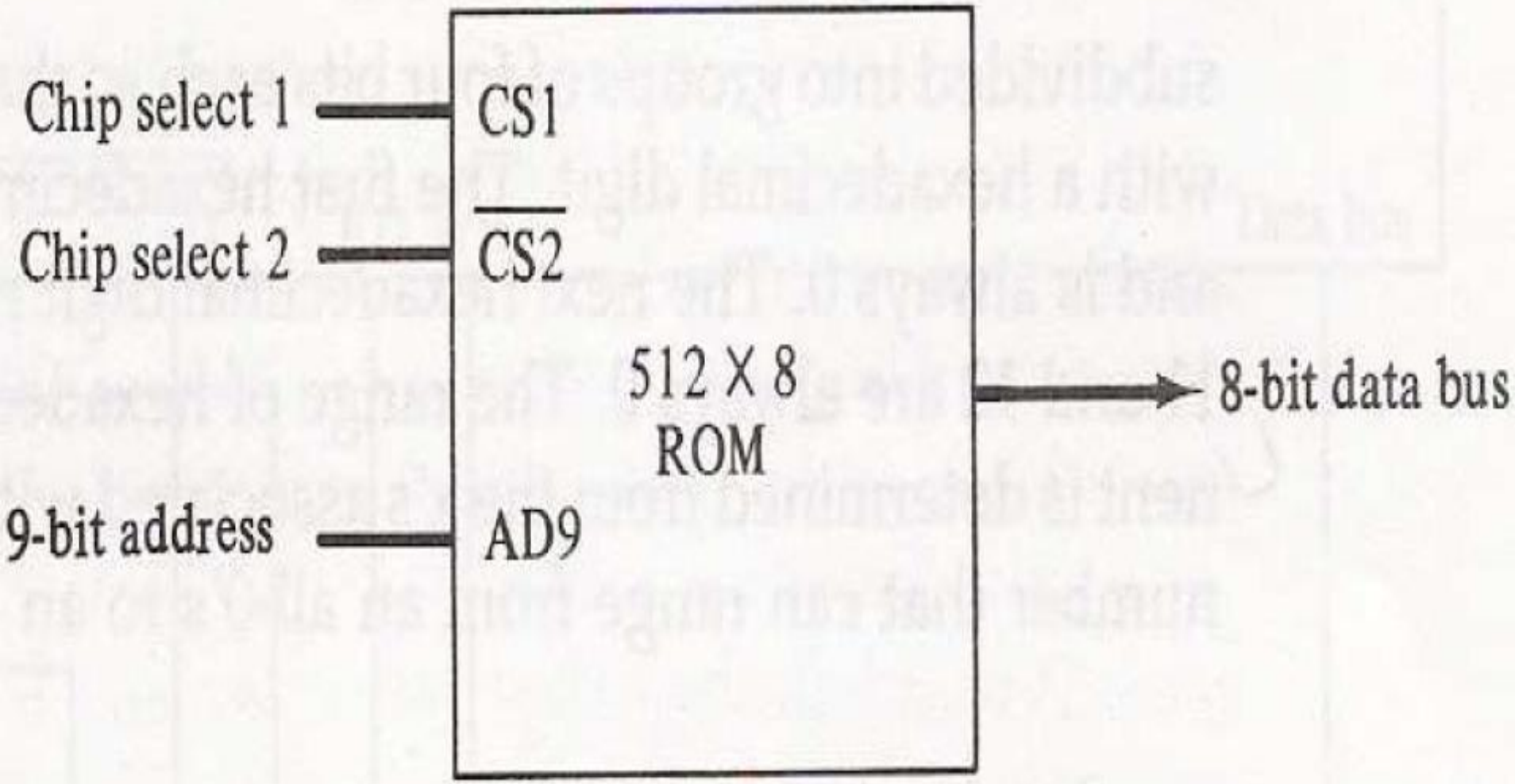


Figure Typical ROM chip.

- 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 512 byte ROM and 128 bytes RAM.

Memory address Map

- Designer must specify the size and the type(RAM or ROM) of memory to be used for particular application.
- The addressing of the memory is then established by means of table called memory address map that specifies the memory address assign to each chip.
- Let us consider an example in which computer needs 512 bytes of RAM and ROM as well and we have to use the chips of size 128 bytes for RAM and 512 bytes for ROM.

TABLE Memory Address Map for Microprocomputer

Component	Hexadecimal address	Address bus 10	9	8	7	6	5	4	3	2	1
RAM 1	0000–007F	0	0	0	x	x	x	x	x	x	x
RAM 2	0080–00FF	0	0	1	x	x	x	x	x	x	x
RAM 3	0100–017F	0	1	0	x	x	x	x	x	x	x
RAM 4	0180–01FF	0	1	1	x	x	x	x	x	x	x
ROM	0200–03FF	1	x	x	x	x	x	x	x	x	x

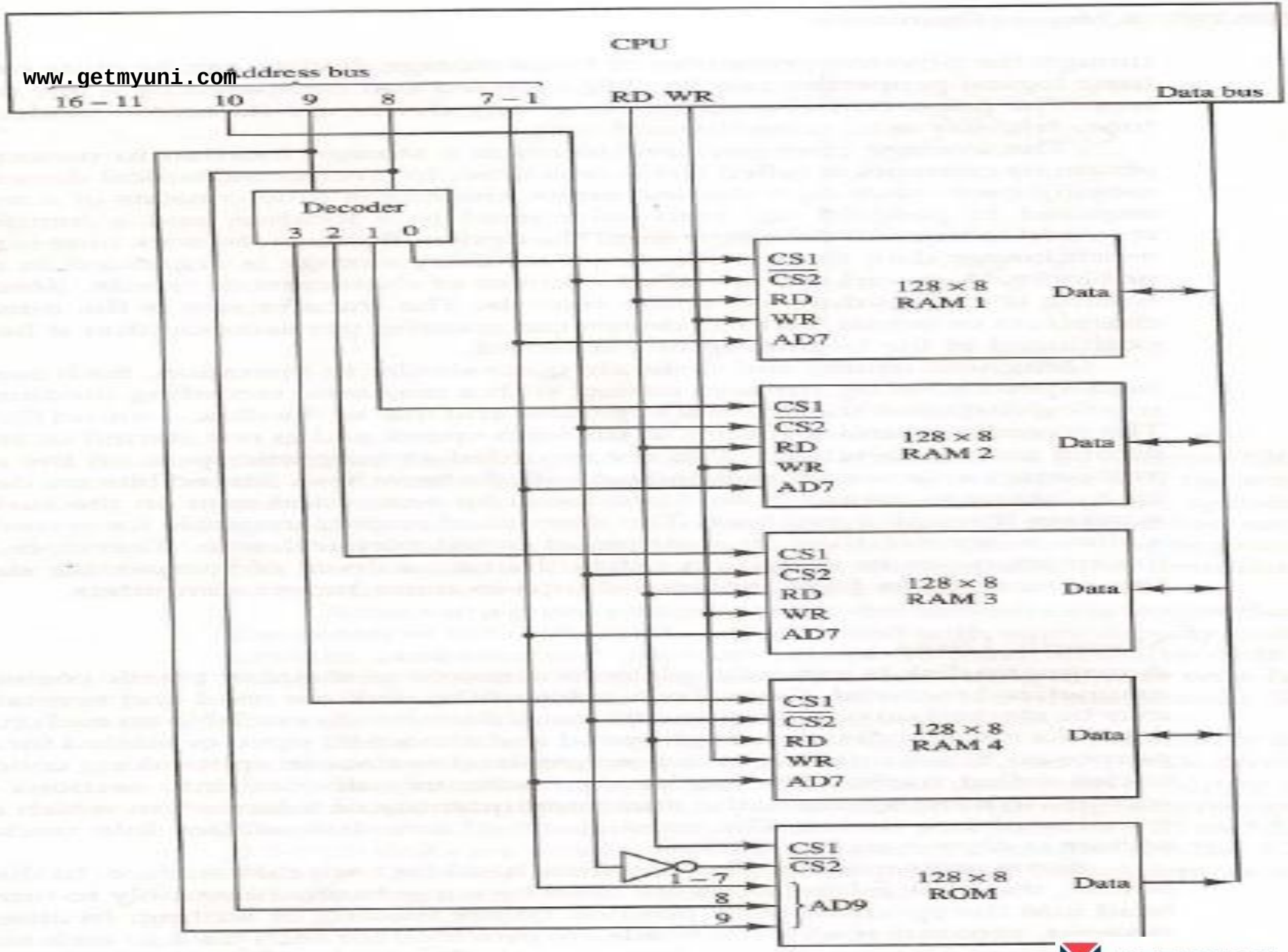


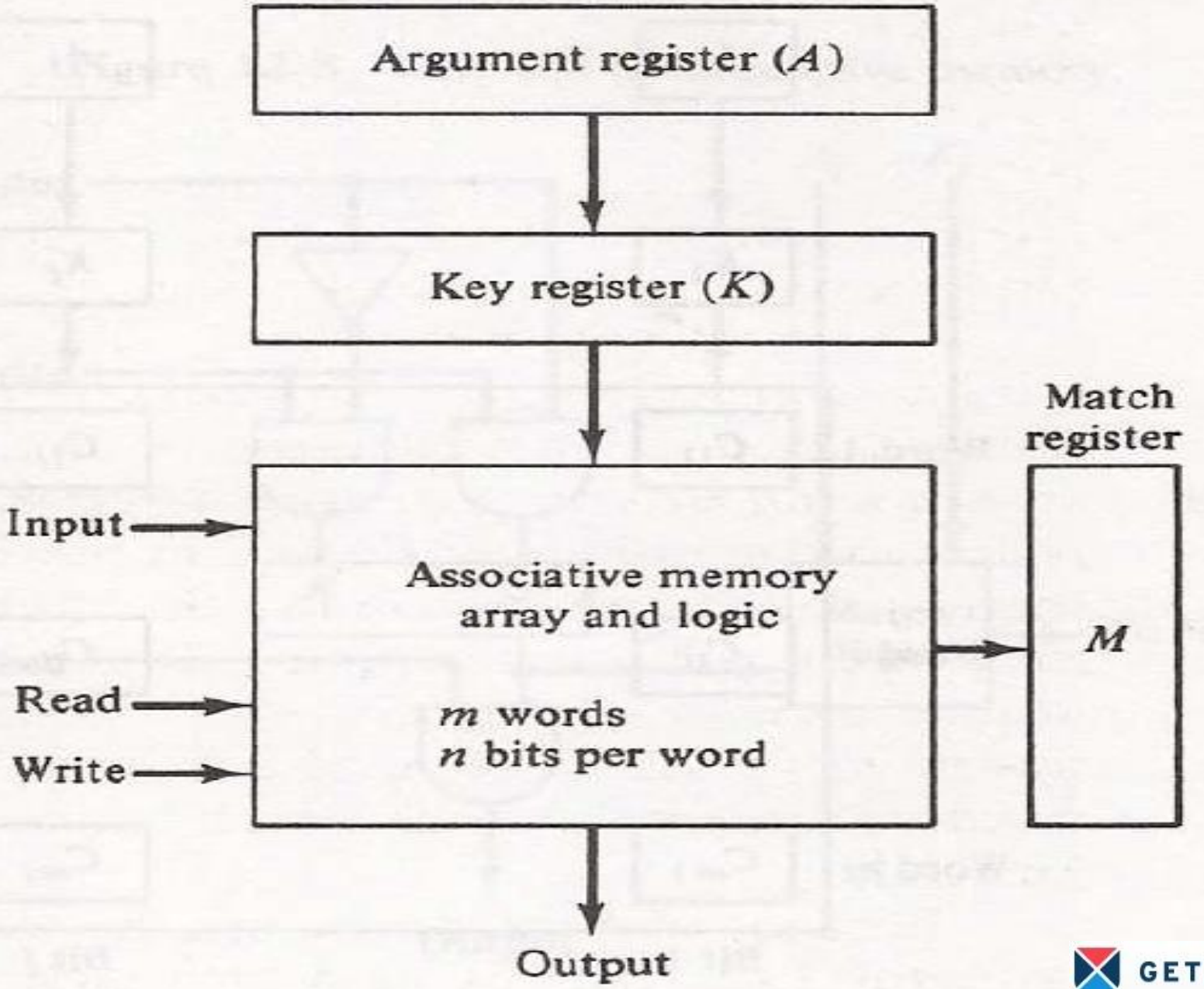
Figure 12-4 Memory connection to the CPU.

Associative Memory

- To search particular data in memory, data is read from certain address and compared if the match is not found content of the next address is accessed and compared.
- This goes on until required data is found. The number of access depend on the location of data and efficiency of searching algorithm.
- The searching time can be reduced if data is searched on the basis of content.

- A memory unit accessed by content is called associative memory or content addressable memory(CAM)
- This type of memory is accessed simultaneously and in parallel on the basis of data content.
- Memory is capable of finding empty unused location to store the word.
- These are used in the application where search time is very critical and must be very short.

Figure Block diagram of associative memory.

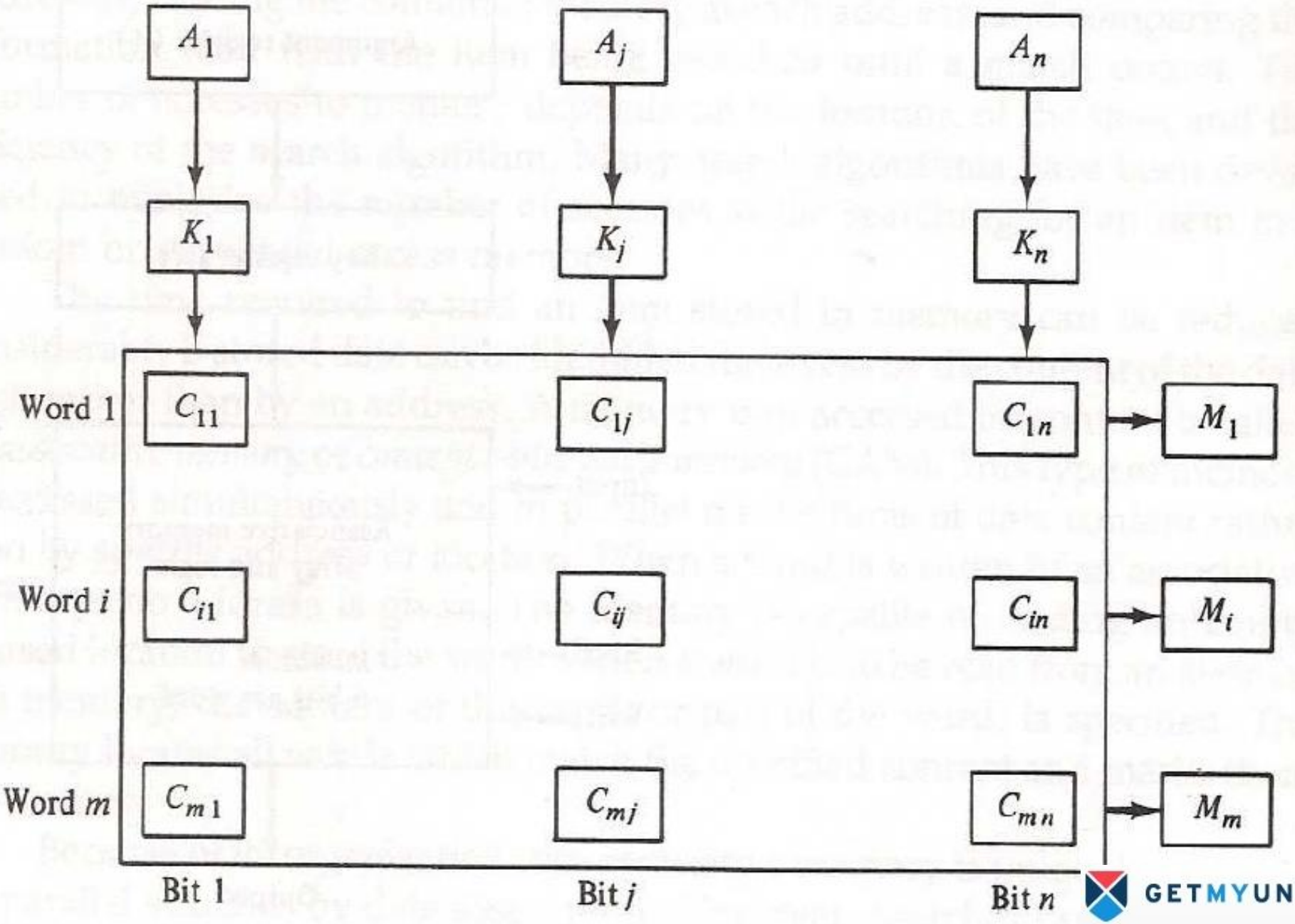


- It consists memory array of m words with n bits per words
- Argument register A and key register K have n bits one for each bit of word.
- Match register has m bits, one for each memory word.
- Each word in memory is compared in parallel with the content of the A register. For the word that match corresponding bit in the match register is set.

- Key register provide the mask for choosing the particular field in A register.
- The entire content of A register is compared if key register content all 1.
- Otherwise only bit that have 1 in key register are compared.
- If the compared data is matched corresponding bits in the match register are set.
- Reading is accomplished by sequential access in memory for those words whose bit are set.

A	101 111100	
K	111 000000	
Word 1	100 111100	no match
Word 2	101 000001	match

Figure Associative memory of m word, n cells per word.



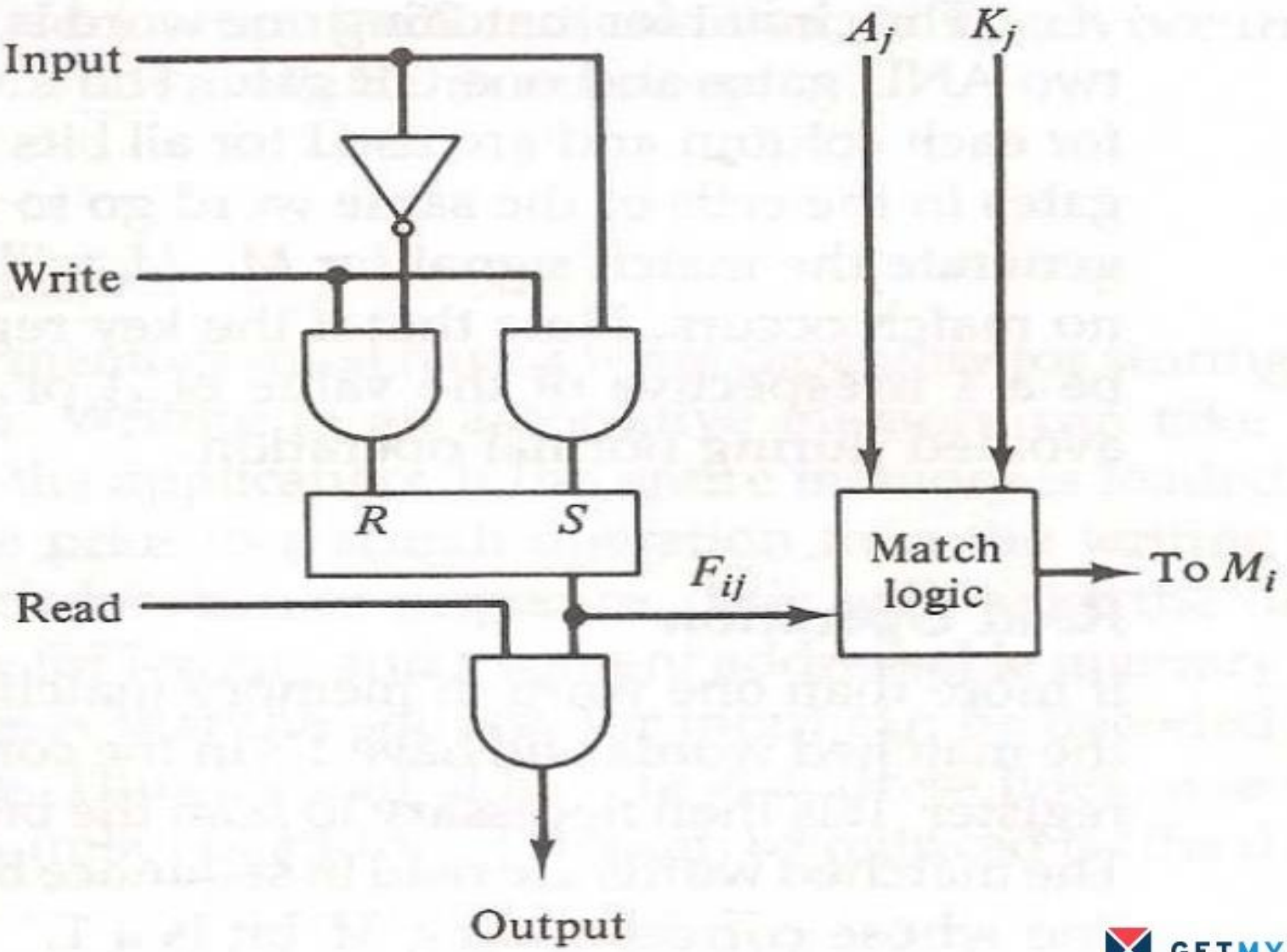
Match Logic

- Let us neglect the key register and compare the content of argument register with memory content.
- Word i is equal to argument in A if $A_j = F_{ij}$ for j=1,2,3,4……..n
- The equality of two bits is expressed as

- x_j =1 if bits are $x_j = A_j F_{ij} + A_j' F_{ij}'$

$$M_i = x_1 x_2 x_3 \cdots x_n$$

Figure One cell of associative memory.



- Let us include key register. If K_j=0 then there is no need to compare A_j and F_{ij}.
- Only when K_j=1, comparison is needed.
- This achieved by ORing each term with K_j.

$$M_i = (x_1 + K_1')(x_2 + K_2')(x_3 + K_3') \cdots (x_n + K_n')$$

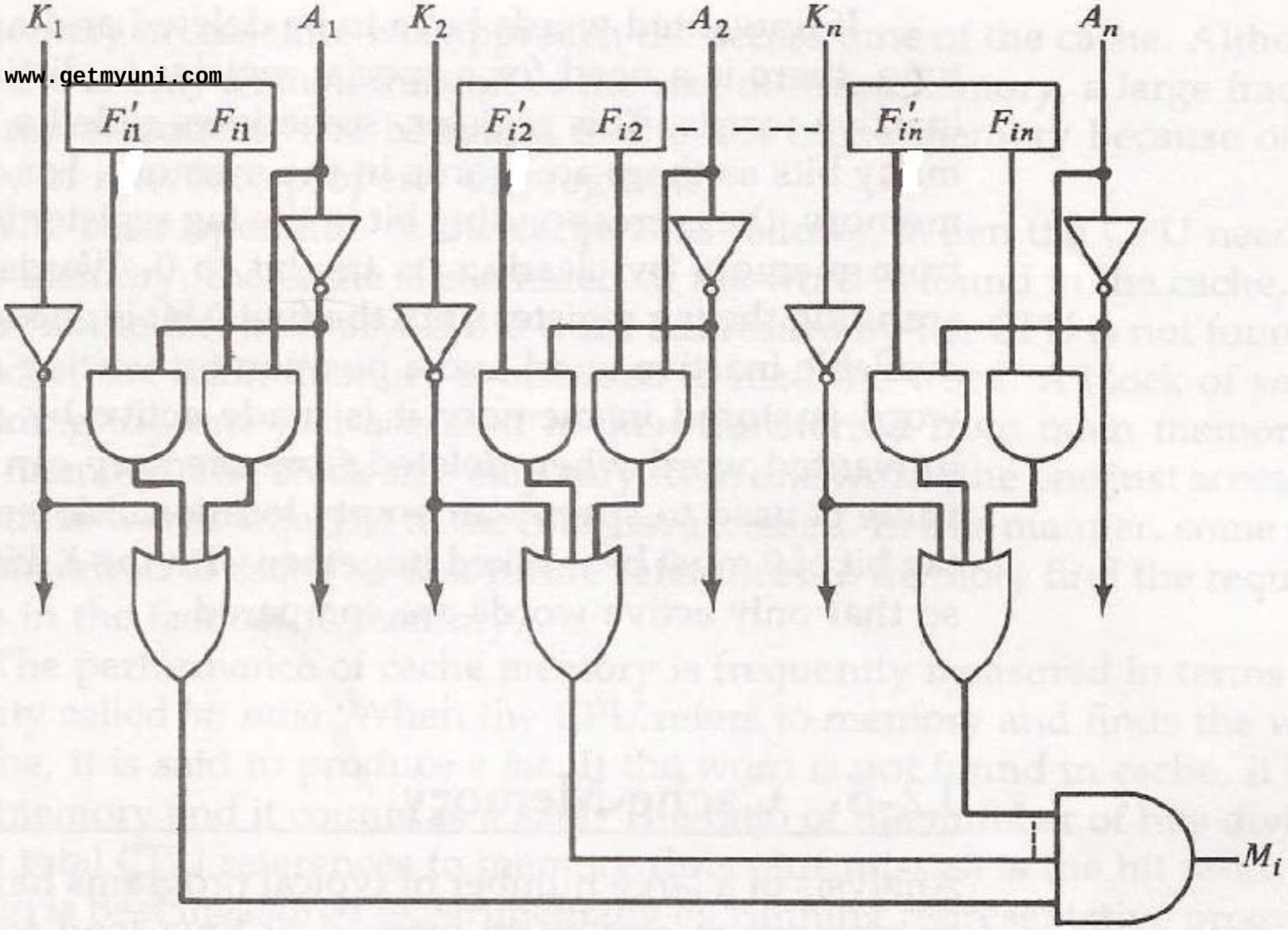


Figure Match logic for one word of associative men

Read Operation

- If more than one word match with the content, all the matched words will have 1 in the corresponding bit position in match register.
- Matched words are then read in sequence by applying a read signal to each word line.
- In most application, the associative memory stores a table with no two identical items under a given key.

Write Operation

- If the entire memory is loaded with new information at once prior to search operation then writing can be done by addressing each location in sequence.
- Tag register contain as many bits as there are words in memory.
- It contain 1 for active word and 0 for inactive word.
- If the word is to be inserted, tag register is scanned until 0 is found and word is written at that position and bit is change to 1.

Cache Memory

- Analysis of large number of program shows that reference to memory at any given interval of time tend to be confined to few localized area in memory. This is known as locality of reference.
- If the active portion of program and data are placed in fast memory, then average execution time of the program can be reduced. Such fast memory is called cache memory.
- It is placed in between the main memory and the CPU.

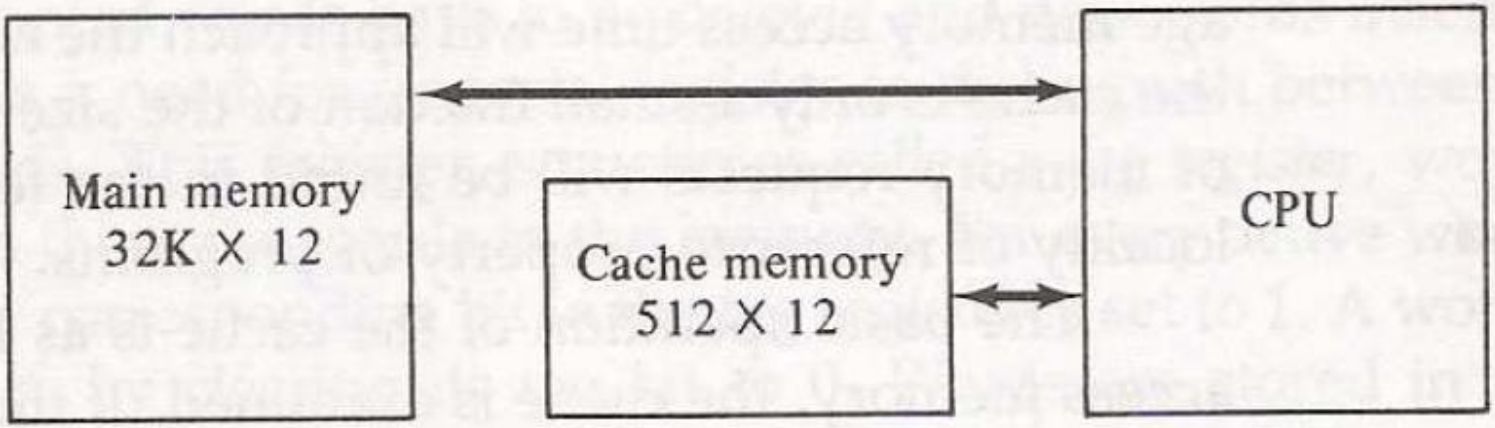


Figure Example of cache memory.

- When the CPU need to access the memory it first search in cache. If word is found, it is read.
- If the word is not found, it is read from main memory and a block of data is transferred from main memory to cache which contain the current word.
- If the word is found in cache, it is said hit. If the word is not found, it is called miss.
- Performance of cache is measured in terms of hit ratio which ratio of total hit to total memory access by CPU.

Mapping Techniques

- The transformation of data from main memory to cache is known as mapping process. Three types of mapping procedures are:
  - Associative Mapping
  - Direct Mapping
  - Set-Associative Mapping

Associative Mapping

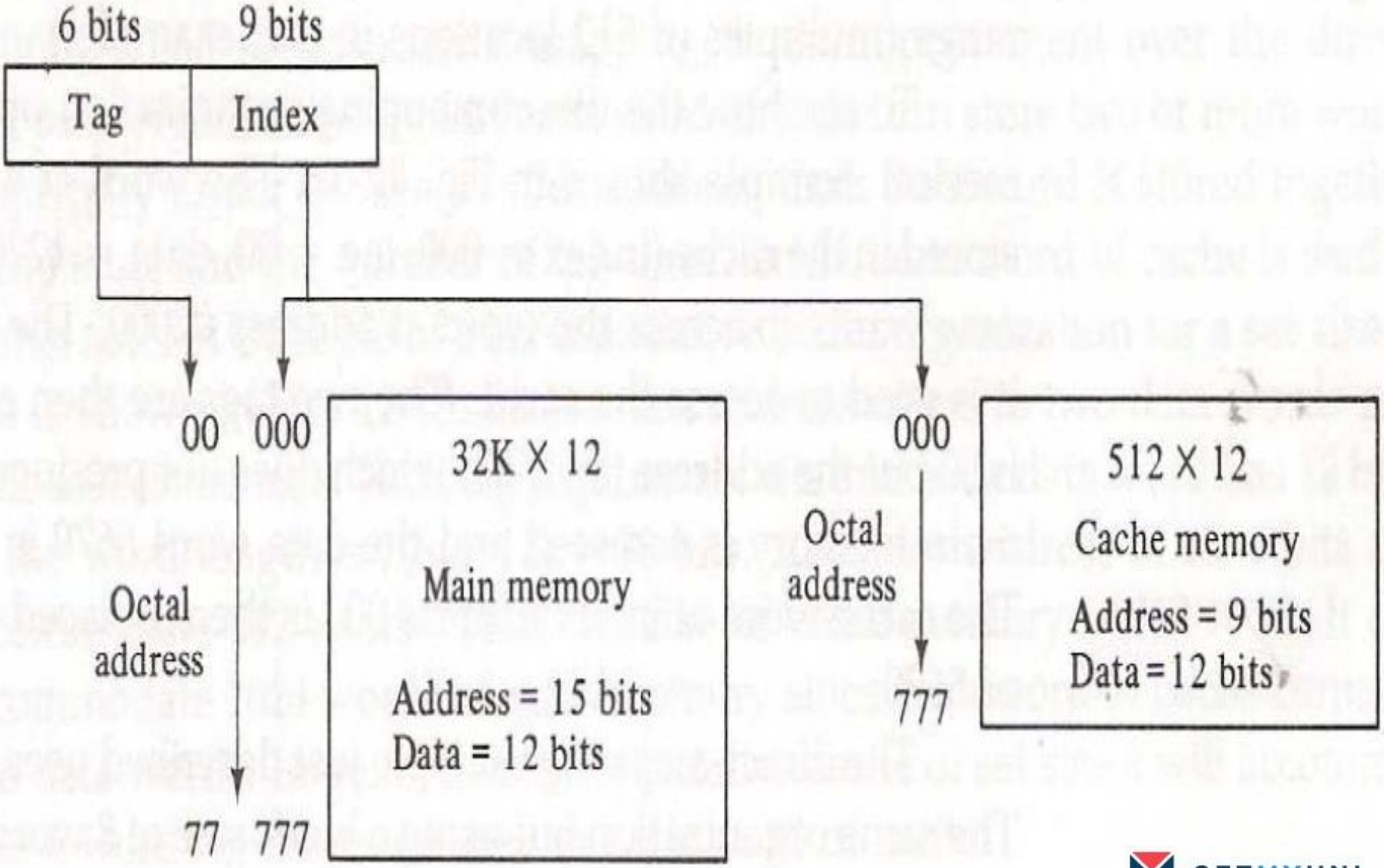
- Fastest and most flexible cache organization uses associative memory.
- It stores both address and content of memory word.
- Address is placed in argument register and memory is searched for matching address.
- If address is found corresponding data is read.
- If address is not found, it is read from main memory and transferred to cache.

- If the cache is full, an address- word pair must be displaced.
- Various algorithm are used to determine which pair to displace. Some of them are FIFO(First In First Out), LRU(Least Recently Used) etc.

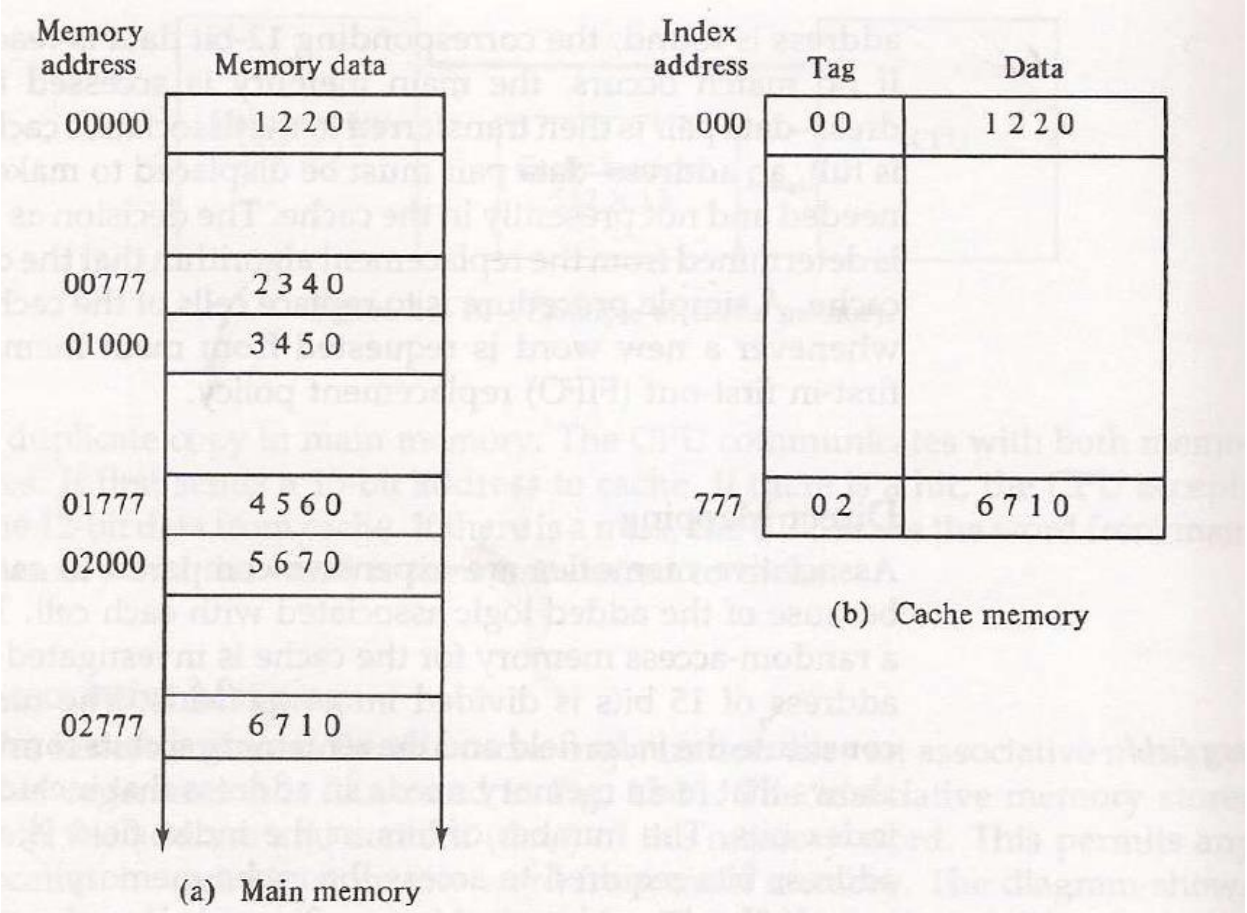
Direct Memory

- CPU address is divided into two fields tag and index.
- Index field is required to access cache memory and total address is used to access main memory.
- If there are 2^k words in cache and 2^n words in main memory, then n bit memory address is divided into two parts. k bits for index field and n-k bits for tag field.

Figure Addressing relationships between main and cache memories.



Direct Mapping Cache Organization



- When CPU generates memory request, index field is used to access the cache.
- Tag field of the CPU address is compared with the tag in the word read. If the tag match, there is hit.
- If the tag does not match, word is read from main memory and updated in cache.
- This example use the block size of 1.
- The same organization can be implemented for block size 8.

- The index field is divided into two parts: block field and word field.
- In 512 word cache there are 64 blocks of 8 words each(64*8=512).
- Block is specified with 6 bit field and word within block with 3 bit field.
- Every time miss occur, entire block of 8 word is transferred from main memory to cahche.

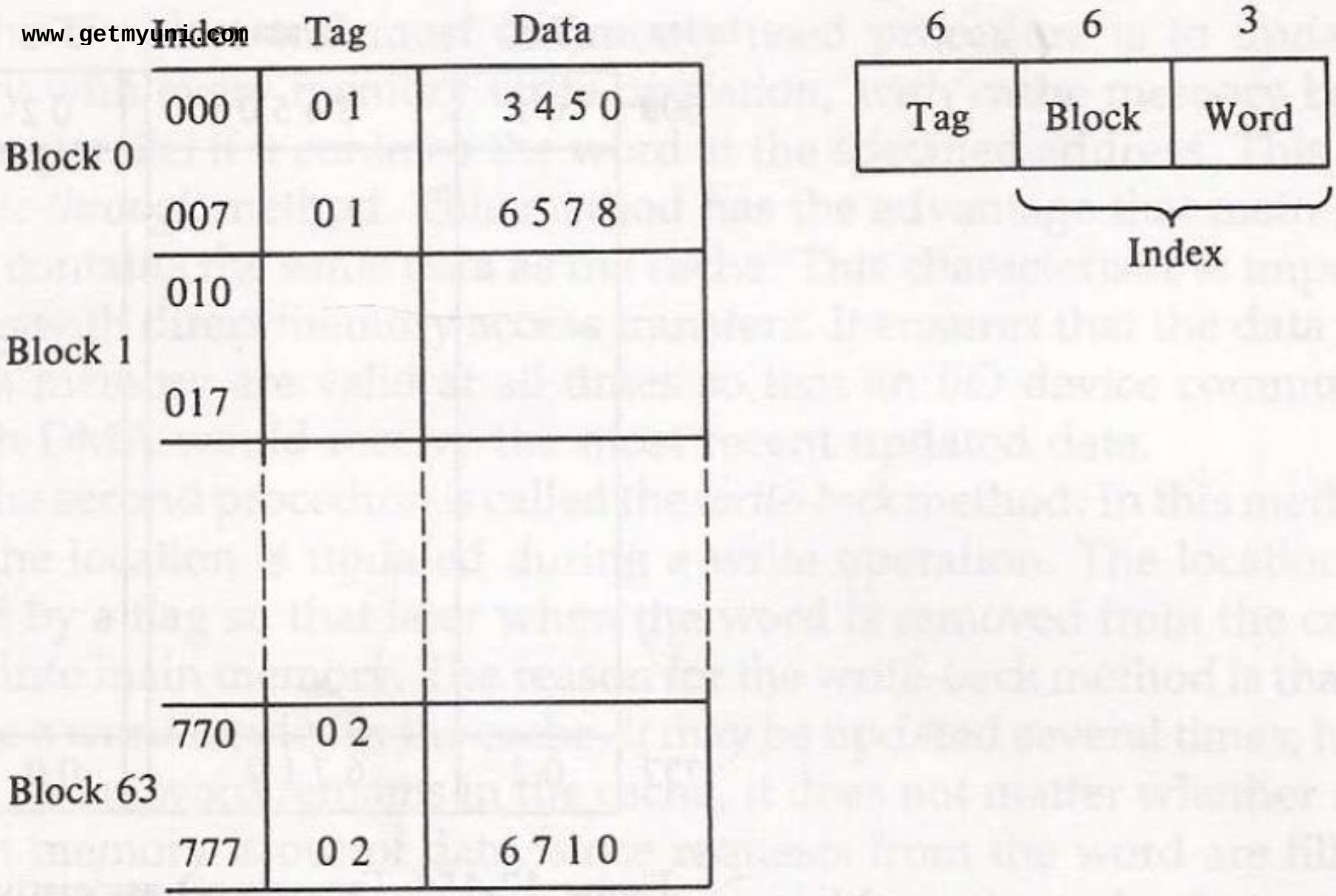
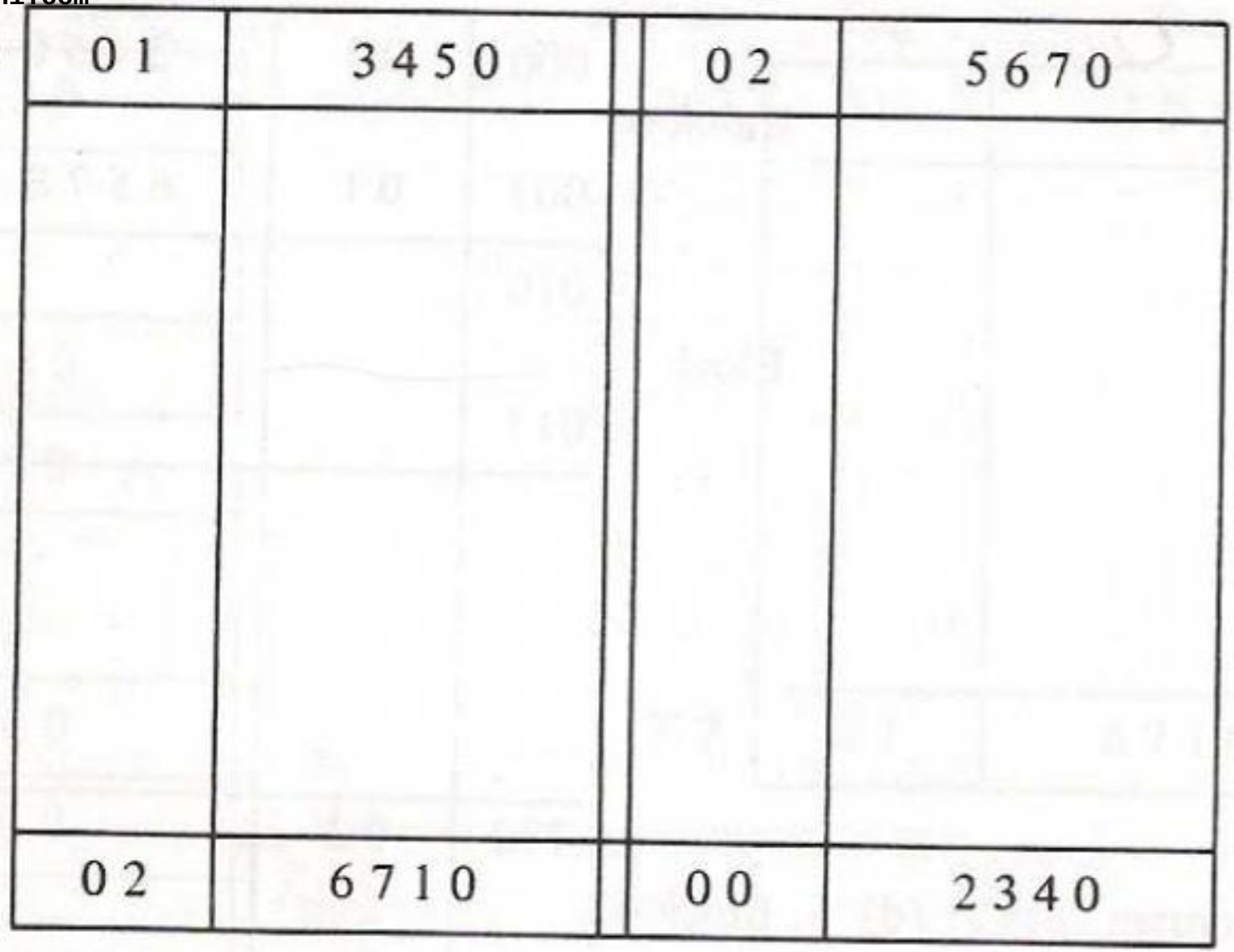


Figure Direct mapping cache with block size of 8

Set-Associative Mapping

- In direct mapping two words with same index in their address but different tag values can’t reside simultaneously in memory.
- In this mapping, each data word is stored together with its tag and number of tag-data items in one word of the cache is said to form set.
- In general, a set associative cache of set size k will accommodate k words of main memory in each word of cache.



Index	Tag	Data	Tag	Data
000	0 1	3 4 5 0	0 2	5 6 7 0
777	0 2	6 7 1 0	0 0	2 3 4 0

Figure Two-way set-associative mapping ca

- When a miss occur and the set is full, one of the tag data item is replaced with new value using various algorithm.

Writing into Cache

- Writing into cache can be done in two ways:
  - Write through
  - Write Back
- In write through, whenever write operation is performed in cache memory, main memory is also updated in parallel with the cache.
- In write back, only cache is updated and marked by the flag. When the word is removed from cache, flag is checked if it is set the corresponding address in main memory is updated.

Cache Initialization

- When power is turned on, cache contain invalid data indicated by valid bit value 0.
- Valid bit of word is set whenever the word is read from main memory and updated in cache.
- If valid bit is 0, new word automatically replace the invalid data.

Virtual Memory

- Virtual memory is a concept used in computer that permit the user to construct a program as though large memory space is available equal to auxiliary memory.
- It give the illusion that computer has large memory even though computer has relatively small main memory.
- It has mechanism that convert generated address into correct main memory address.

Address Space and Memory Space

- An address used by the programmer is called virtual address and set of such address is called address space.
- An address in main memory is called physical address. The set of such location is called memory space.

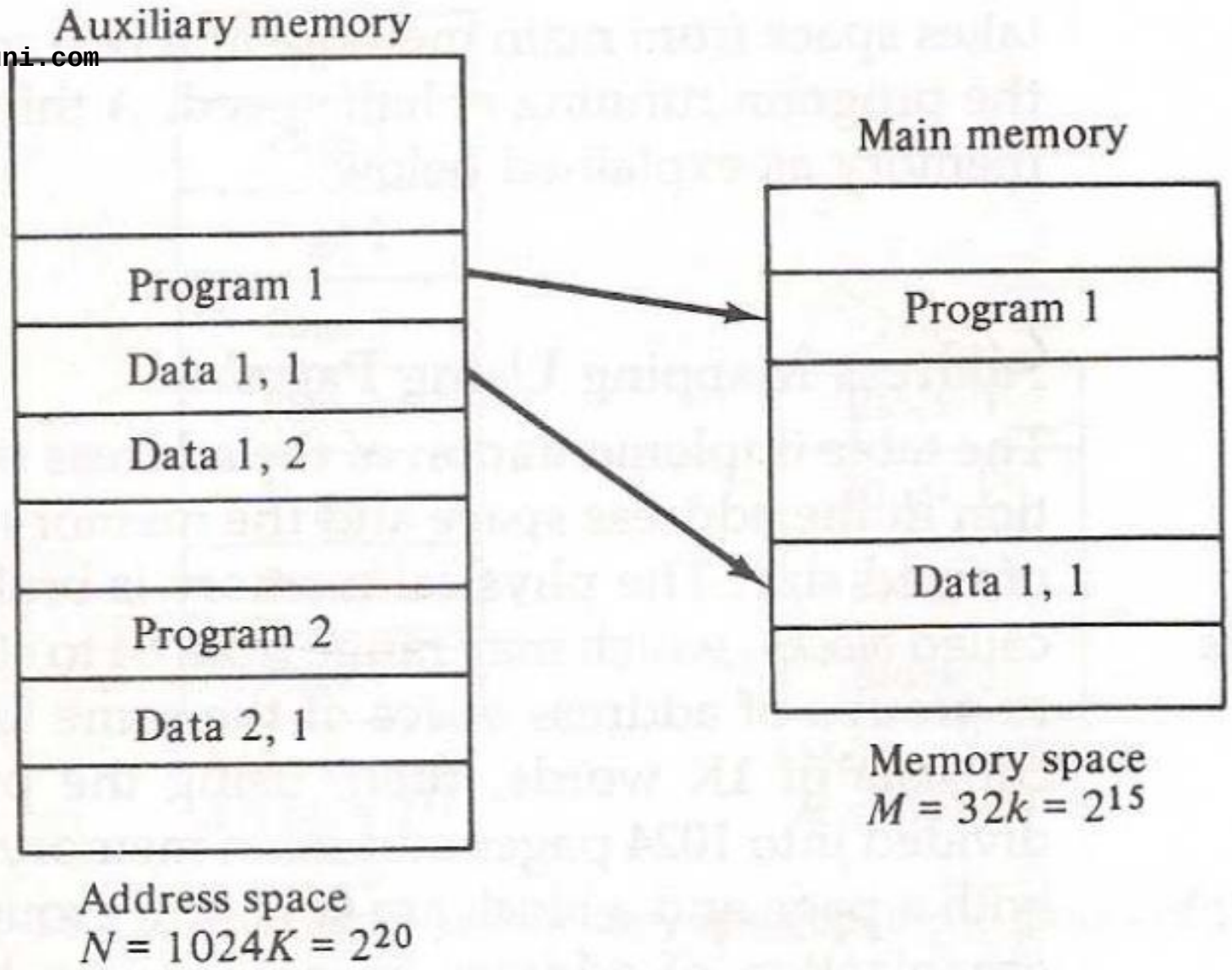
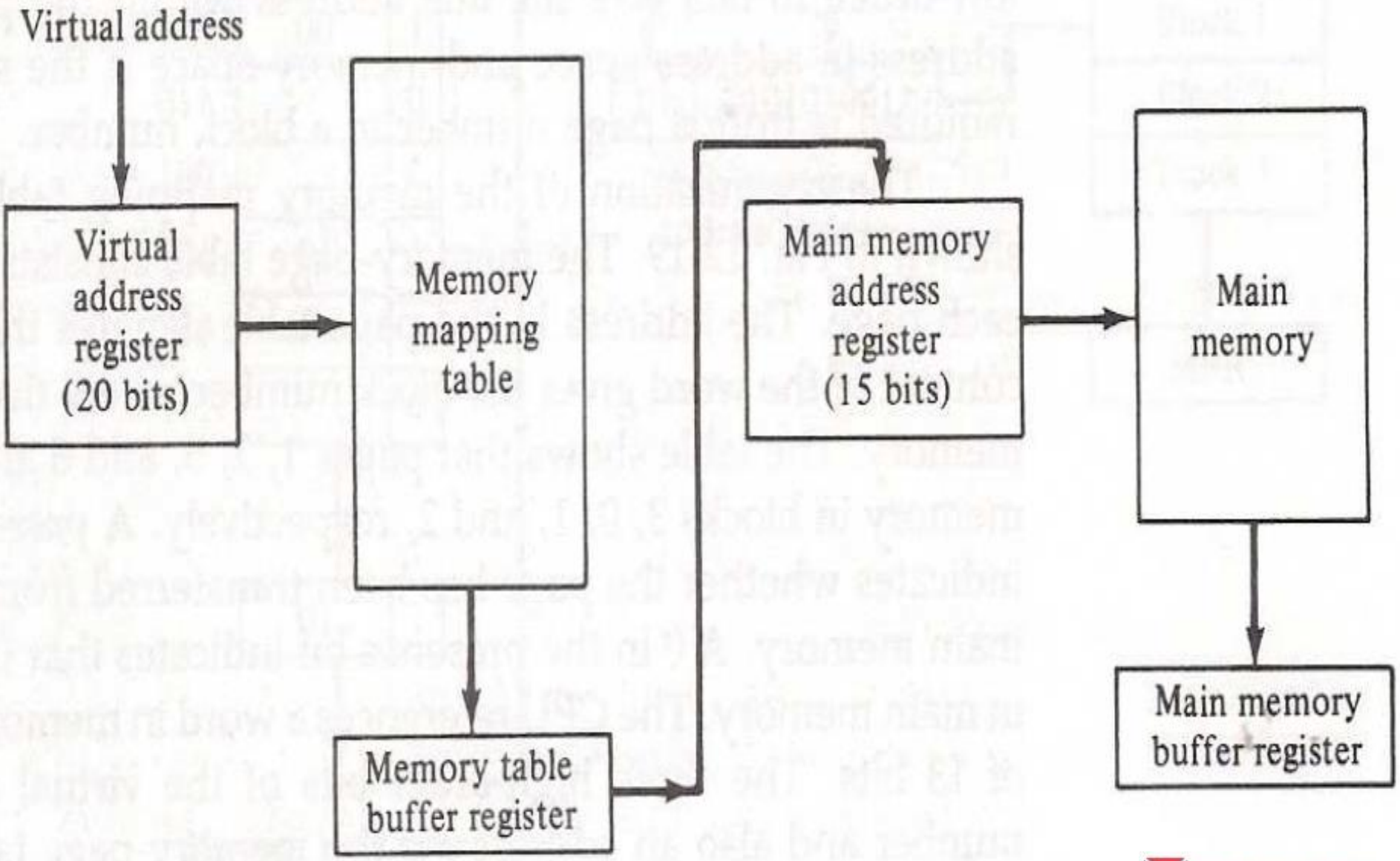


Figure Relation between address and memory space in a virtual memory system.

Figure Memory table for mapping a virtual address.



Address Mapping Using Pages

- The main memory is broken down into groups of equal size called blocks.
- Term pages refers to groups of address space of same size.
- Although page and block are of equal size, page refer to organization of address space and block represent the organization of memory space.
- The term page frame is sometimes used to denote block.

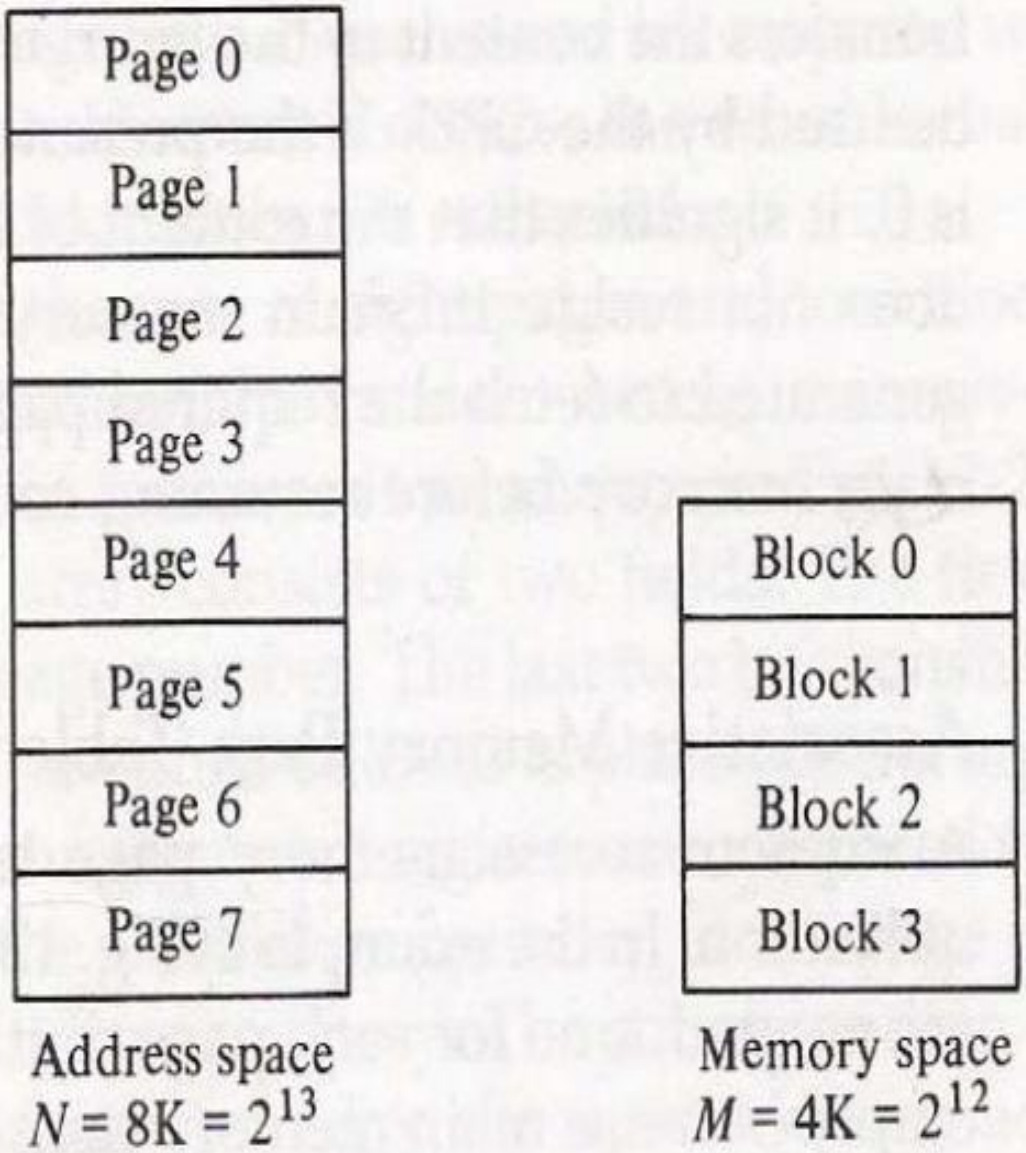


Figure Address space and memory space split into groups of 1K words.

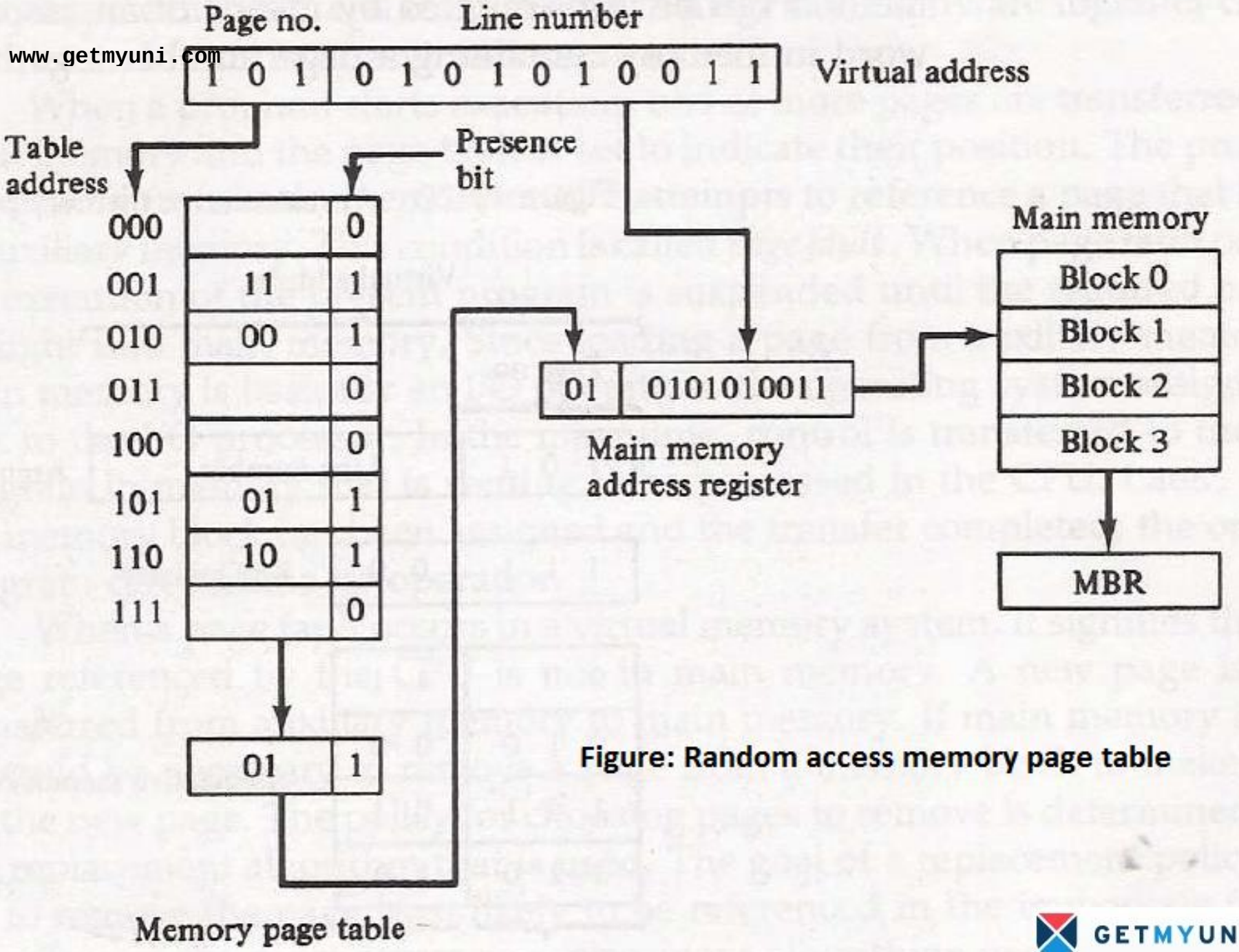
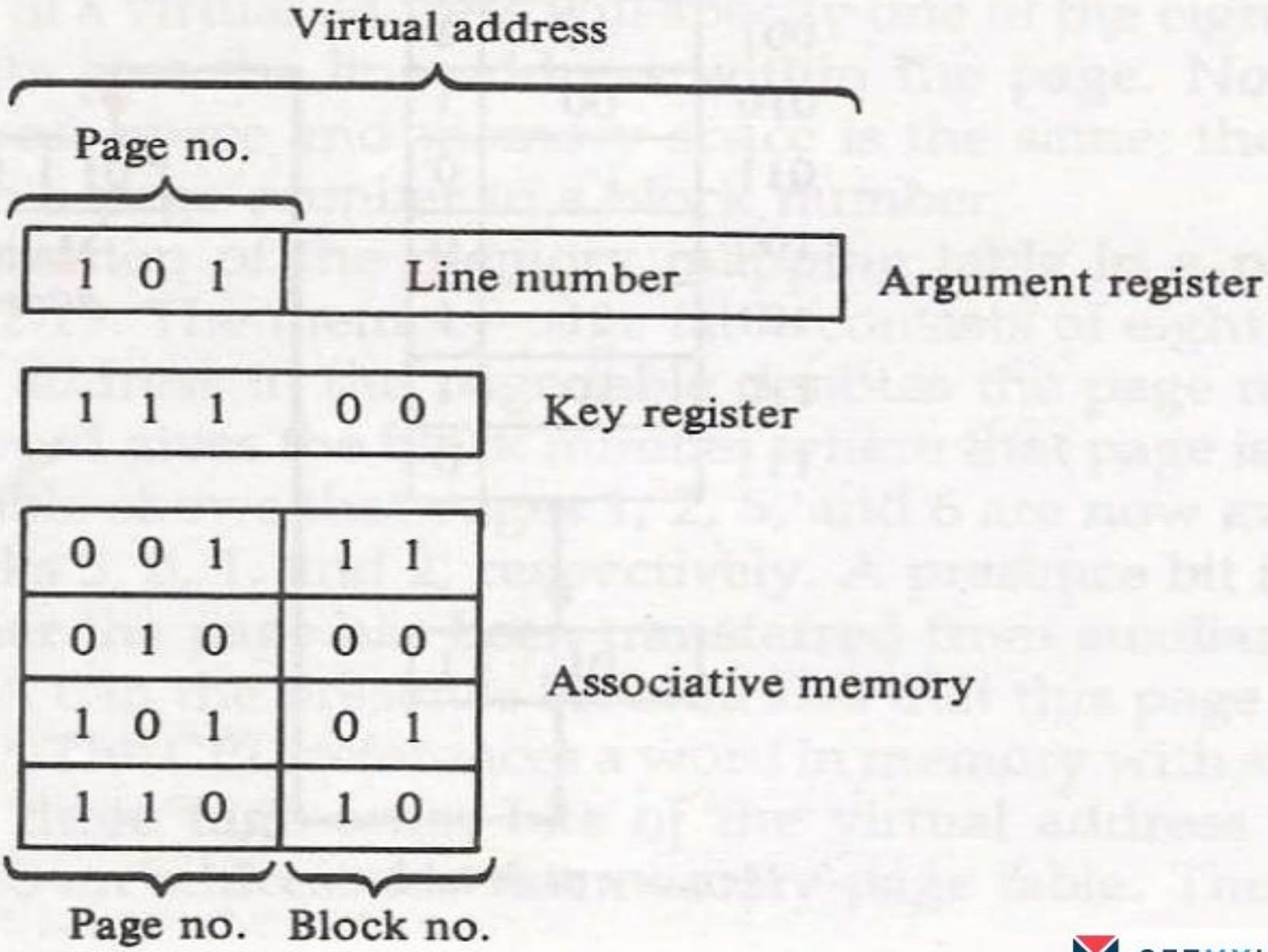


Figure: Random access memory page table

Figure An associative memory page table.



Page Replacement

- The program is executed from main memory until page required is not available.
- If page is not available, this condition is called page fault. When it occurs, present program is suspended until the page required is brought into main memory.
- If main memory is full, pages to remove is determined from the replacement algorithm used.