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

- Storage Capacity increases.
- Access time increases

- Cost per bit decreases.
- Frequency of access by the processor decreases.

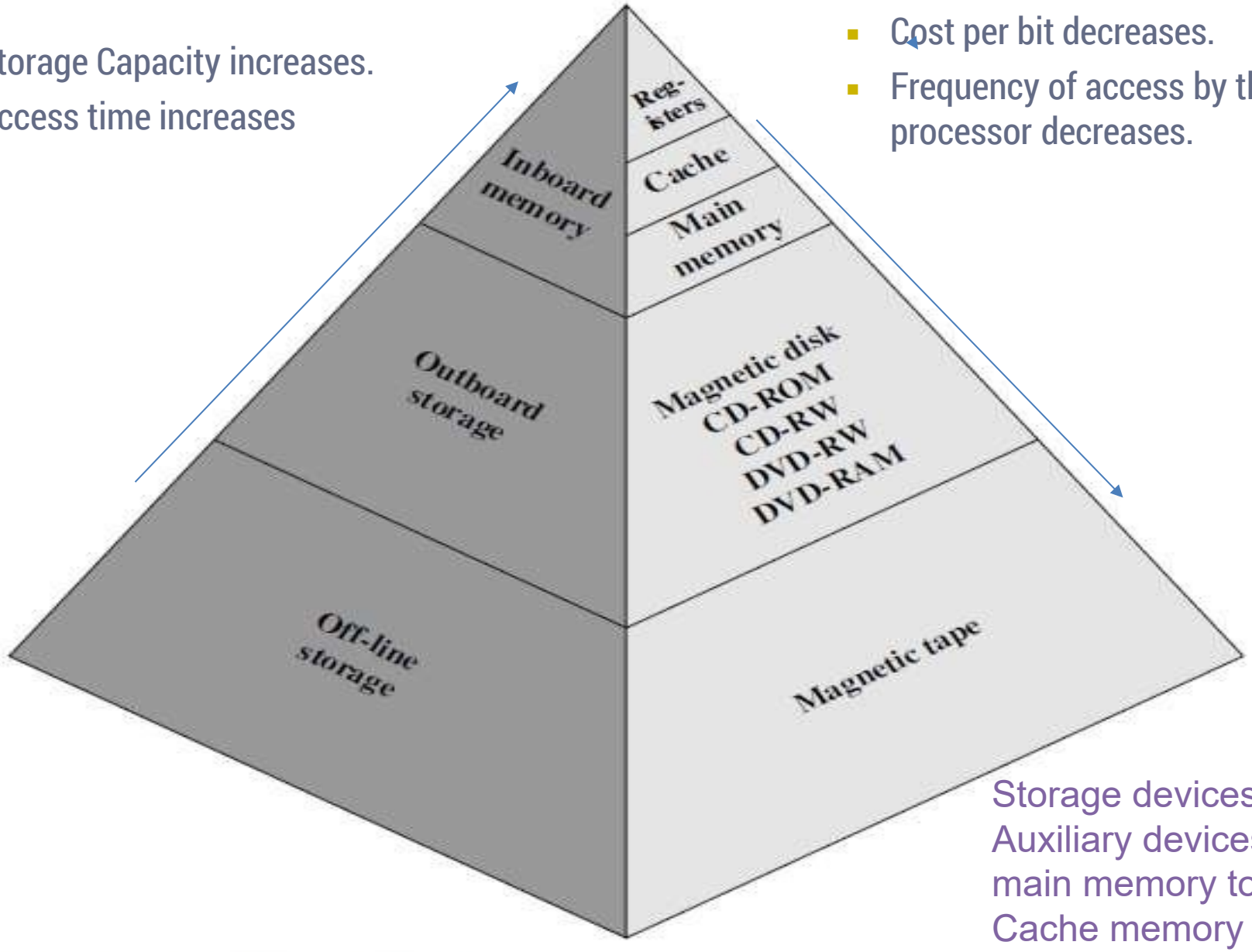
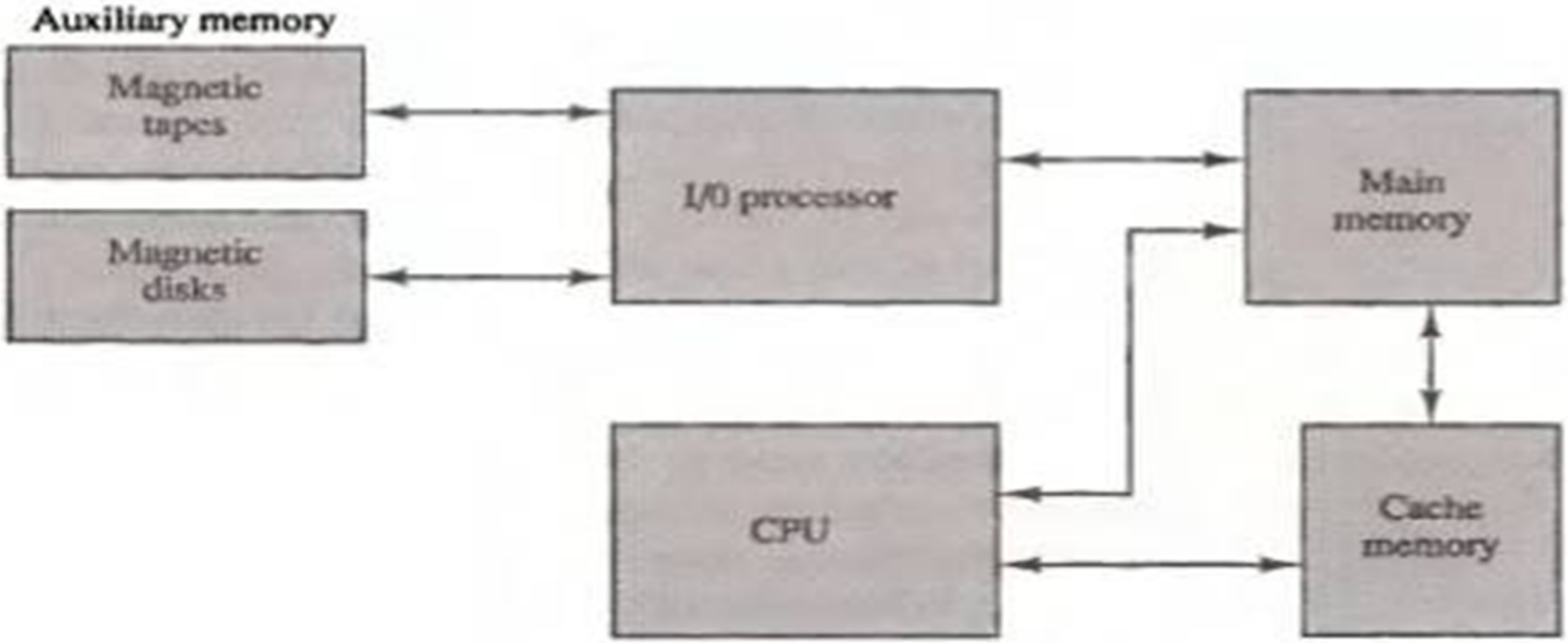


Figure 4.1 The Memory Hierarchy

Storage devices from Auxiliary devices to main memory to Cache memory

Memory hierarchy in a computer system



1. Programs and data currently needed by processor reside by main memory and transferred whenever needed.I/O processor manages the transfers between auxiliary memory and main memory
2. Cache memory transfer of information between main memory and CPU(frequently used).
3. CPU has direct access to both cache and main memory

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.

Main Memory-RAM-Random Access Memory

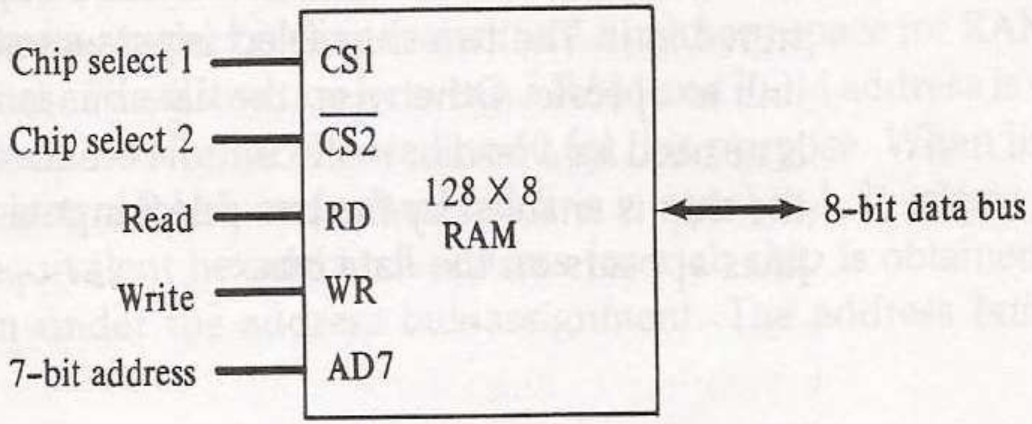
SRAM	DRAM
Uses capacitor for storing information	Uses Flip flop
More cells per unit area due to smaller cell size.	Needs more space for same capacity
Cheap and smaller in size	Expensive and bigger in size
Slower and analog device	Faster and digital device
Requires refresh circuit	No need
Used in main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

Main Memory-ROM

- ROM is uses Read Only Memory.
- 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.

Main Memory-RAM Chip

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

Bidirectional bus allows transfer to and from the CPU to RAM.
Bus is constructed with three state buffer.
Capacity is 128 words of eight bits per word(128 x 8)
AD7-128=2^7= 7 bit address bus
8 bit bidirectional data bus
CS-Chip select
This works when CS1=1,CS2'=0
WR is enabled stores 8 bit word into a location specified by address lines.
RD is enabled the content selected is placed into the data bus

- 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.

Main memory ROM Chip

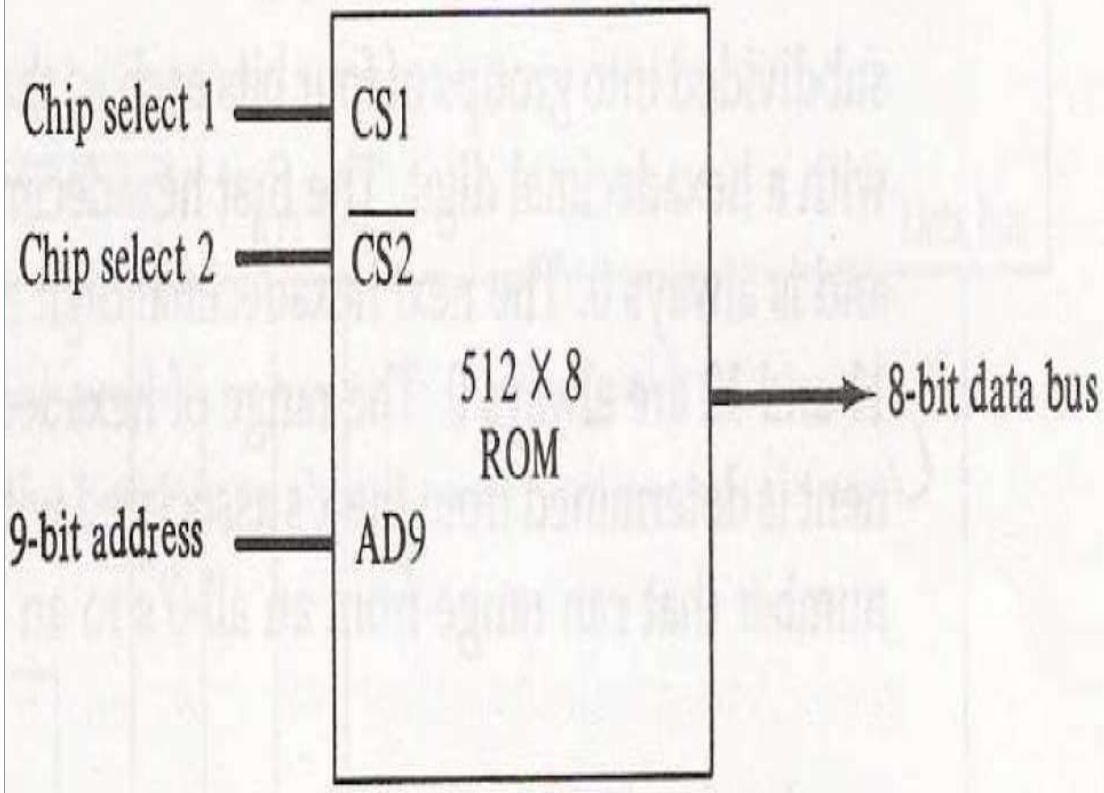


Figure Typical ROM chip.

ROM Chip can read only , the data bus can only be in an output mode.

AD9=512=2^9

9 bit Address bus

8 bit unidirectional Data bus

CS1=1 CS2'=0

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.
- Design 1024 bytes RAM and 1024 bytes of 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

To Design 512 RAM- 4 X 128 - 4 RAM's
RAM-128 X 8- 7 bit Address line
8,9 bits are used to select among 4 RAM's
00-RAM1
01-RAM2
10-RAM3
11-RAM4
10 bit for ROM chip

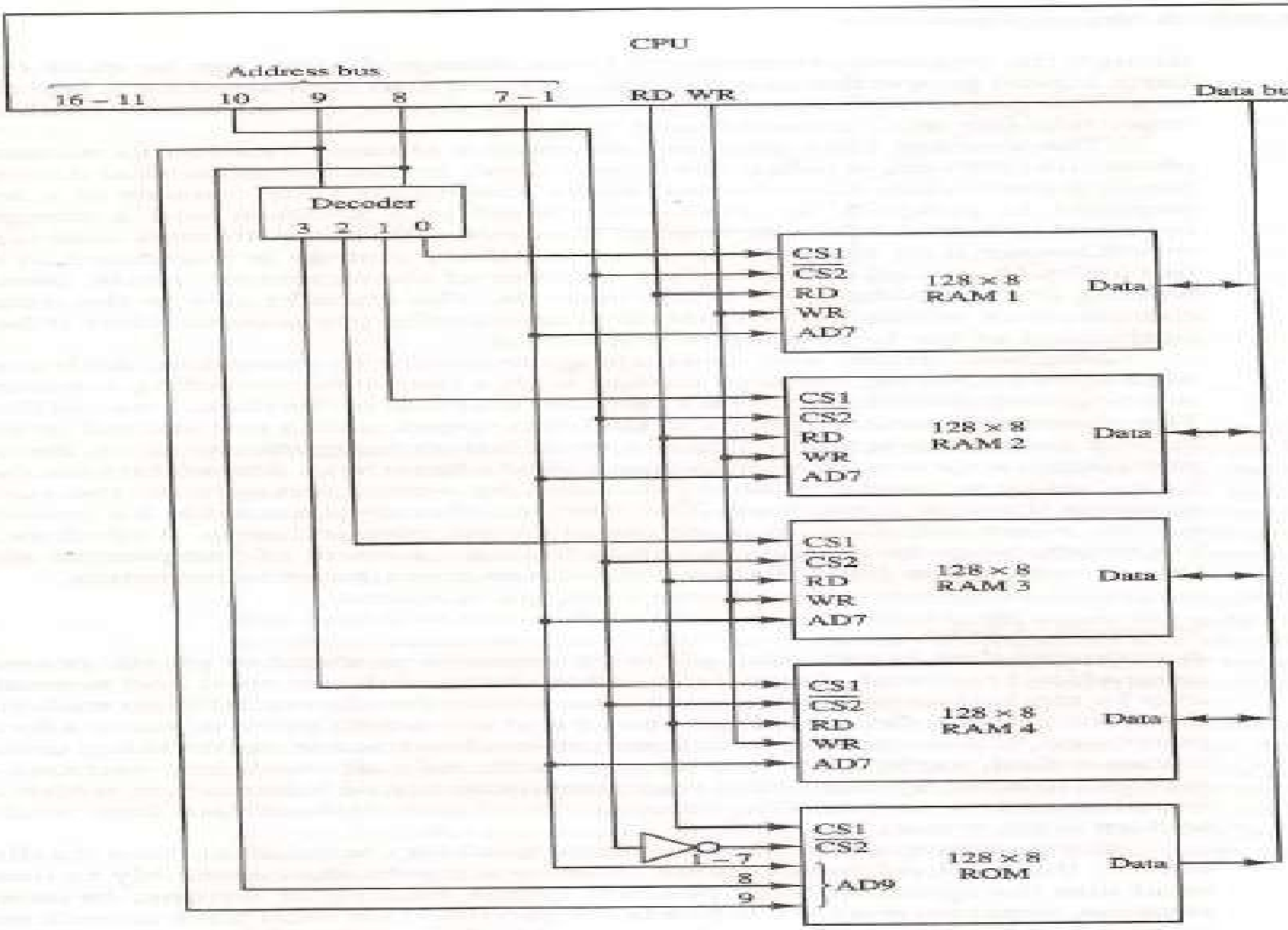


Figure 12-4 Memory connection to the CPU.

Auxiliary Memory

qAn Auxiliary memory is known as the lowest-cost, highest-capacity and slowest-access storage in a computer system.

qIt is where programs and data are kept for long-term storage or when not in immediate use.

qThe most common examples of auxiliary memories are magnetic tapes and magnetic disks.

Auxiliary Memory

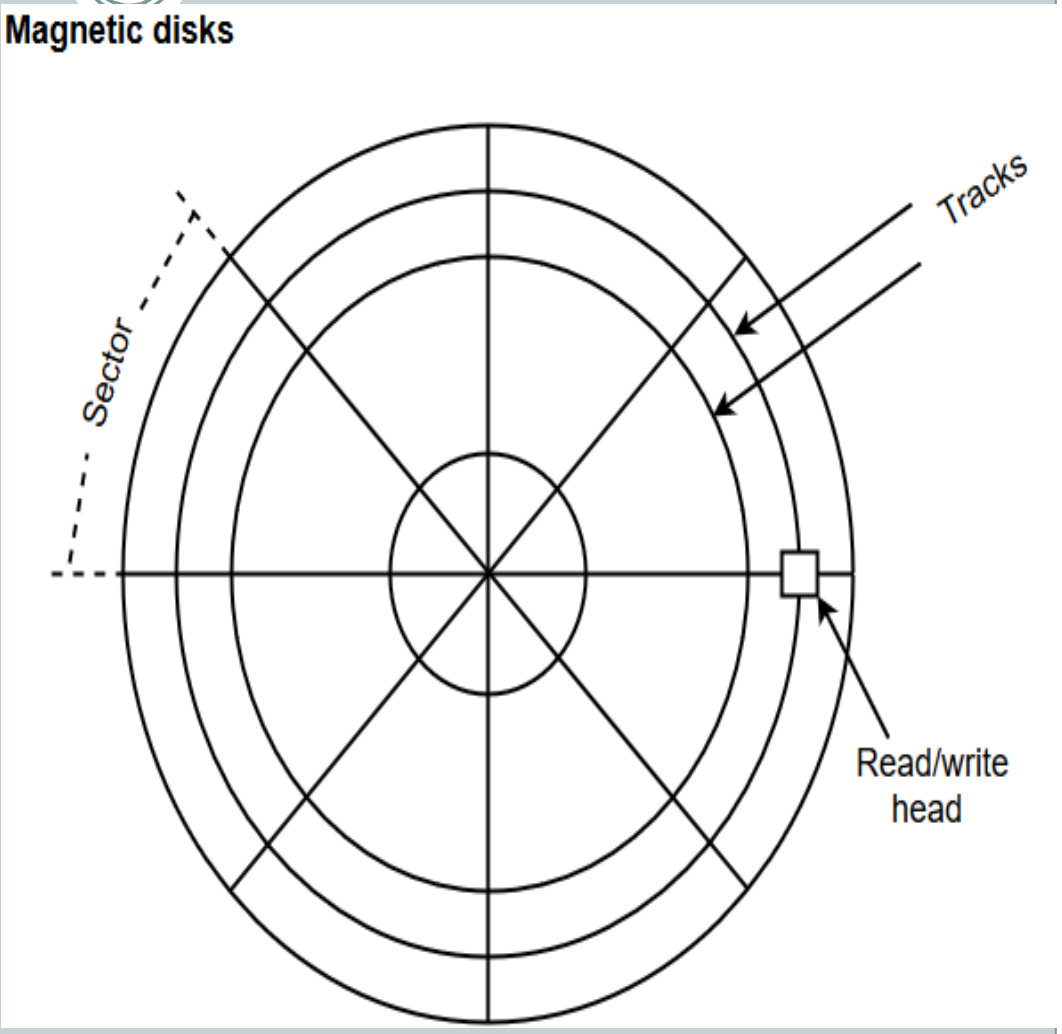
qA magnetic disk is a type of memory constructed using a circular plate of metal or plastic coated with magnetized materials.

qUsually, both sides of the disks are used to carry out read/write operations.

qHowever, several disks may be stacked on one spindle with read/write head available on each surface.

Auxiliary Memory

- The memory bits are stored in the magnetized surface in spots along the concentric circles called tracks.
- The concentric circles (tracks) are commonly divided into sections called sectors.



Magnetic Tape

qMagnetic tape is a storage medium that allows data archiving, collection, and backup for different kinds of data.

qThe magnetic tape is constructed using a plastic strip coated with a magnetic recording medium.

qThe bits are recorded as magnetic spots on the tape along several tracks.

Magnetic Tape

qUsually, seven or nine bits are recorded simultaneously to form a character together with a parity bit.

qMagnetic tape units can be halted, started to move forward or in reverse, or can be rewound.

q However, they cannot be started or stopped fast enough between individual characters.

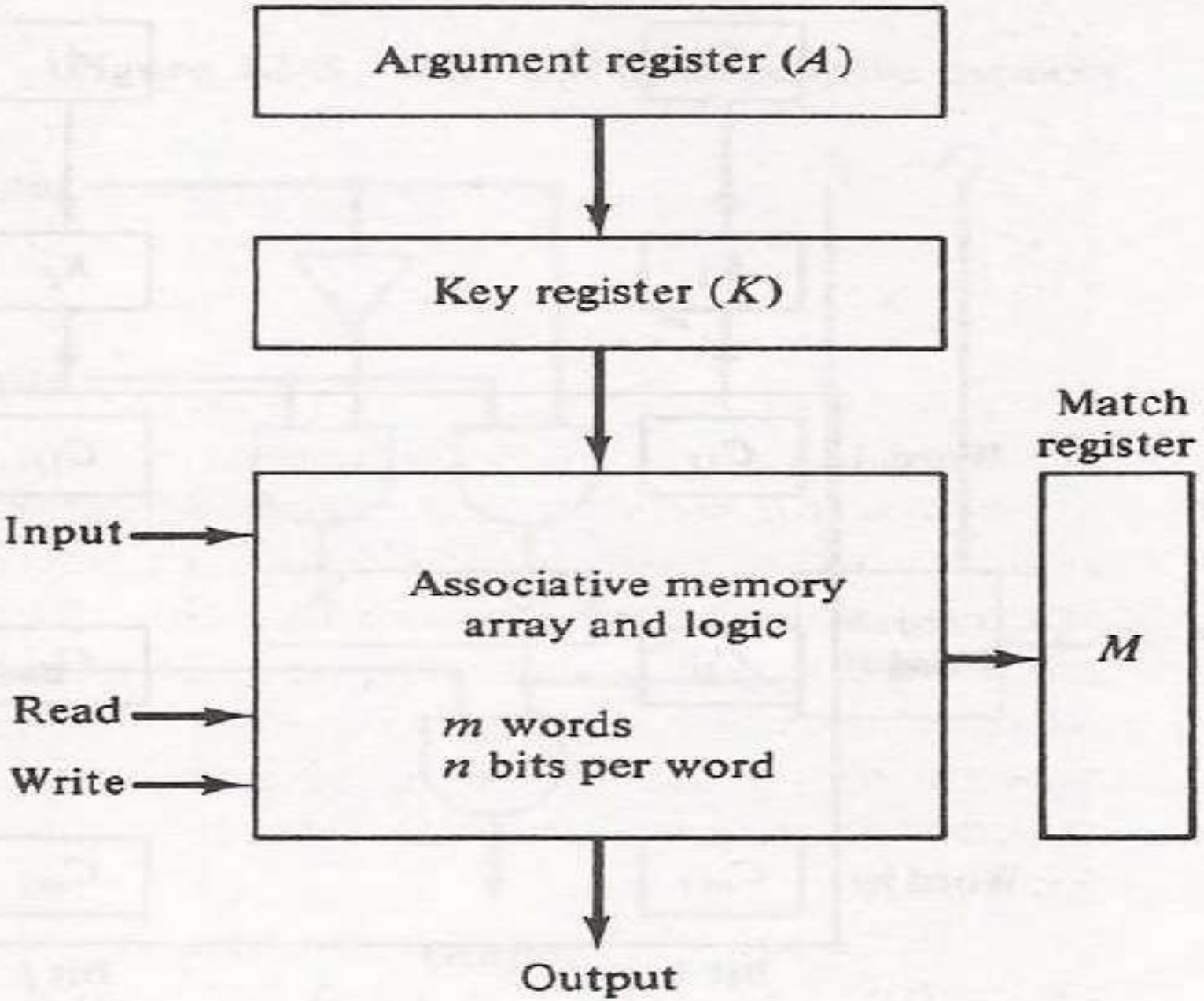
qFor this reason, information is recorded in blocks referred to as records.

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.

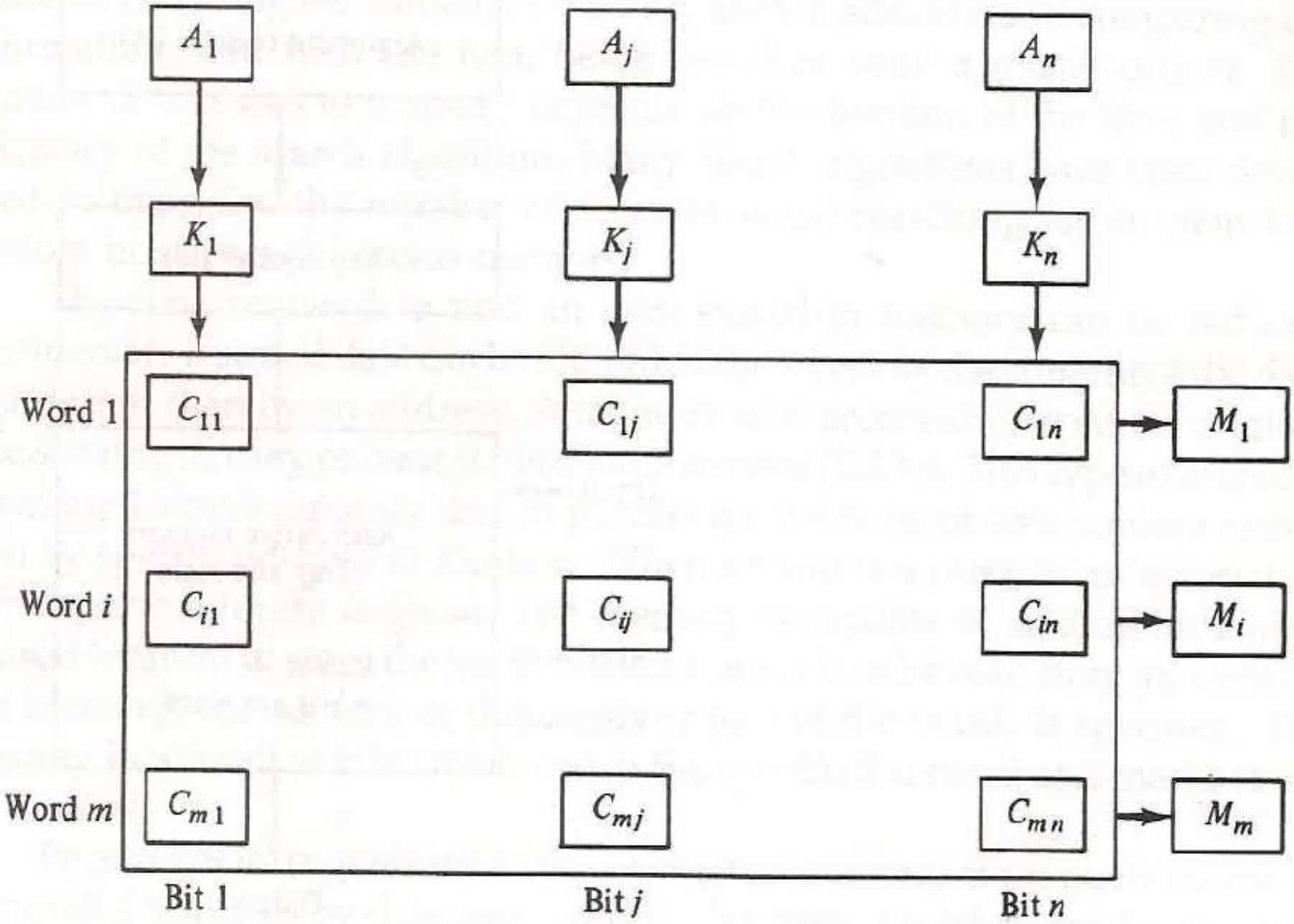


A	101 111100	
K	111 000000	
Word 1	100 111100	no match
Word 2	101 000001	match

- 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.

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



Match Logic

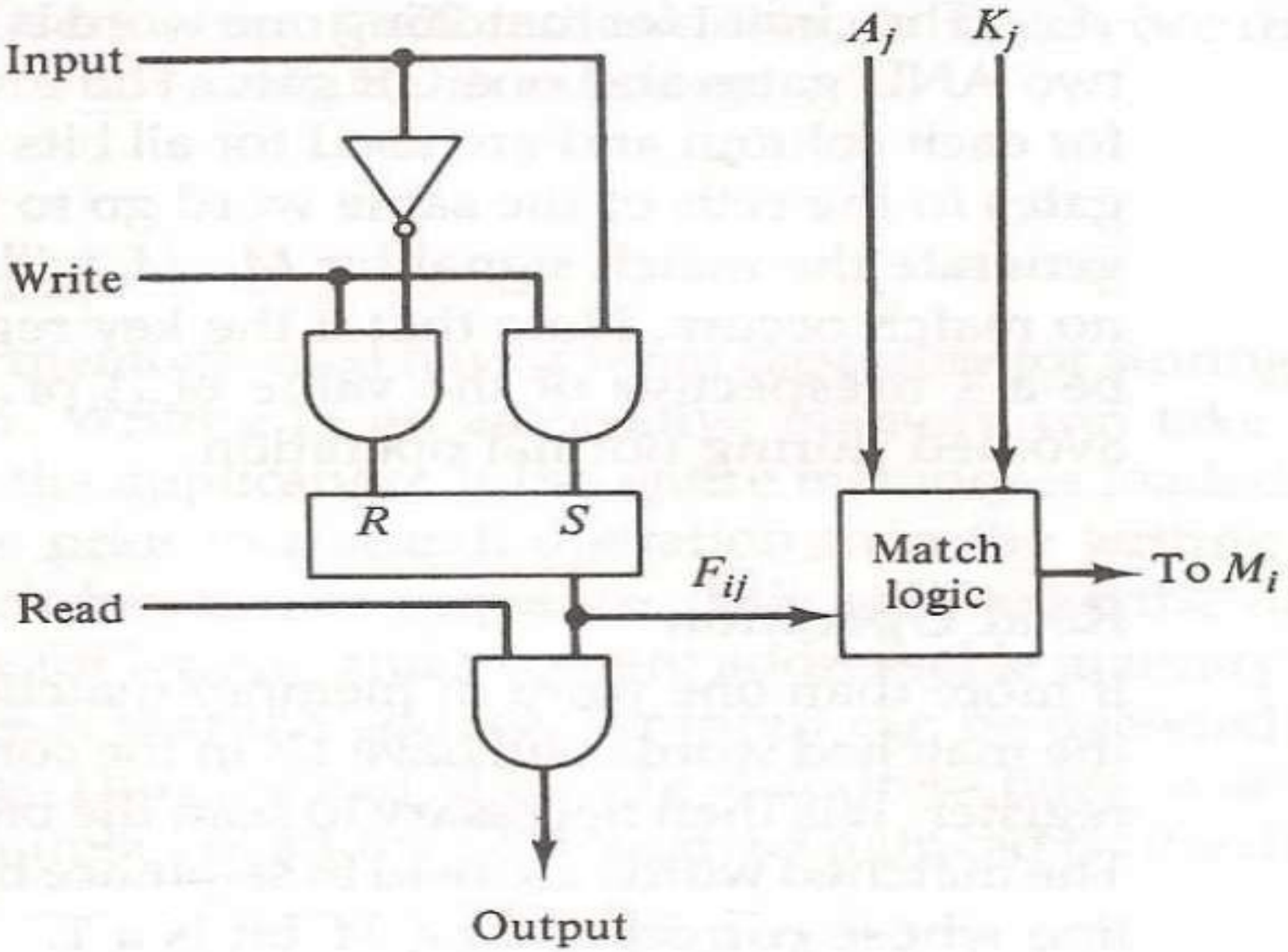
- Let us neglect the key register and compare the content of argument register with memory content.
- Word i is equal to argument in A if $A_j = F_{ij}$ for j=1,2,3,4……..n
- The equality of two bits is expressed as

$$x_j = A_j F_{ij} + A_j' F_{ij}'$$

- x_j =1 if bits are equal and 0 otherwise.

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

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.

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

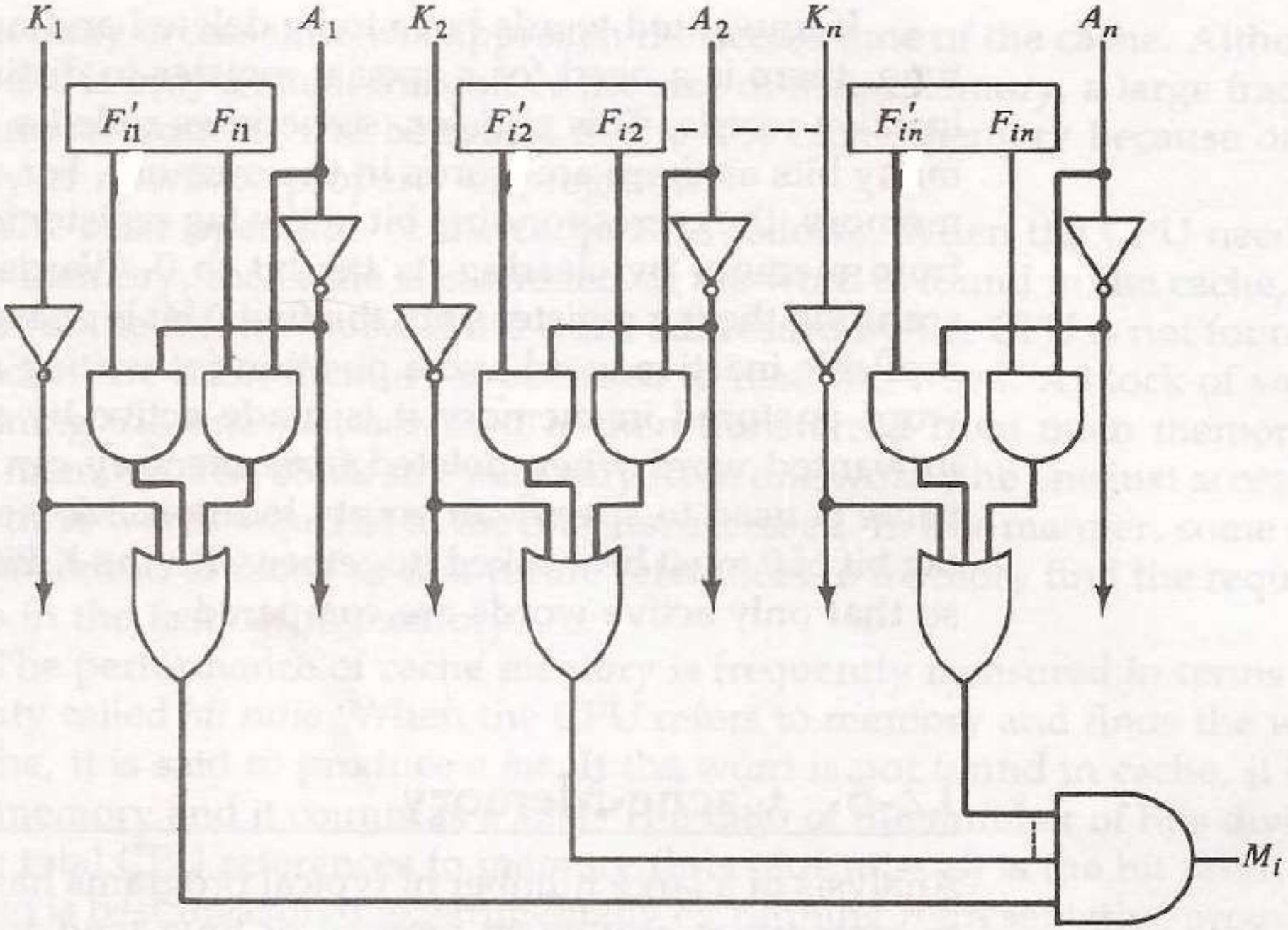


Figure Match logic for one word of associative memory.

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.

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.
- Loops and subroutines tends to localize the memory.
- 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.

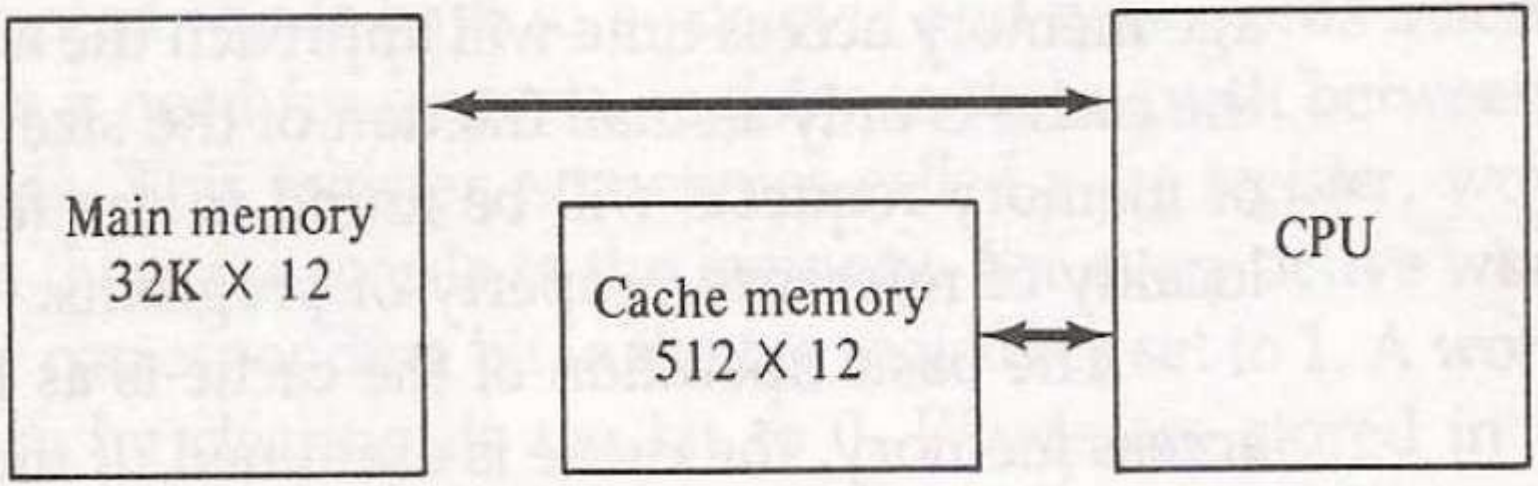


Figure Example of cache memory.

- 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

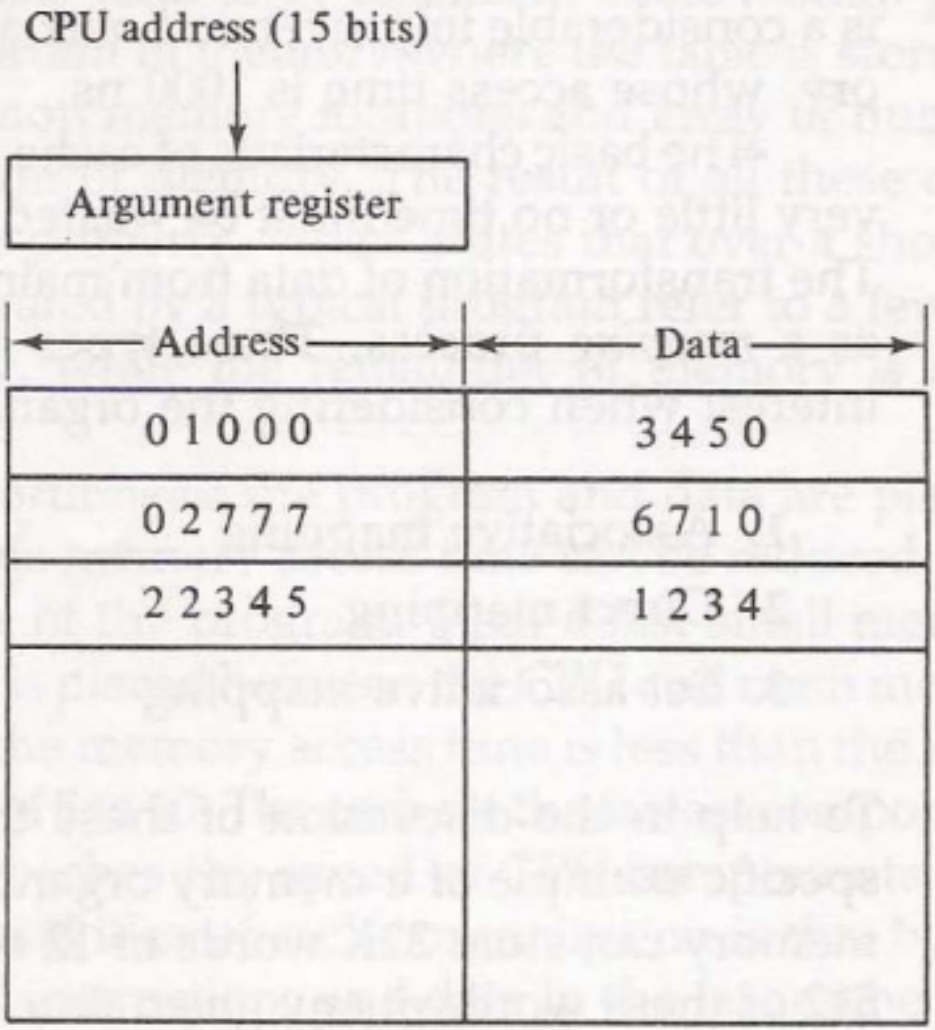
- The transformation of data from main memory to cache is known as mapping process. Three types of mapping procedures are:
  - Associative Mapping
  - Direct Mapping
  - Set-Associative Mapping

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.

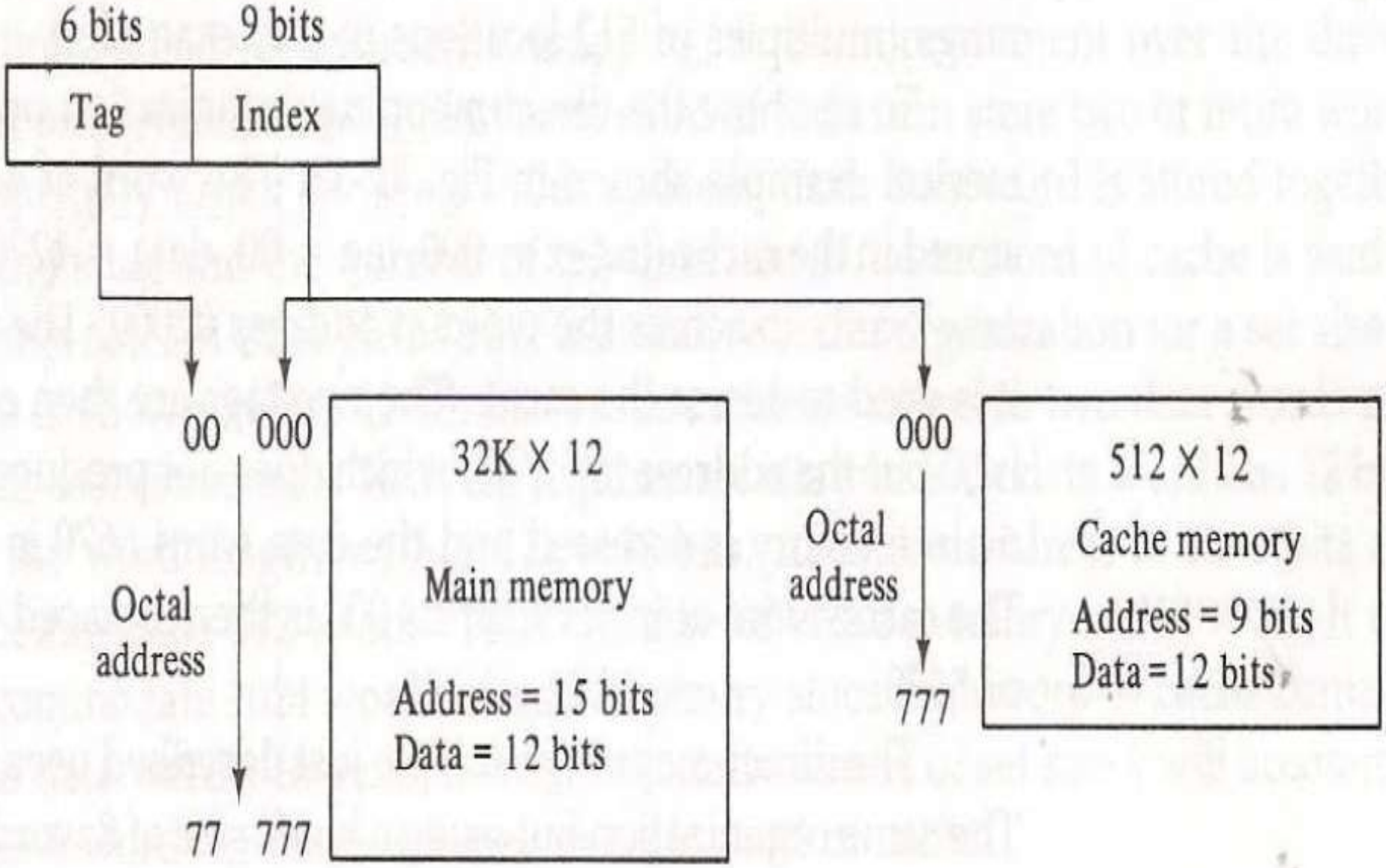
Figure 12-11 Associative mapping cache (all numbers in octal).



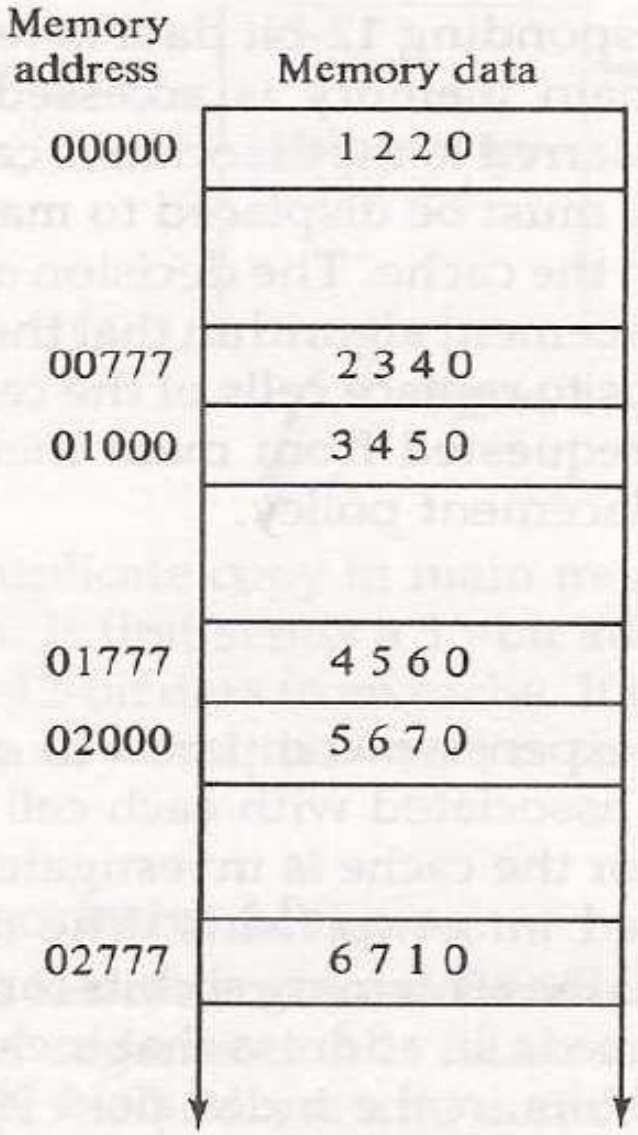
Direct Mapping

- 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.

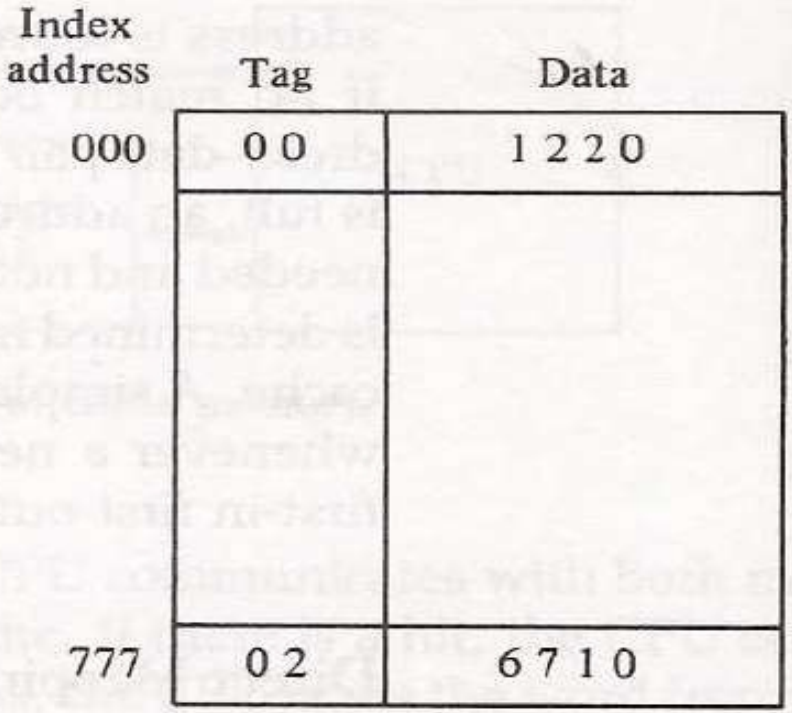
Figure Addressing relationships between main and cache memories.



Direct Mapping Cache Organization



(a) Main memory



(b) Cache memory

- 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 cache.

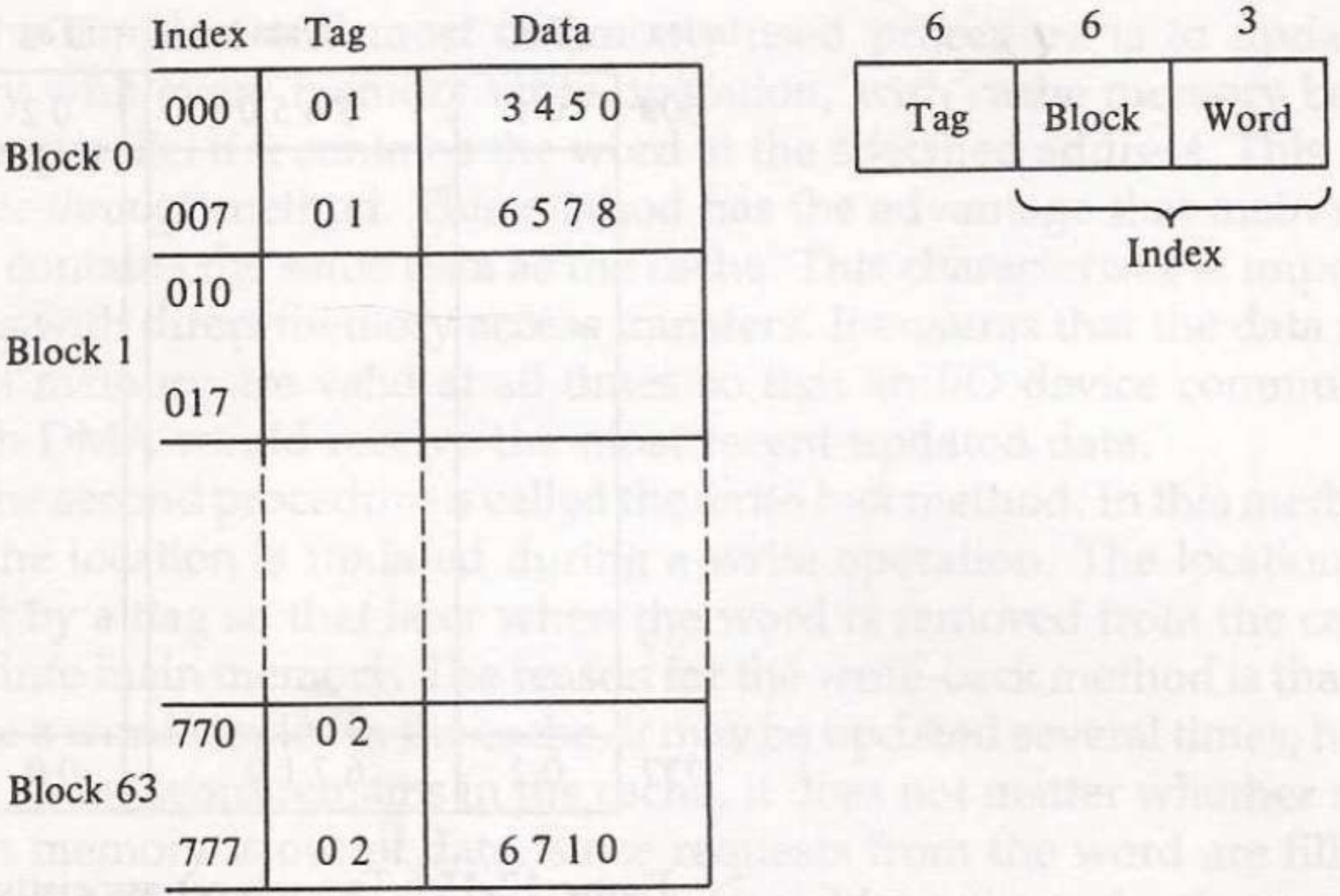


Figure Direct mapping cache with block size of 8 words.

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.

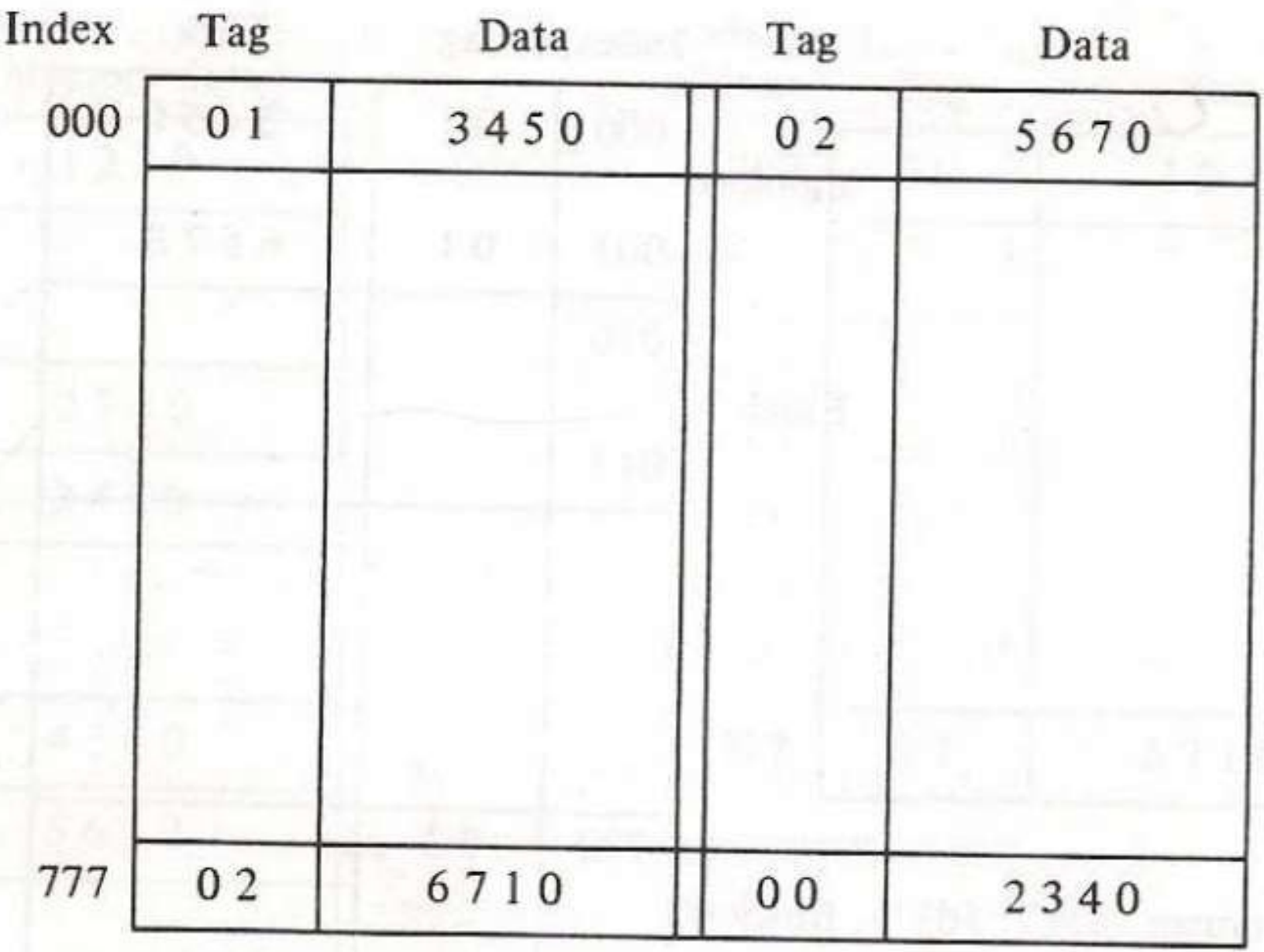


Figure Two-way set-associative mapping cache.

- 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.

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.

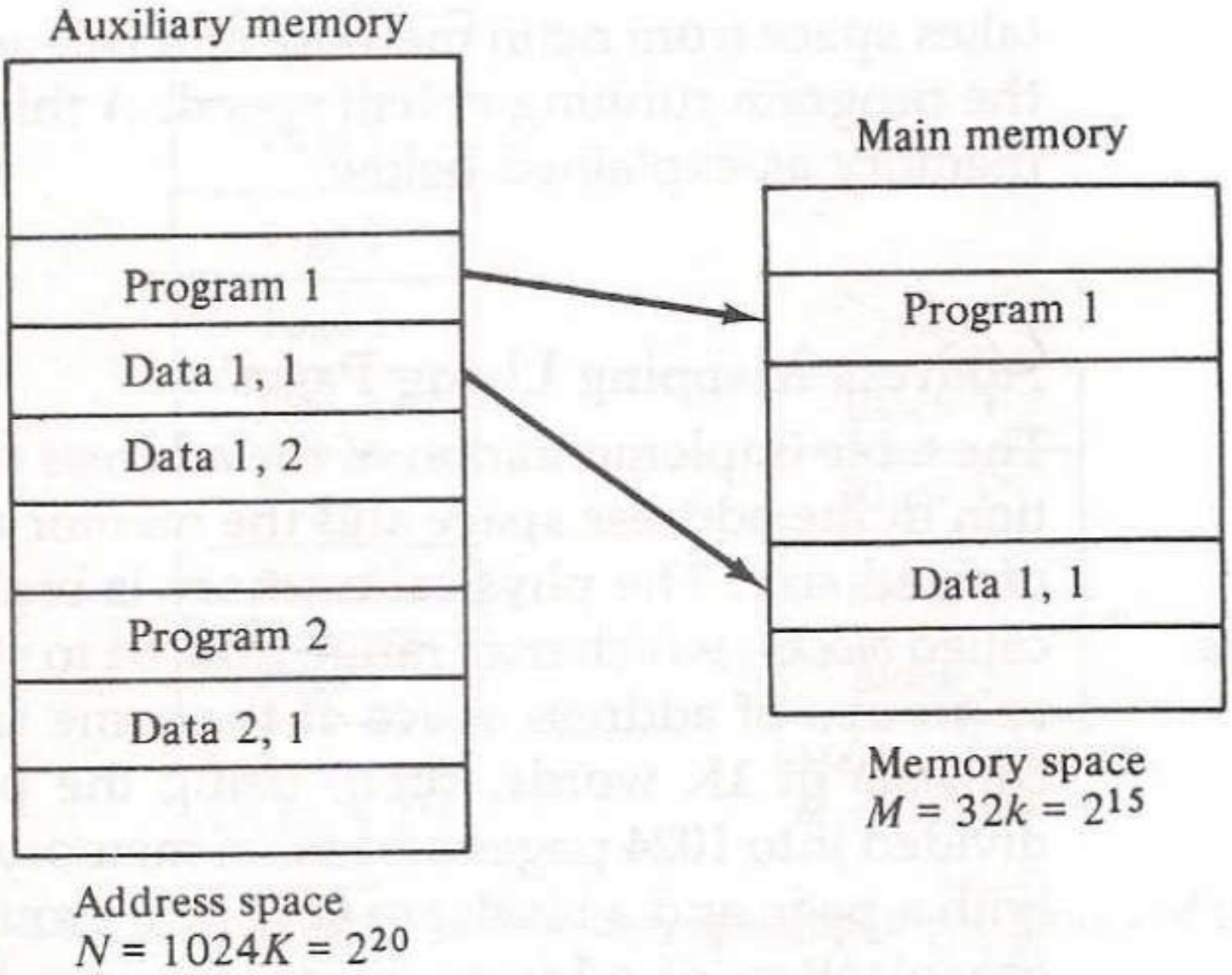
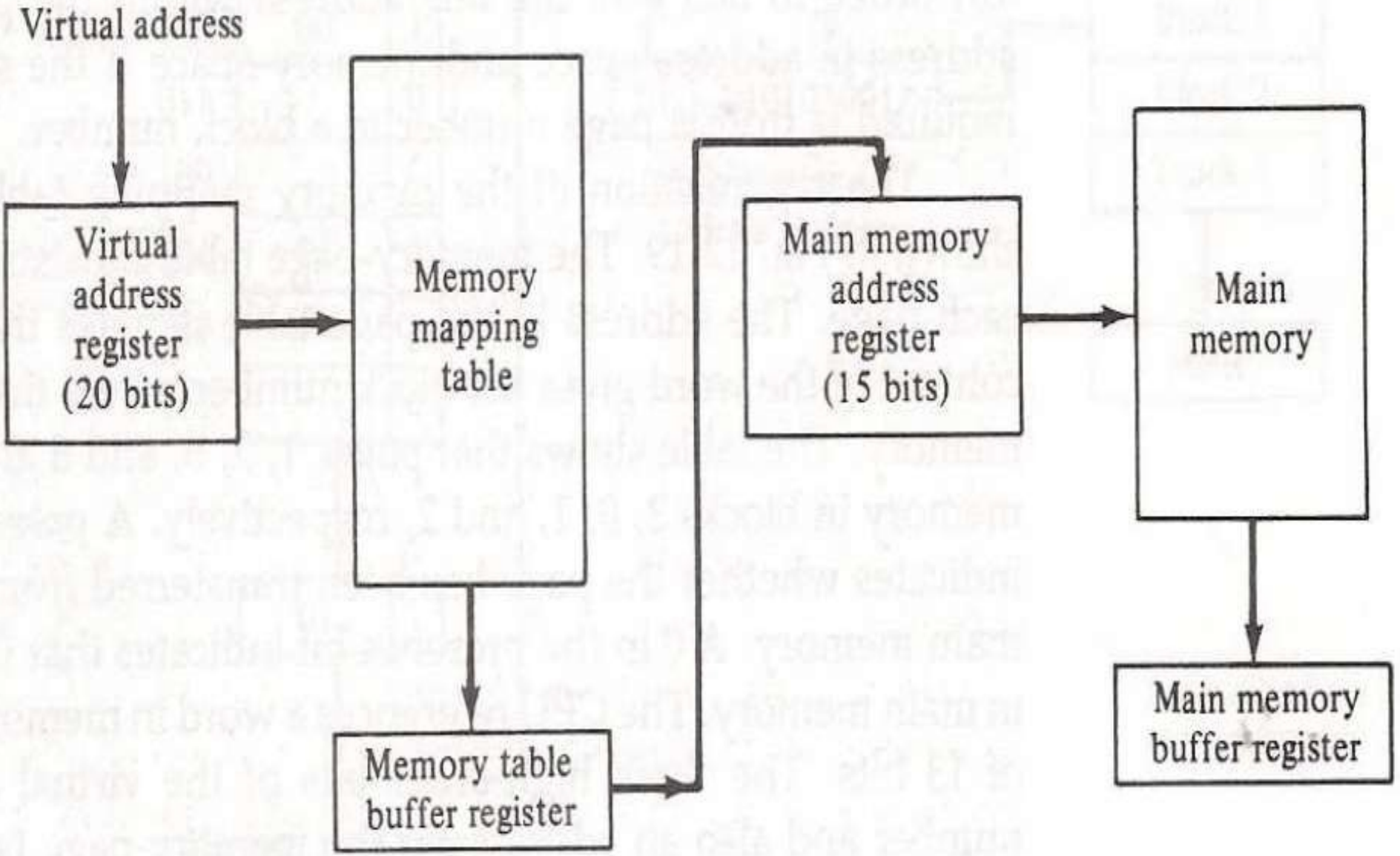


Figure Relation between address and memory space in a virtual memory system.

Figure Memory table for mapping a virtual address.



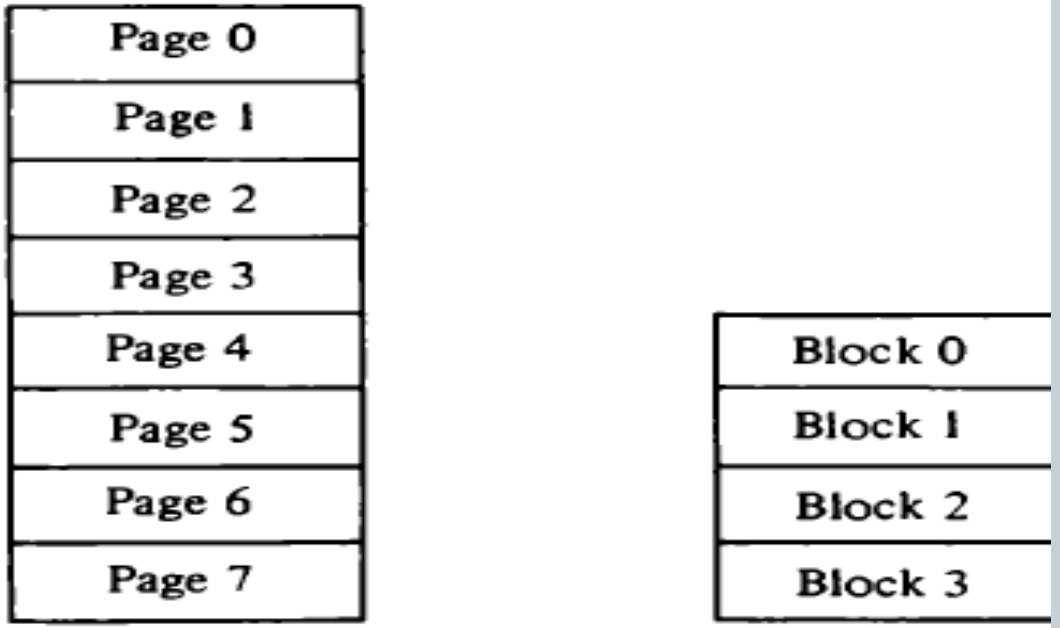
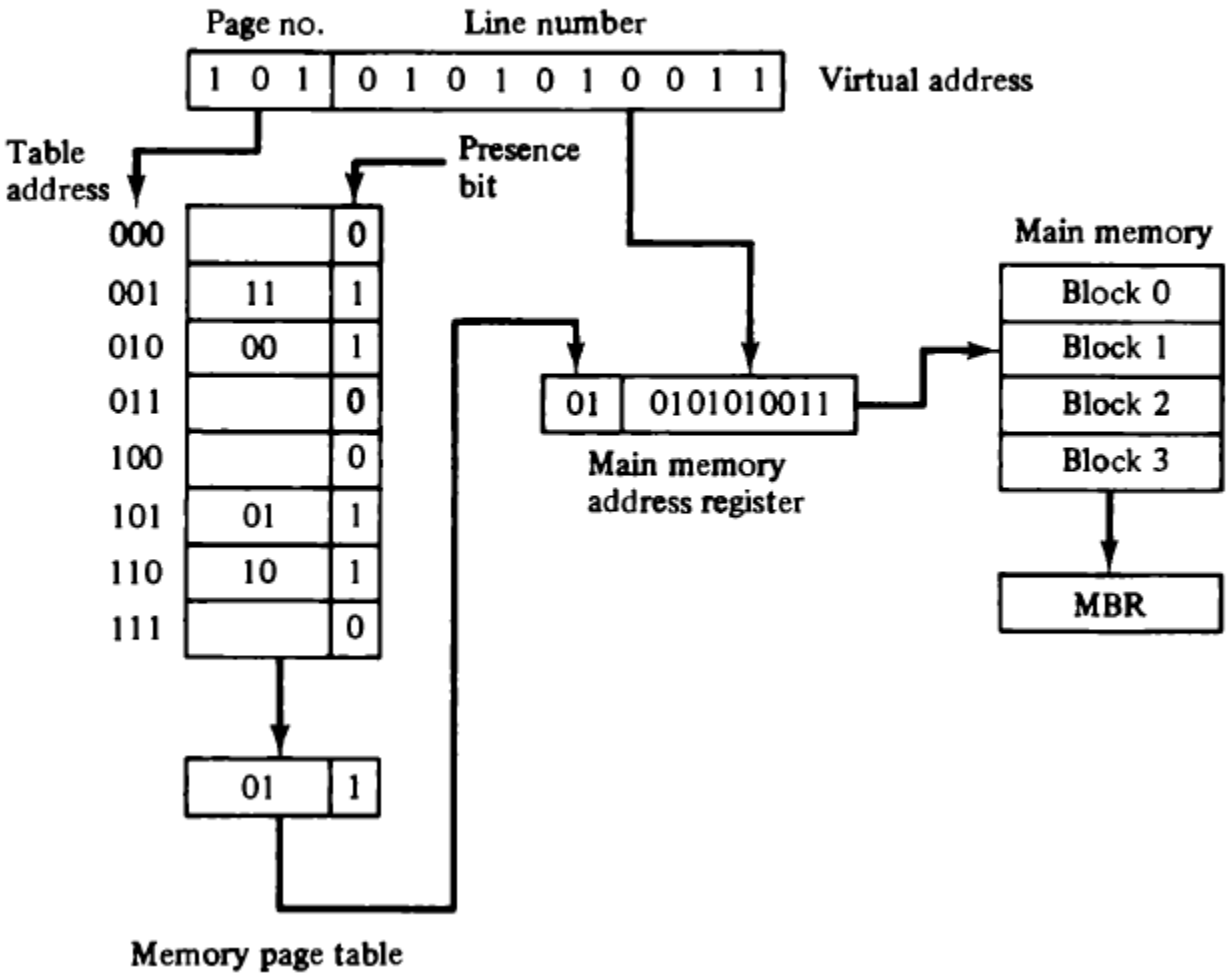


Figure 12-18 Address space and memory space split into gro

the memory page table at the page number add

Figure 12-19 Memory table in a paged system.



Page Replacement

A virtual memory organization is a consolidation of hardware and software systems. It can make efficient utilization of memory space all the software operations are handled by the memory management software.

The hardware mapping system and the memory management software together form the structure of virtual memory.

When the program implementation starts, one or more pages are transferred into the main memory and the page table is set to denote their location. The program is implemented from the main memory just before a reference is created for a page that is not in memory. This event is defined as a page fault

When a page fault appears, the program that is directly in execution is stopped just before the required page is transferred into the main memory. Because the act of loading a page from auxiliary memory to main memory is an I/O operation, the operating framework creates this function for the I/O processor.

In this interval, control is moved to the next program in the main memory that is waiting to be prepared in the CPU. Soon after the memory block is assigned and then moved, the suspended program can resume execution.

If the main memory is full, a new page cannot be moved in. Therefore, it is important to remove a page from a memory block to hold the new page. The decision of removing specific pages from memory is determined by the replacement algorithm.

There are two common replacement algorithms used are the first-in, first-out (FIFO) and least recently used (LRU). The FIFO algorithm chooses to replace the page that has been in memory for the highest time. Every time a page is weighted into memory, its identification number is pushed into a FIFO stack.

FIFO will be complete whenever memory has no more null blocks. When a new page should be loaded, the page least currently transports in is removed. The page to be removed is simply determined because its identification number is at the high of the FIFO stack.

The FIFO replacement policy has the benefit of being simple to execute. It has the drawback that under specific circumstances pages are removed and loaded from memory too frequently.

The LRU policy is more complex to execute but has been more interesting on the presumption that the least recently used page is an excellent applicant for removal than the least recently loaded page as in FIFO. The LRU algorithm can be executed by relating a counter with each page that is in the main memory.

When a page is referenced, its associated counter is set to zero. At permanent intervals of time, the counters related to all pages directly in memory are incremented by 1.

The least recently used page is the page with the largest count. The counters are known as aging registers, as their count denotes their age, that is, how long ago their related pages have been referenced.

MEMORY ORGANISATION

Outline

Memory Organization:
Memory Hierarchy,
Main Memory- RAM and ROM chips, Memory Address Map,
Auxiliary memory- Magnetic Disks, Magnetic Tapes,
Associative Memory – Hardware Organization,
Cache Memory – Associative Mapping, Direct Mapping, Set Associative Mapping,
Virtual Memory – Address Space and Memory Space, Address Mapping using pages, A

Memory Hierarchy

Section - 1

General memory Hierarchy

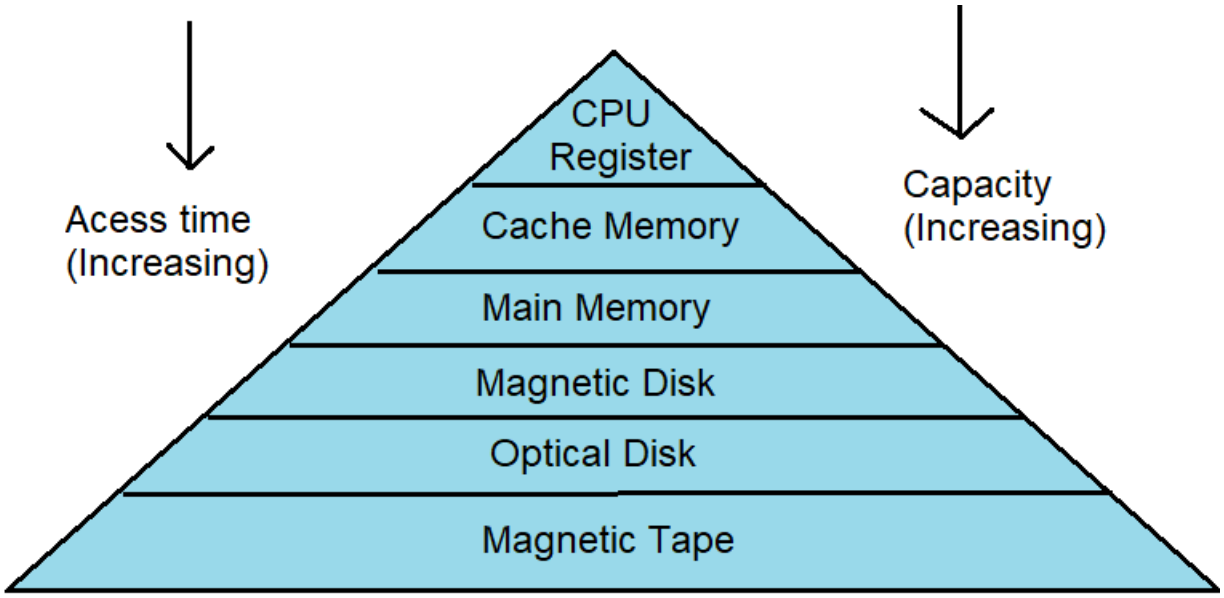
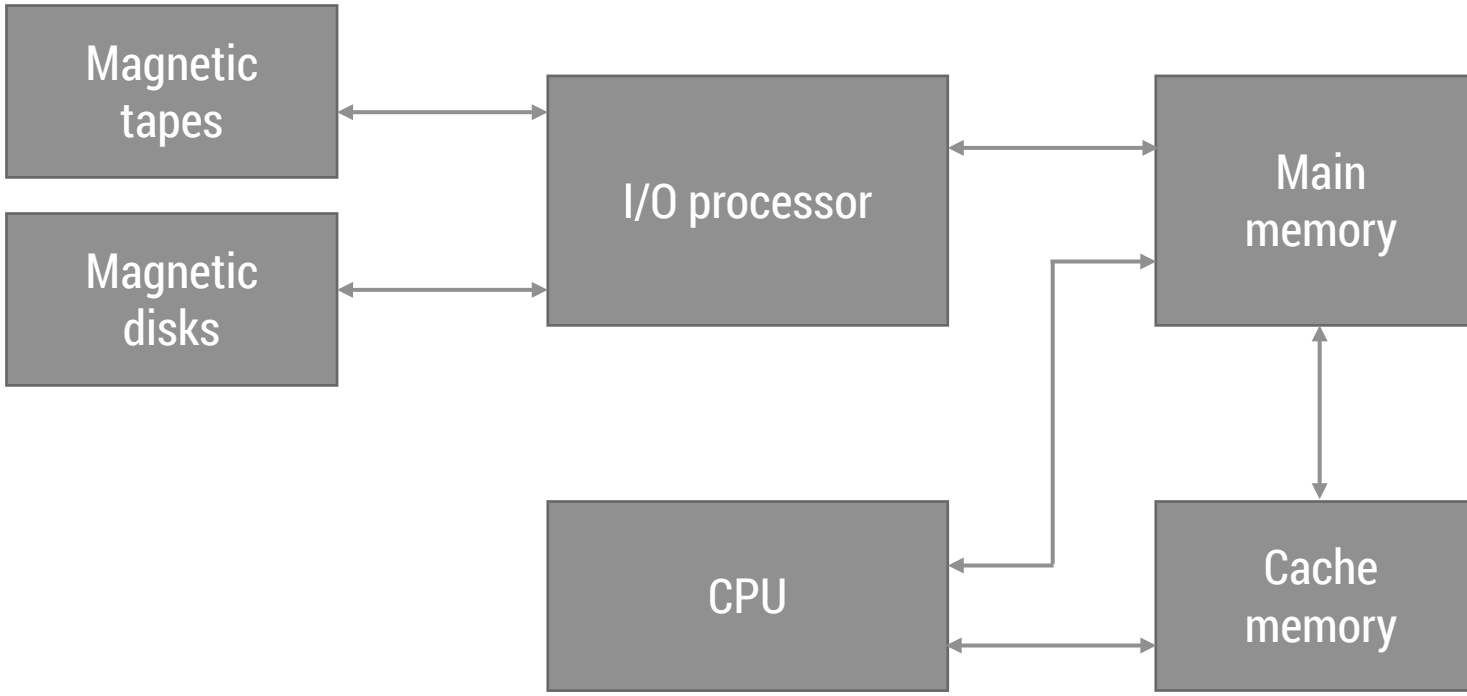


Fig:- Memory Hierarchy

Memory Hierarchy



The overall goal of using memory hierarchy is to obtain the highest possible average acces
minimising the total cost of the entire memory system

Memory Hierarchy

- The memory unit that communicates directly with the CPU is called main
- Devices that provide backup storage are called Auxilary memory .The auxilary memory devices used in computer system are magnetic disks an
- They are used for storing system programs,large data files and other back
- Only the data and programs currently needed by the processor reside in th
- All other information is stored in the auxilary memory and transferred needed
- A special very high speed memory called Cache is sometimes used to processing
- The cache is used for storing segments of programs currently being ex temporary data frequently needed in present calculations

Main Memory

Section - 2

Random access memory (RAM)

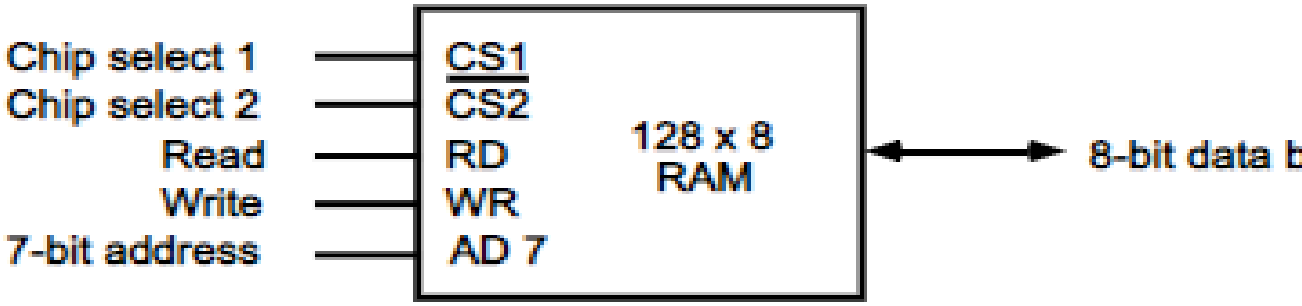
- Used in computers for the temporary storage of programs and data.
- Read and write both operations are performed by RAM which requires fa
  slow down the computer operation.
- It is volatile and lose all stored information if power is interrupted or turne
- It can be expanded by combining several memory chips.

Read-Only Memory (ROM)

- A read-only memory (ROM) is essentially a memory device in wh information is stored.
- A ROM which can be programmed is called a PROM. The process of en ROM is known as programming.
- ROMs are used to store information which is of fixed type, such as table fixed data and instructions.
- ROMs can be used for designing combinational logic circuits.

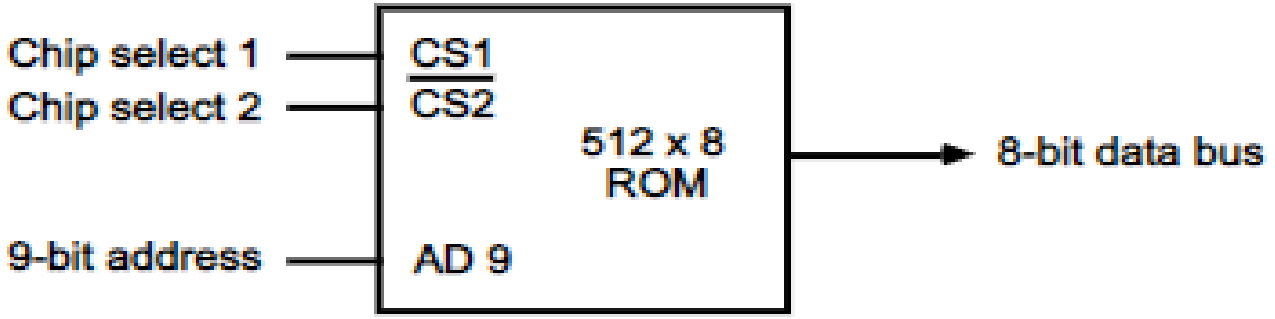
RAM and ROM Chips

Typical RAM chip



CS1	$\overline{CS2}$	RD	WR	Memory function	State of data bus
0	0	x	x	Inhibit	High-impedence
0	1	x	x	Inhibit	High-impedence
1	0	0	0	Inhibit	High-impedence
1	0	0	1	Write	Input data to RAM
1	0	1	x	Read	Output data from
1	1	x	x	Inhibit	High-impedence

Typical ROM chip



Memory Address Map

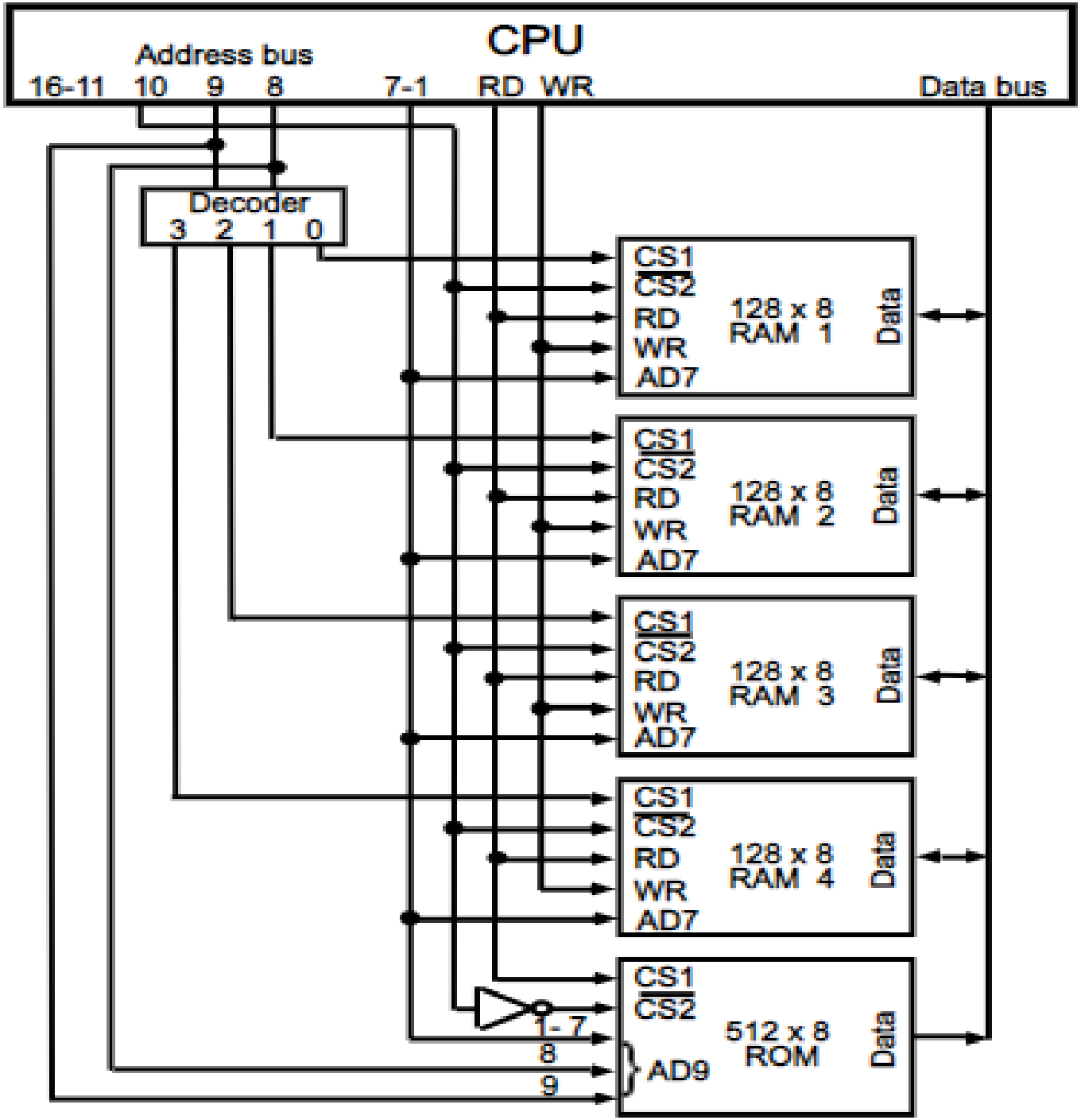
- The addressing of memory can establish by means of a table that specifie assigned to each chip.
- The table, called a memory address map, is a pictorial representation of as for each chip in the system, shown in the table.♦
- To demonstrate with a particular example, assume that a computer syster RAM and 512 bytes of ROM.

TABLE 1 Memory Address Map for Microprocomputer

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

Memory Connection to CPU

- RAM and ROM Chips are connected to a CPU through the data and addres

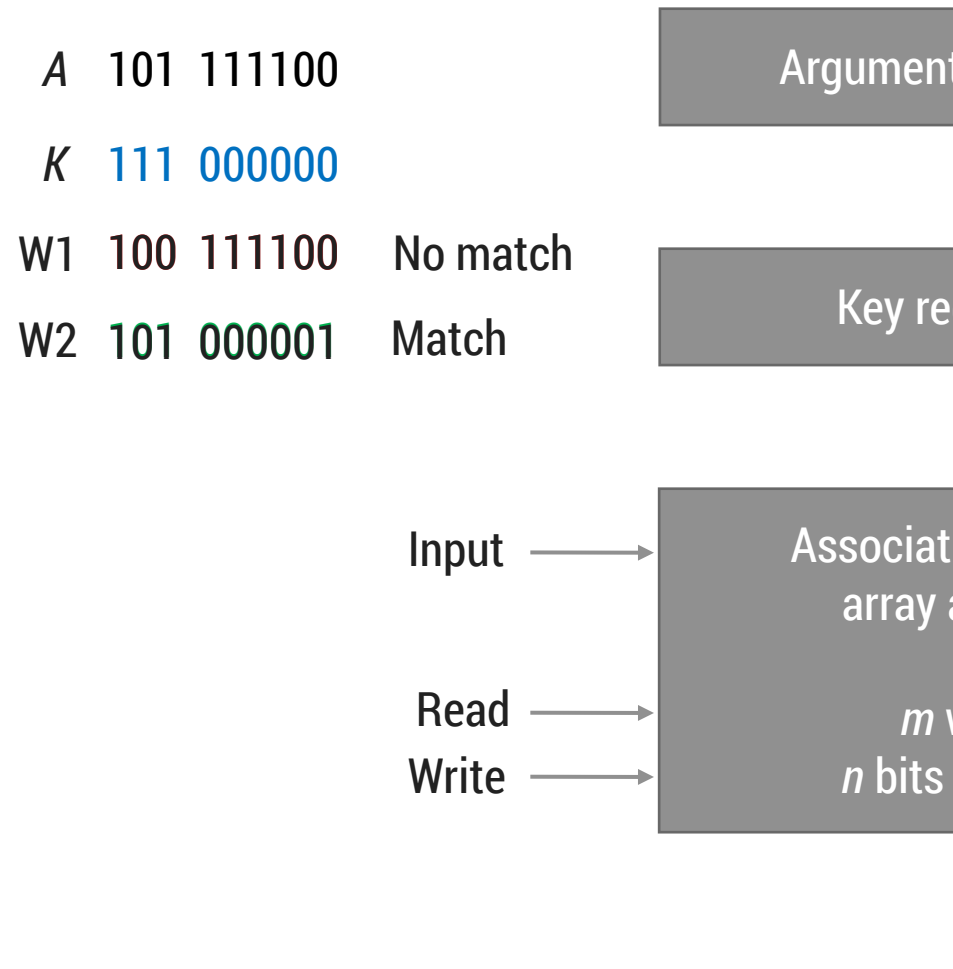

CPU
Address bus
16-11
10
9
8
7-1
RD
WR
Data bus
Decoder
3
2
1
0
CS1
CS2
RD
WR
AD7
128 x 8
RAM 1
Data
CS1
CS2
RD
WR
AD7
128 x 8
RAM 2
Data
CS1
CS2
RD
WR
AD7
128 x 8
RAM 3
Data
CS1
CS2
RD
WR
AD7
128 x 8
RAM 4
Data
CS1
CS2
1 - 7
8
9
AD9
512 x 8
ROM
Data

Associative Memory

Section - 4

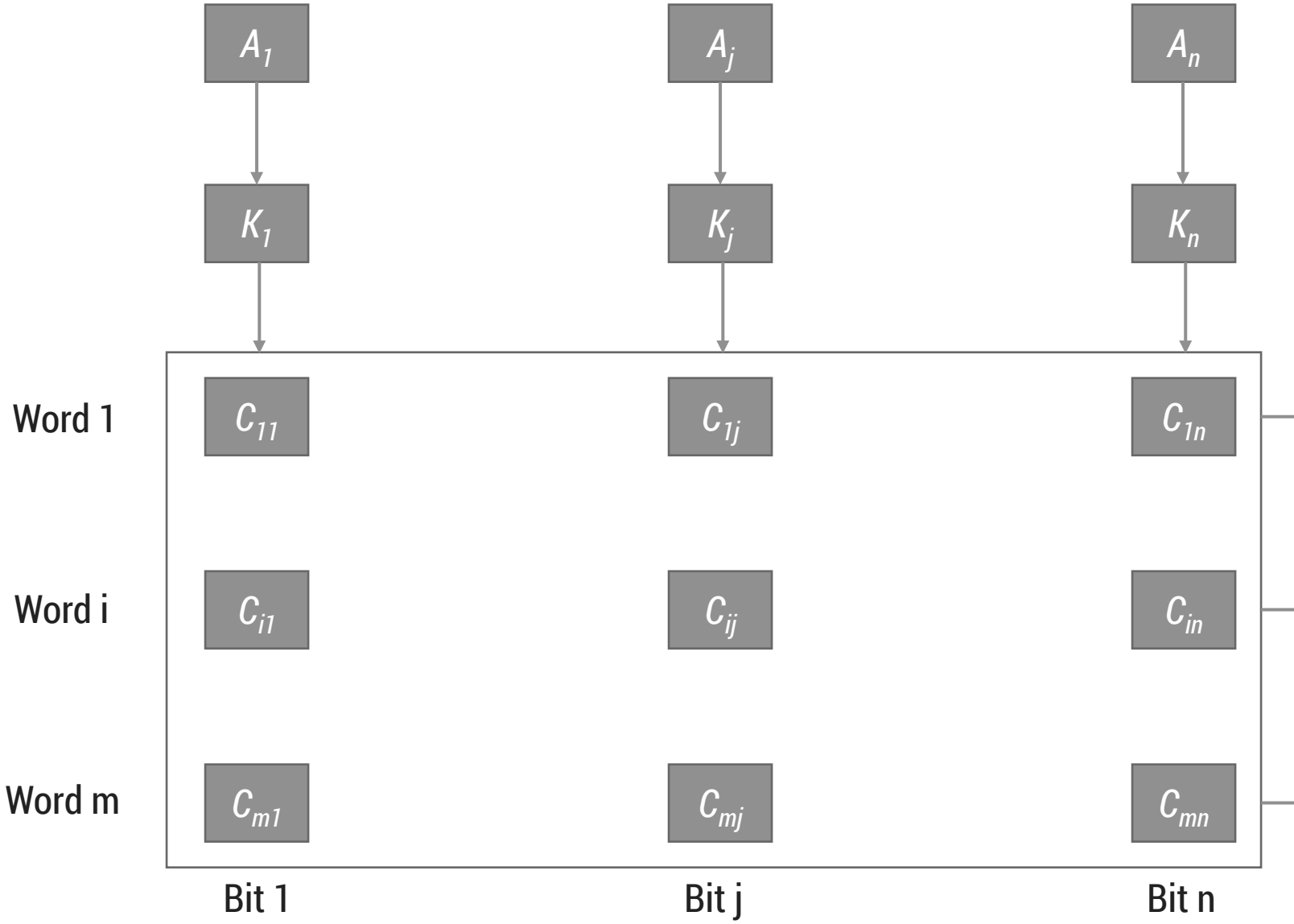
Associative Memory(Content Addressable Memory)

- 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 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 to be read from an associative memory, then the content of the word or part of word is specified



- Argument Register: This is the register that which stores Actual conten search in the available associative memory.
- Key Register: The key Register will provide a mask for choosing a particu Argument register.
- The Argument bits, where corresponding 1's in the key register are o content for searching a required word from the Memory.

Associative Memory of m words n cells per word



- **The cells in the array** are marked by the letter C with two subcripts. The word number and the second specifies the bit position in the word.

- **Thus cell** C_{ij} is the cell for bit j in word i. A bit A_j in the argument register bits in column j of the array provided that $K_j = 1$.

- **This is done for all columns** j = 1, 2, . . . , n. If a match occurs between a the argument and the bits in word i, the corresponding bit M_i in the match

- **If one or more unmasked bits** of the argument and the word do not match

Cache Memory

Section - 5

Cache Memory

- Cache is a fast small capacity memory that should hold those information which are most likely to be accessed.
- **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 locality of reference.

- The basic operation of the cache is, when the 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.
- The transformation of data from main memory to cache memory is referred to as a mapping process.

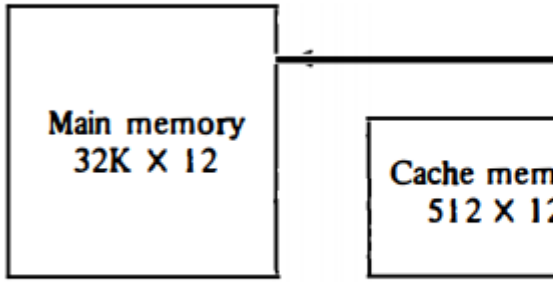


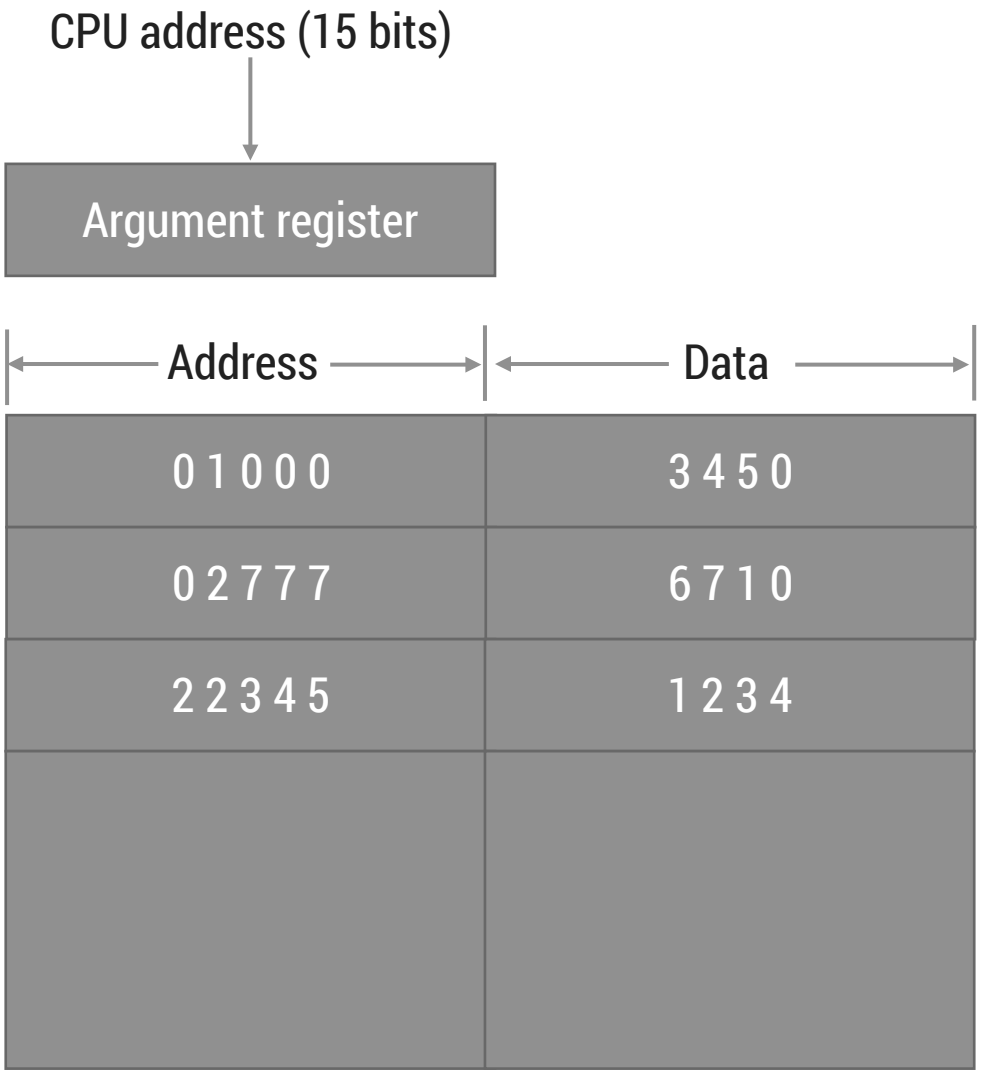
Figure 10 Example

The CPU communicates
first sends a 15-bit addre
a hit, the CPU accepts th
cache. If there is a miss,
from main memory and t
transferred to cache.

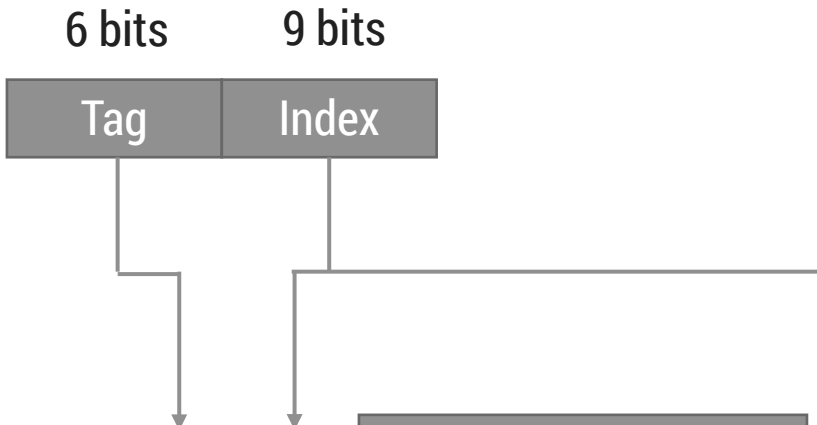
- The performance of the cache memory is frequently measured in terms *ratio*.
- When the CPU refers to memory and finds the word in cache, it is said to p
- If the word is not found in cache, it is in main memory and it counts as a *m*
- The ratio of the number of hits divided by the total CPU references to me is the *hit ratio.*
- Hit ratios of 0.9 and higher have been reported.

Associative Mapping

- The diagram shows three words presently stored in cache
- The address value of 15 bits is shown as 5 digit octal number and corresponding 12-digit word is shown as 4 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.



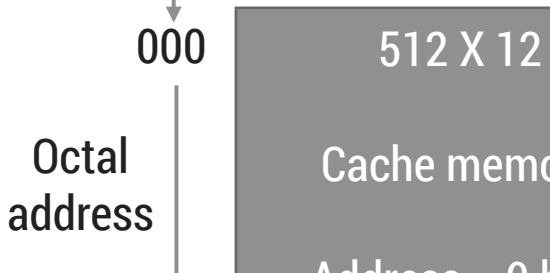
Direct Mapping



In the general case, there are 2
words in main memory.

The n-bit memory address is di
the index field and n - k bits for

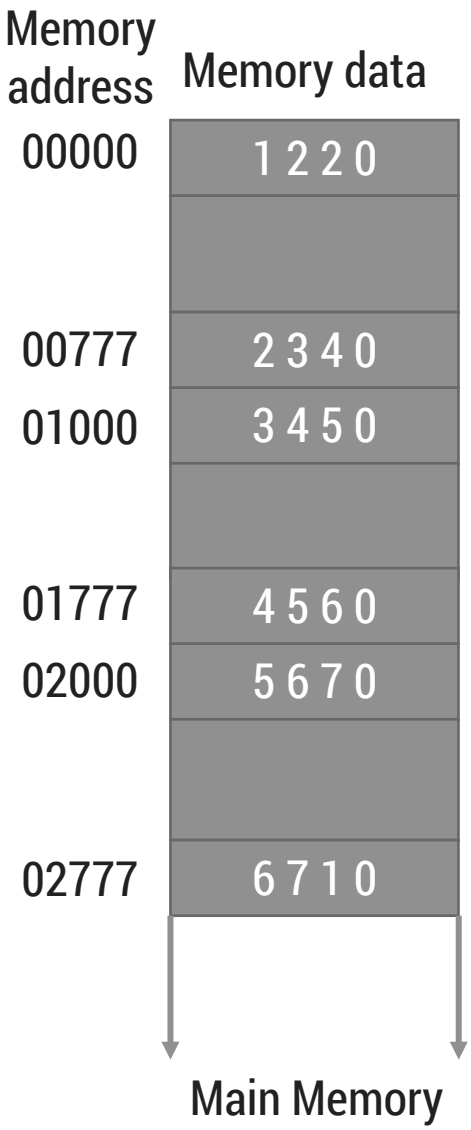
The direct mapping cache orga
access the main memory and th
32k =2 power 5*2 power10=

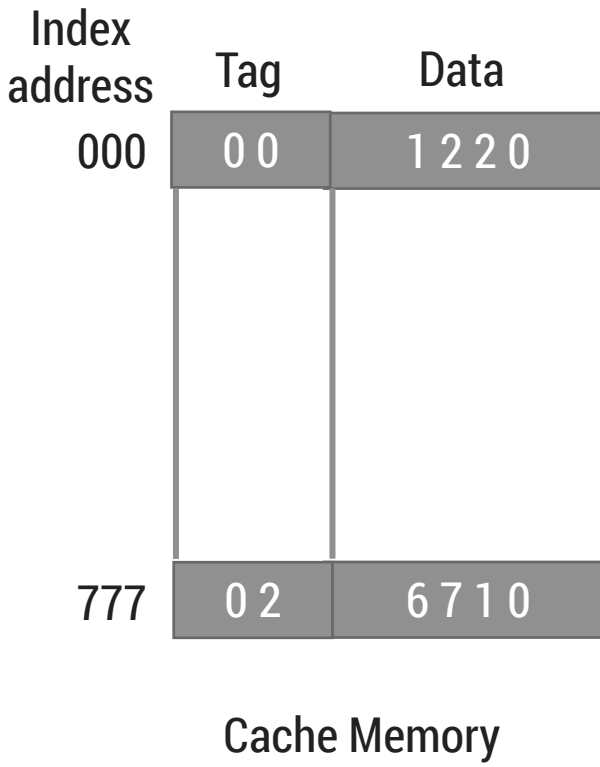


Direct Mapping

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.

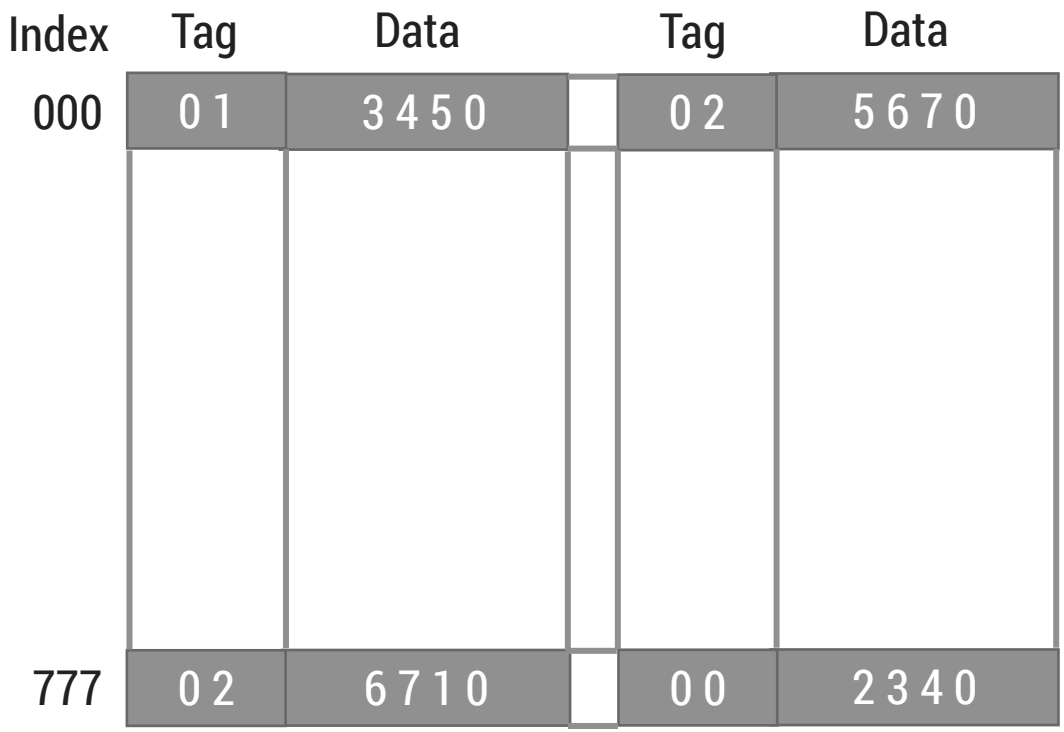




The disadvantage of direct mapping is that the hit ratio can drop considerably if two or more words whose addresses have the same index but differenttags are accessed repeatedly.

Set-Associative Mapping

It can store two or more words 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. An index address of nine bits can accommodate 512 words. Thus the size of cache memory is 512 x 36.



Virtual Memory

Section - 6

Virtual Memory

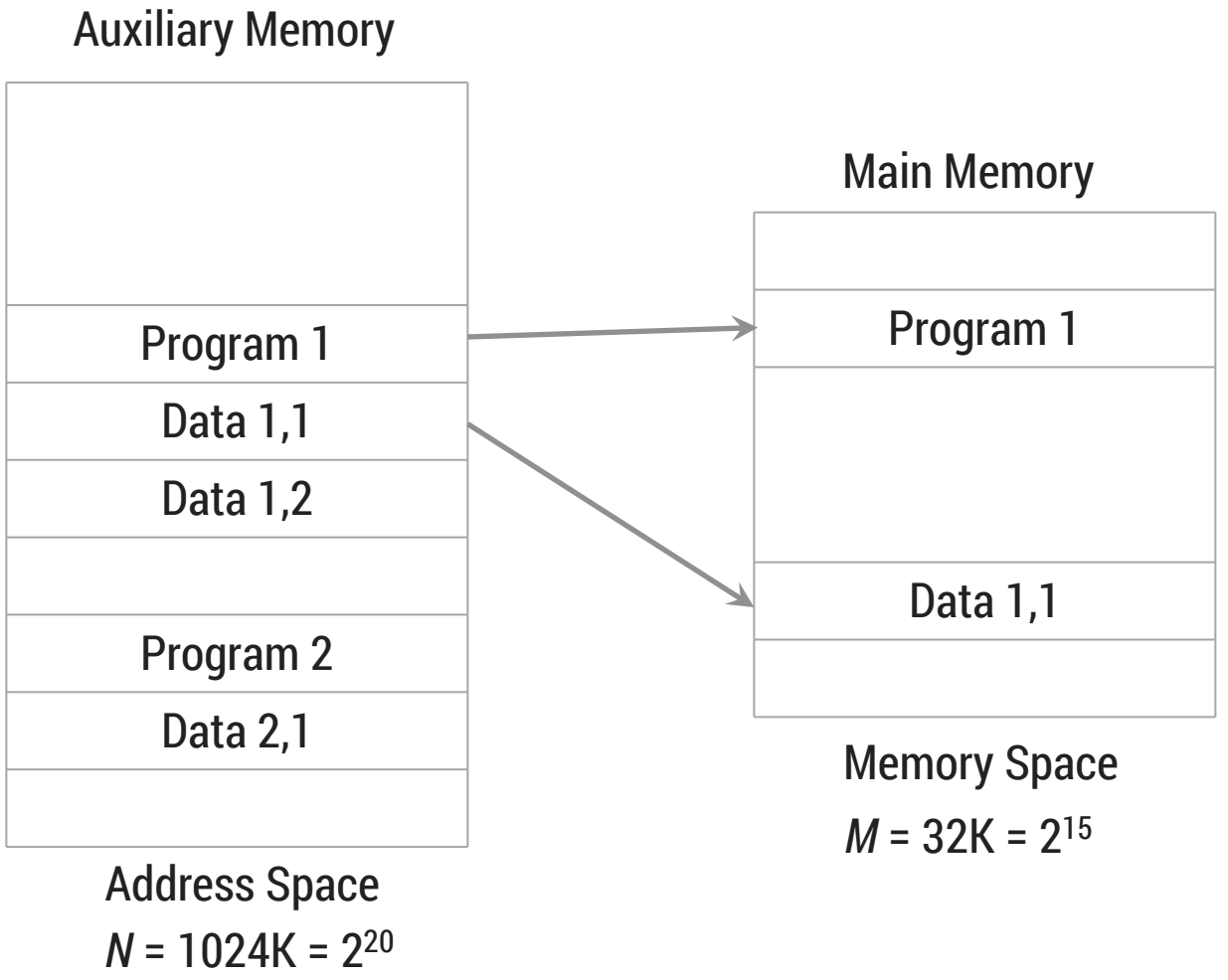
- Virtual memory is used to give programmers the illusion that they have their disposal, even though the computer actually has a relatively small m
- A virtual memory system provides a mechanism for translating progra into correct main memory locations.
- **Address space**

  An address used by a programmer will be called a virtual address, and th is known as address space.
- **Memory space**

  An address in main memory is called a location or physical address. The called the memory space.

Virtual Memory

- Relation between Address space & Memory space



Virtual Memory

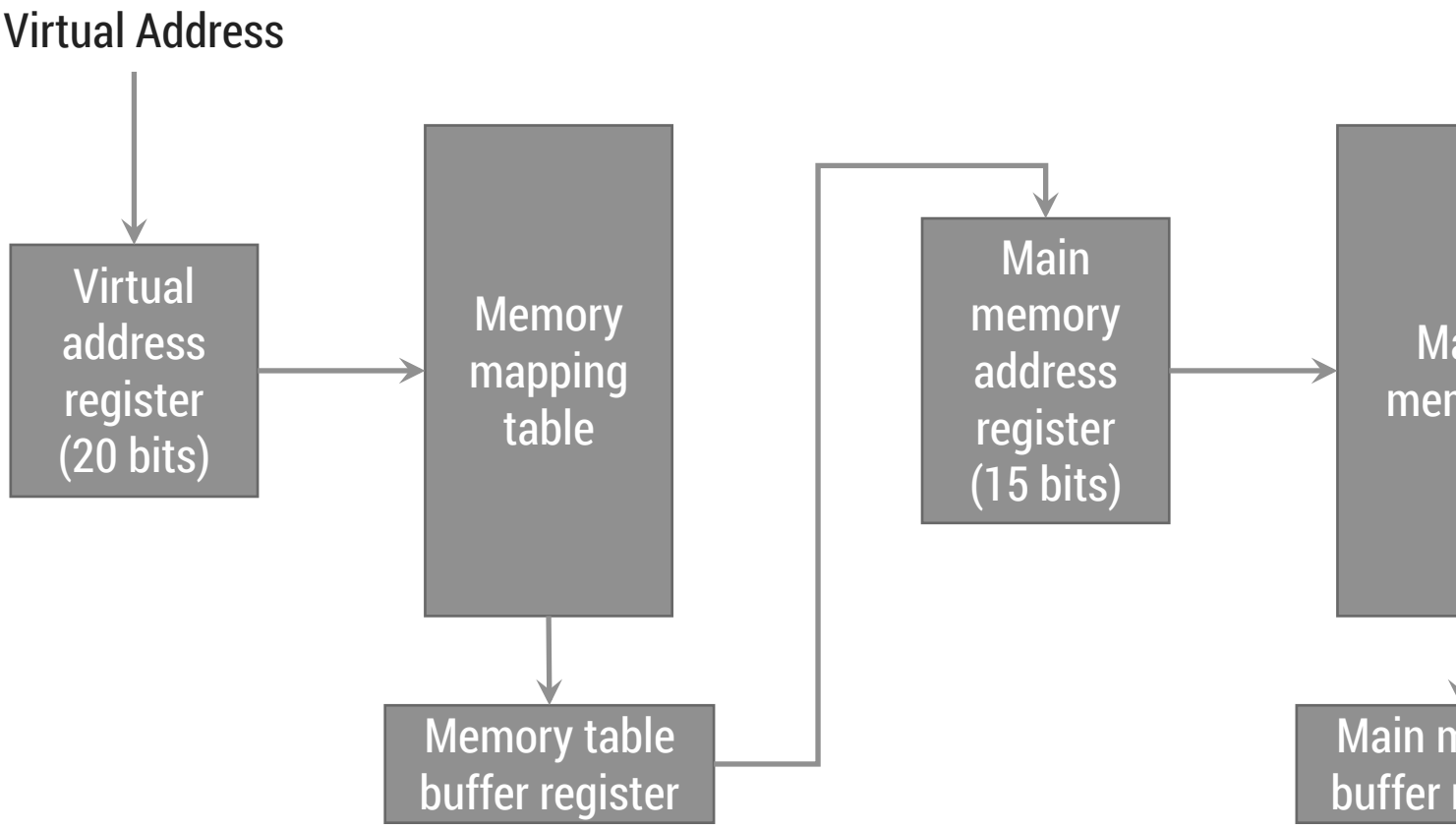
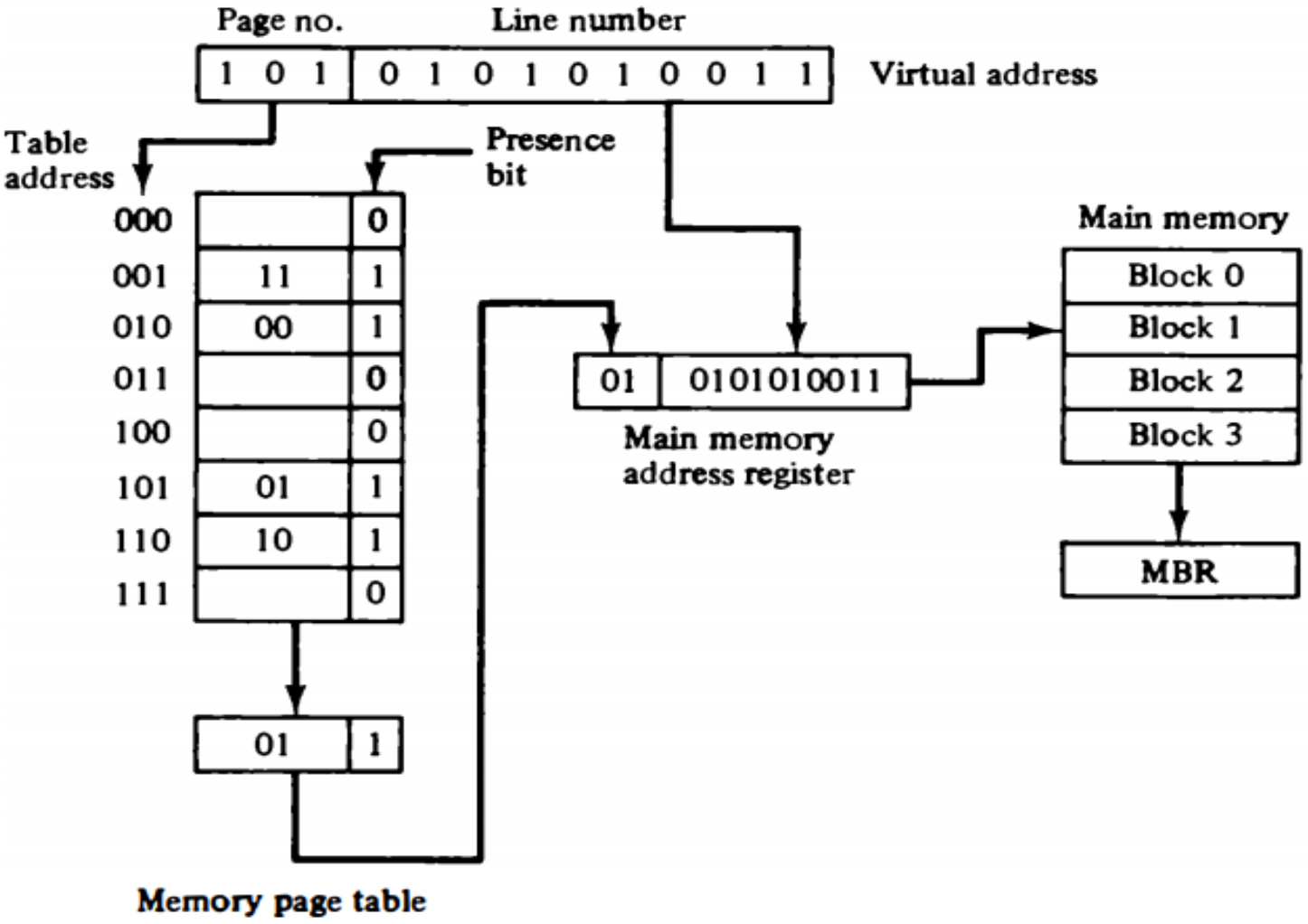


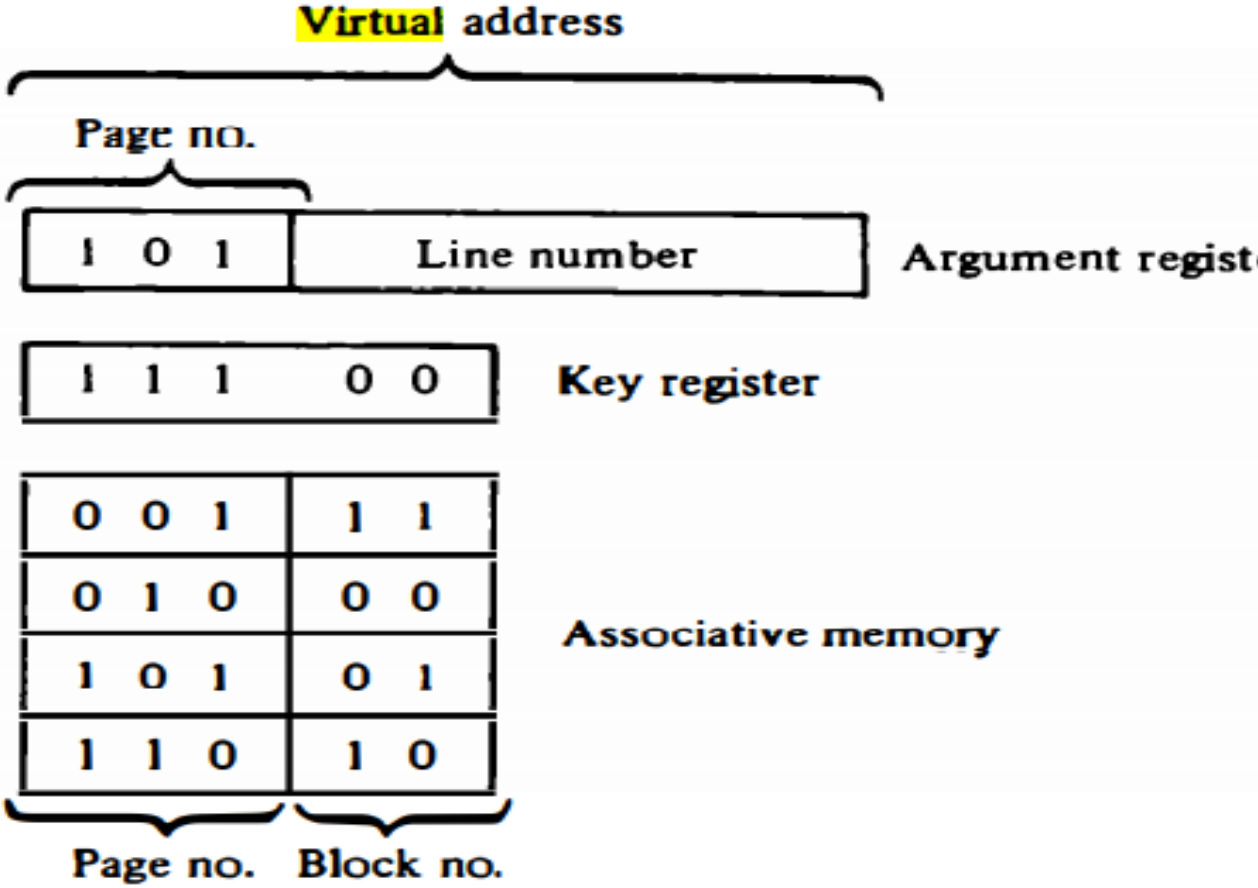
Figure 19 Memory table in a paged system.



Associative Memory Page Table

- A random-access memory page table is inefficient with respect to storage
- consider an address space of 1024K words and memory space of 32K block contains 1K words, the number of pages is 1024 and the number of of the memory-page table must be 1024 words and only 32 locations m equal to 1. At any given time, at least 992 locations will be empty and not
- A more efficient way to organize the page table would be to construct it equal to the number of blocks in main memory.

Figure 12-20 An associative memory page table.



Page Replacement

- The program is executed from main memory until it attempts to referen auxiliary memory. This condition is called page fault.
- When page fault occurs, the execution of the present program is suspe page is brought into main memory.
- The FIFO algorithm selects for replacement the page that has been in mer
- LRU
- 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2