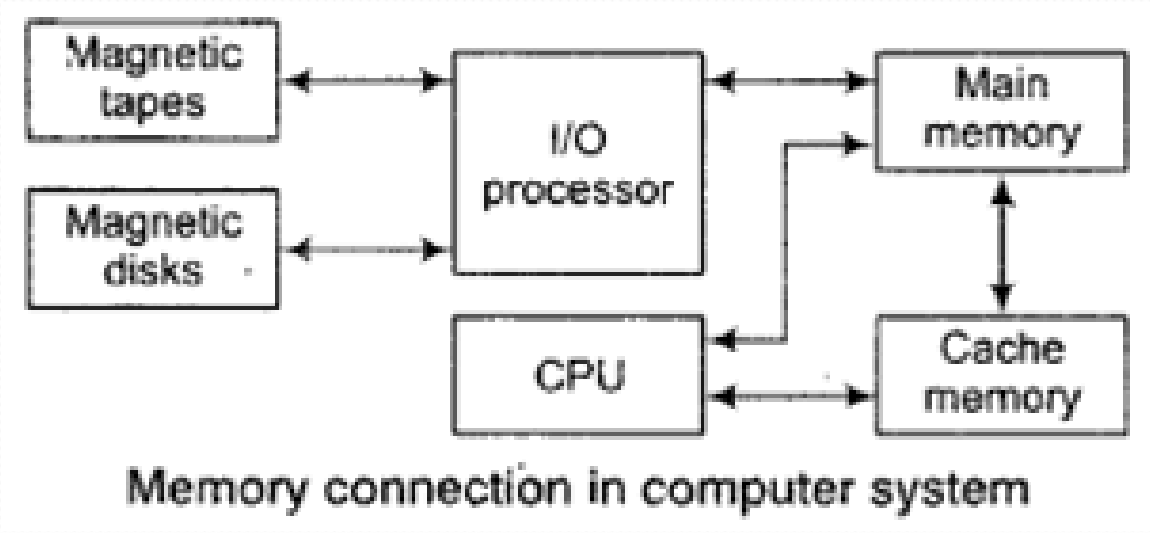
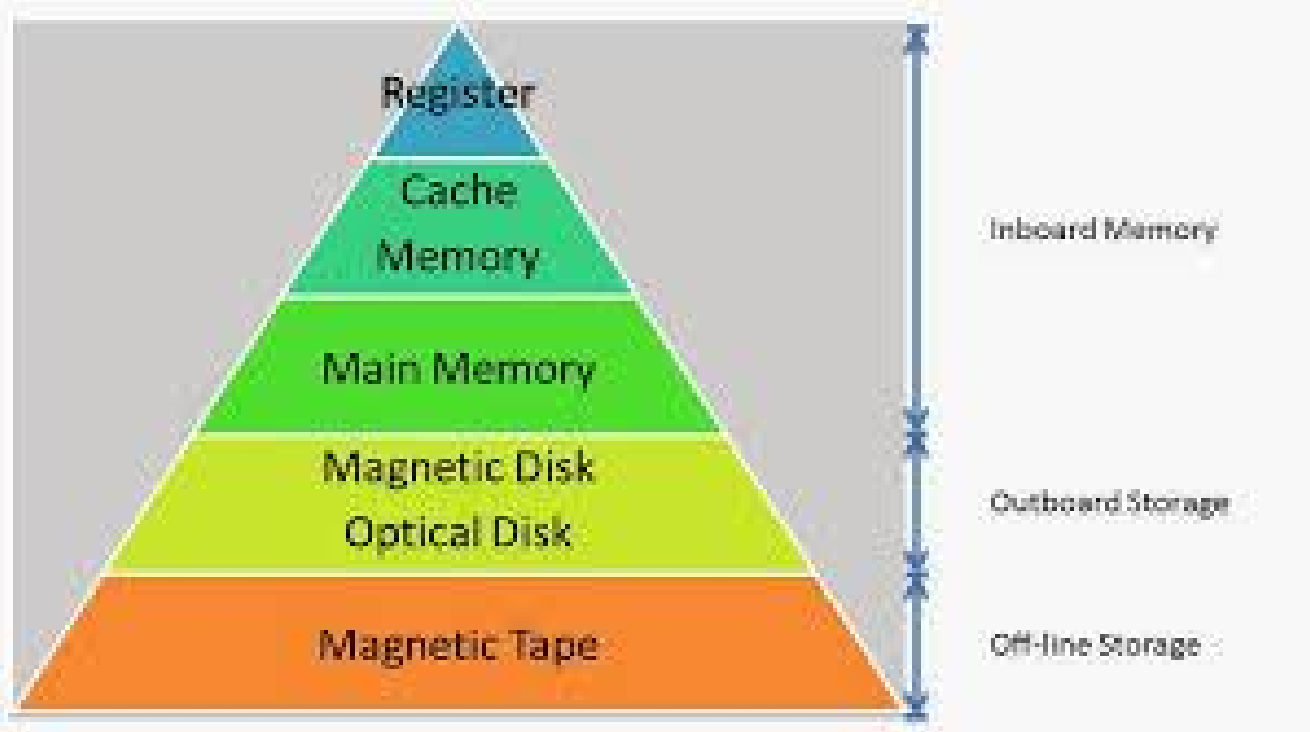
Memory Organization

Memory Hierarchy



Memory connection in computer system

- Memory hierarchy system consists of all storage devices employed in a computer system.



- As the storage capacity increases, cost/bit decreases and access time becomes longer.
- Goal of memory hierarchy: to obtain the highest possible average access speed while minimizing the total cost of entire memory.

- **Main memory:** communicates directly with the CPU
- **Auxiliary memory:** devices that provide backup storage (ex. Magnetic disks, magnetic tapes)
  - Stores programs, large data files and other backup info
- Only the programs and data currently needed by the processor are reside in main memory.
- **Cache Memory:** very high speed memory

Cache Memory

- CPU logic is usually faster than main memory access time.
- To compensate the mismatch in operating speed, a extremely fast, small cache is employed between CPU and main memory.
- Access time of cache is close to processor logic cycle time.
- Small, fast and expensive memory.

- Cache Memory:
  - Holds parts of program that are most heavily used.
  - CPU has direct access to cache and main memory.
  - Access time ratio between cache and main memory is about 1 to 7.

- Auxiliary Memory:
  - Holds parts of program and data that are not presently being used.
  - CPU does not have direct access to auxiliary memory.
  - Average access time is 1000 times that of main memory.
  - Transfer from auxiliary to main memory in blocks of data.

Main Memory

- Relatively large and fast memory used to store programs and data during the computer operation.
- Random access memory(RAM), also referred to as read write memory.
- Volatile: stores info as long as power is applied to the unit.
- IC RAM chips: Static and Dynamic

Static RAM

- Internal flip-flops store the binary info.
- Info valid as long as power is applied to the unit.
- Easier to use and has shorter read and write cycles.
- Application: Cache memory

Dynamic RAM

- Stores binary information in the form of electric charges that are applied to the capacitors.
- Reduced power consumption and larger storage capacity in a single memory chip.
- Used for implementing main memory.

Read Only Memory(ROM)

- A portion of the memory may be constructed with ROM chips.
- Non volatile.
- Used for storing programs that are permanently resident in the computer or information that do not change.
- Ex. To store bootstrap loader.
- When power is turned on, the hardware of the computer sets the PC to the first address of the bootstrap loader.
- This program loads the portion of the OS, which prepares the computer for general use.

Types of ROM

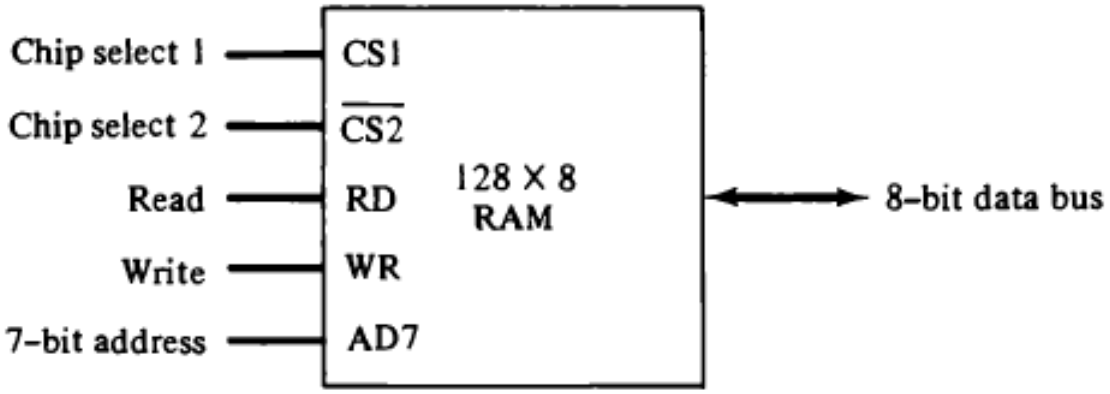
- ROM
- PROM(Programmable ROM)

} Irreversible

- EPROM(Erasable PROM)
- EEPROM(Electrically Erasable PROM)
- Ex. Flash memory is a form of EEPROM

RAM Chip

Figure 12-2 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

ROM Chip

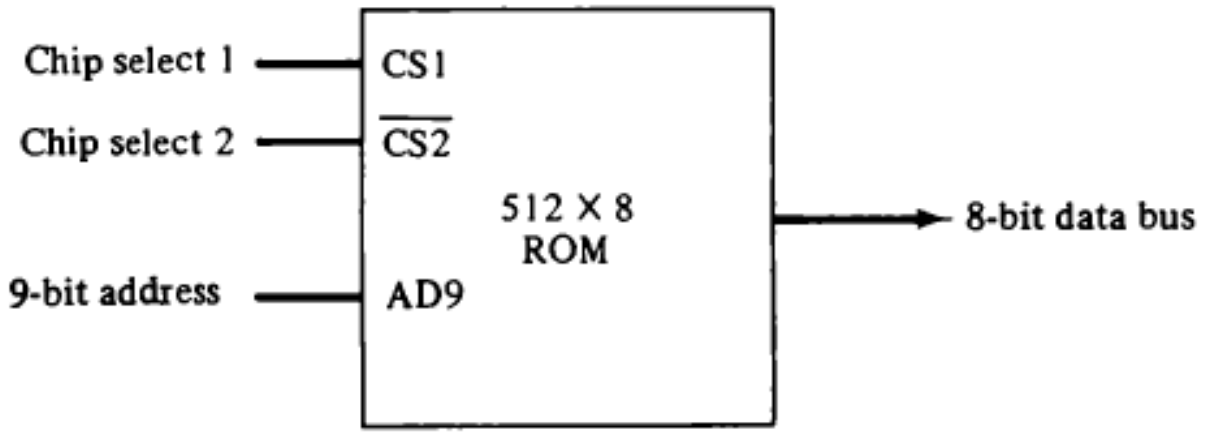
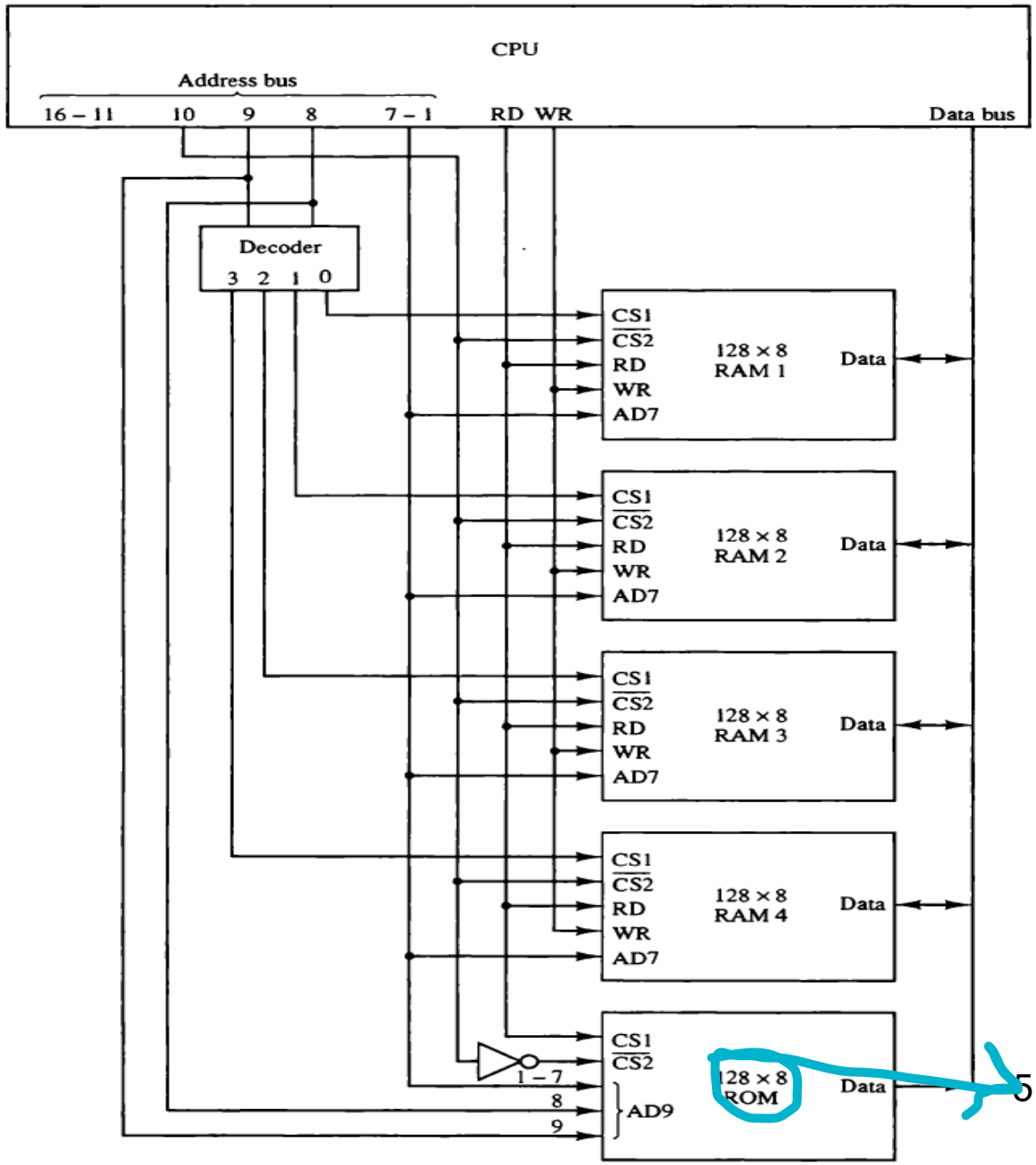


Figure 12-3 Typical ROM chip.

Memory Connection to CPU

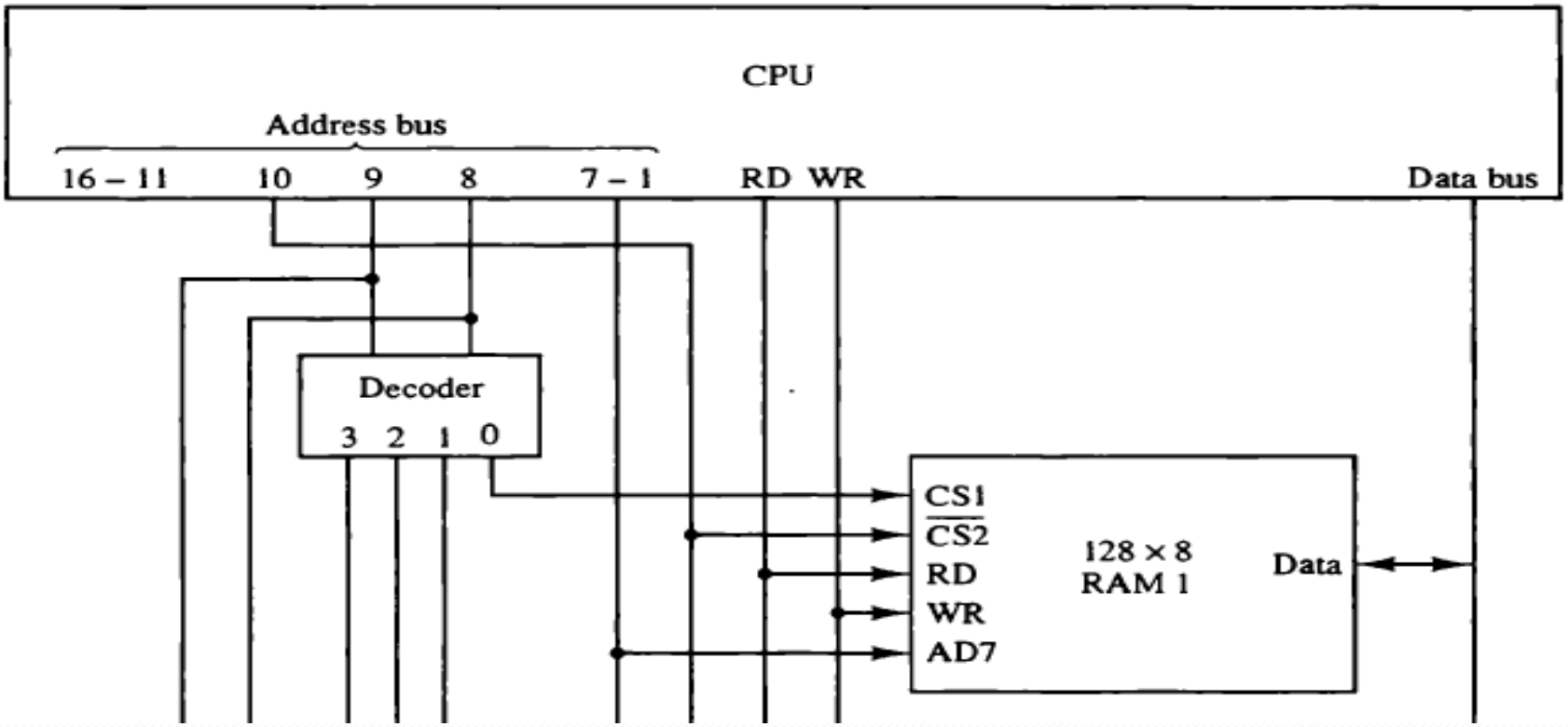
- RAM and ROM chips are connected to the CPU through the data and address buses.
- The low-order lines in the address bus select the byte within the chips.
- Other lines select a particular chip through its chip select inputs.



Total RAM:
512 x 8

512 x 8 ROM

Figure 12-4 Memory connection to the CPU.



- Bus line 10: Selects RAM or ROM
- If 0 – RAM ,1- ROM
- RAM(128 bytes): needs address lines (1-7)
- ROM(512 bytes) : needs address lines 1-9

Associative Memory

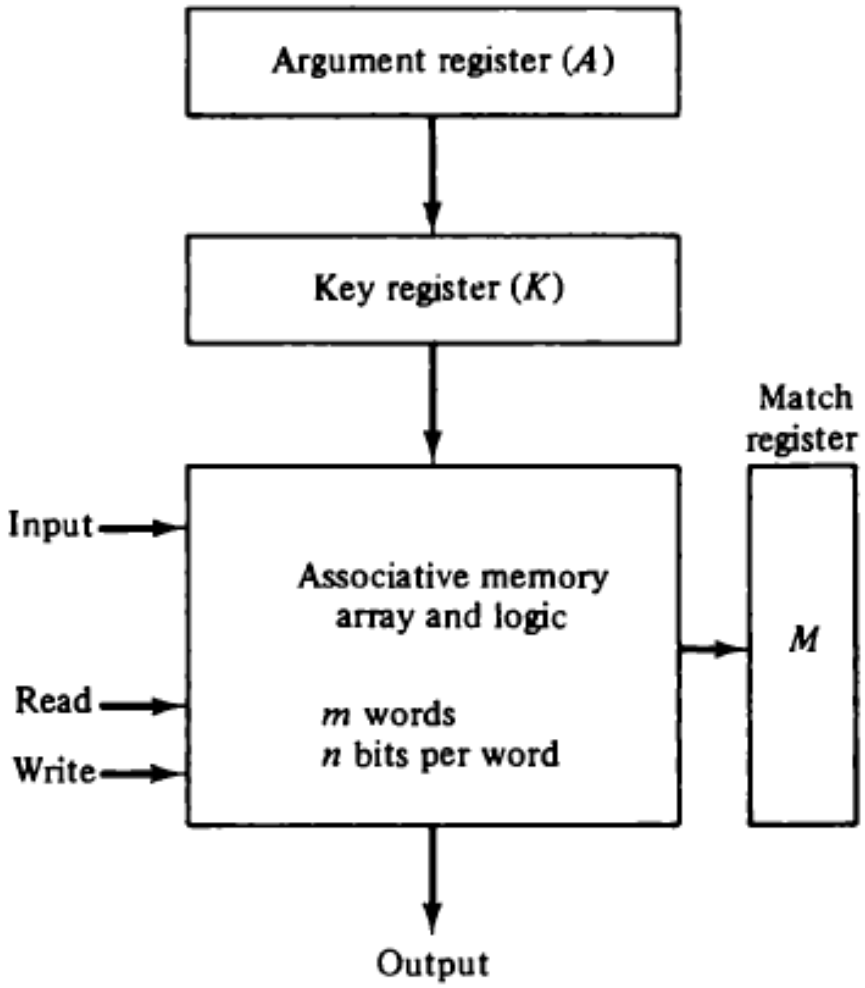
- Many data processing applications require the search of items in a table stored in memory.
- Go to address-read contents-compare with the item being searched- continue until match is found

Associative Memory

- Content Addressable Memory(CAM)
- Reduces time required to search an item stored in the memory by looking by contents rather than by address.
- It is accessed simultaneously and in parallel on the basis of data contents 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 location to store the word.
- Searches can be done on an entire word or on a specific field within a word.
- Key register provides a mask for choosing a particular field.

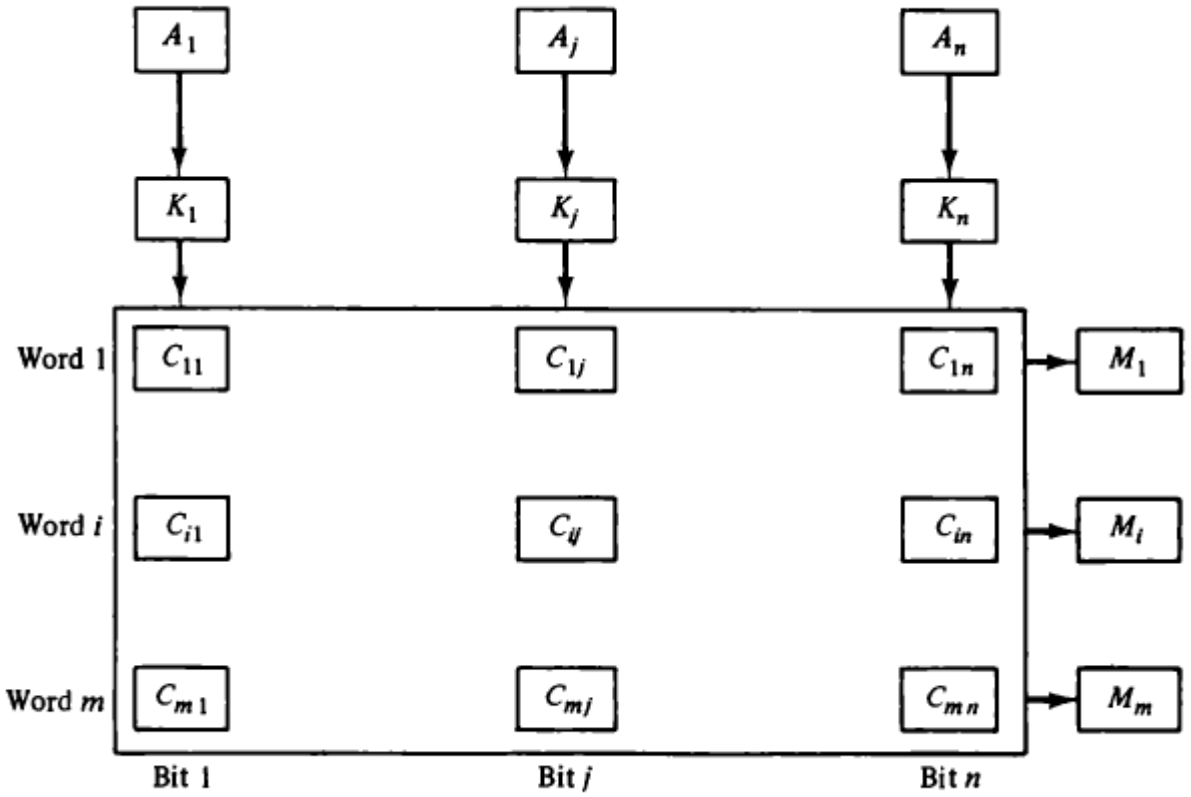
Figure 12-6 Block diagram of associative memory.



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

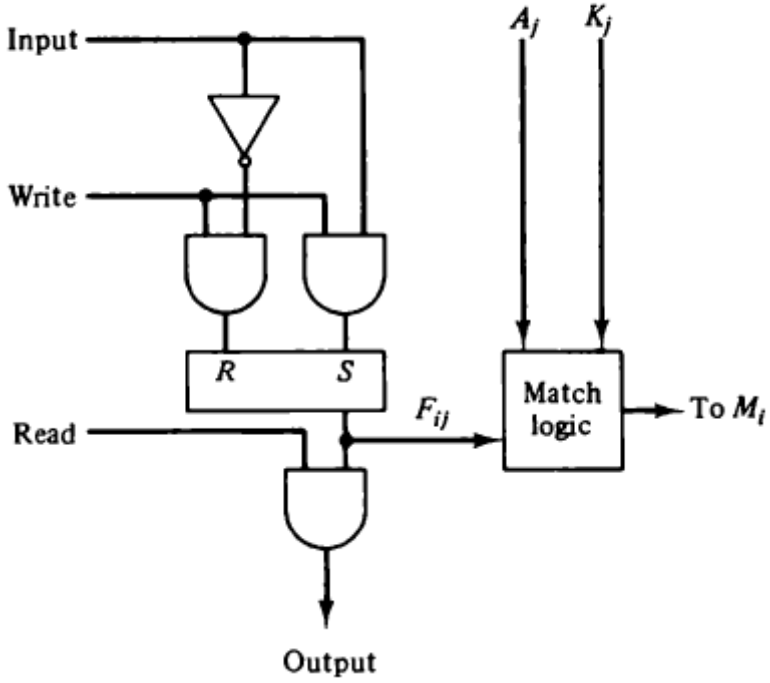
- Argument register and key register have n-bits.
- Match register has m-bits.

Figure 12-7 Associative memory of m word, n cells per word.



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

Figure 12-8 One cell of associative memory.



- More expensive than RAM.
- Used in applications where the search time is very critical and must be very short.

Cache Memory

- **Locality of reference:** references to memory at any given interval of time tend to be confined within a few localized areas in memory.
- Keep the most frequently accessed instructions and data in cache memory.
- Small in size, still helpful because of locality of reference property of programs.

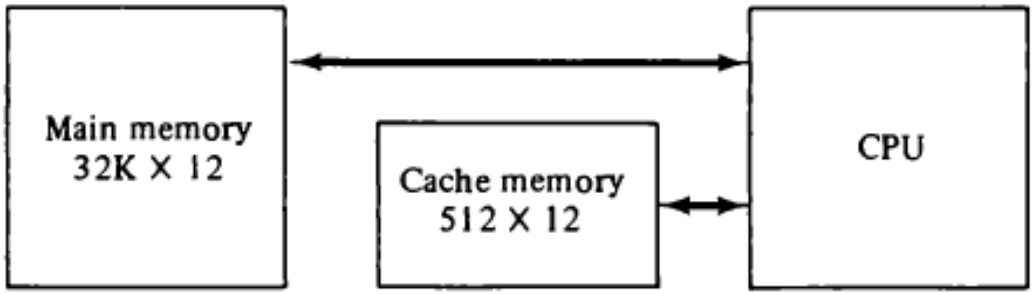


Figure 12-10 Example of cache memory.

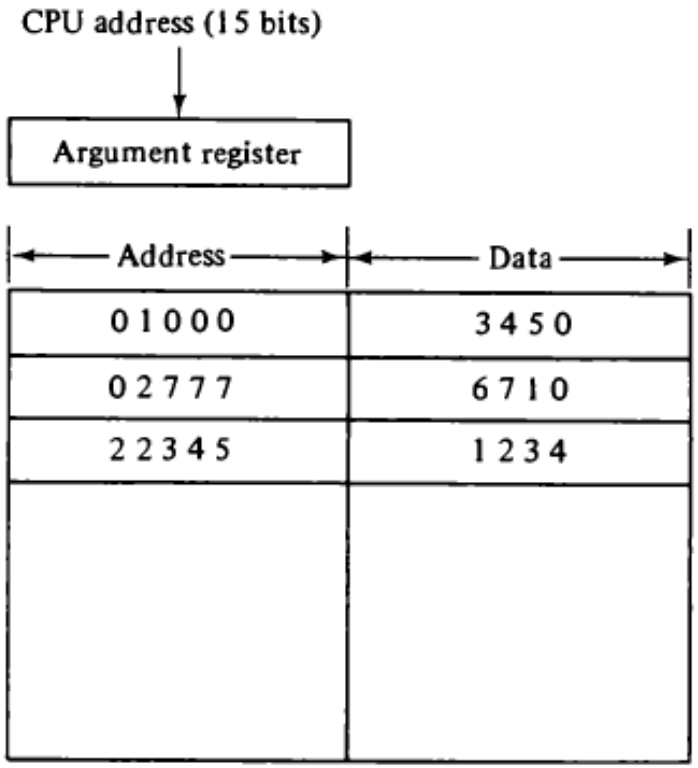
- For every word stored in cache, there is a duplicate copy in main memory.

- **Cache Hit:** When the CPU refers to memory and finds the word in cache.
- **Cache Miss**
- **Hit Ratio:** Ratio of number of hits divided by the total CPU references to memory (hit + miss)
- **Miss Penalty:** Additional cycle required to serve the miss.

Mapping

- Basic characteristic of cache is fast access time.
- Very little time must be wasted when searching for words in cache.
- Mapping function determines which location the block will occupy in cache:
  - Associative Mapping
  - Direct Mapping
  - Set-associative mapping

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



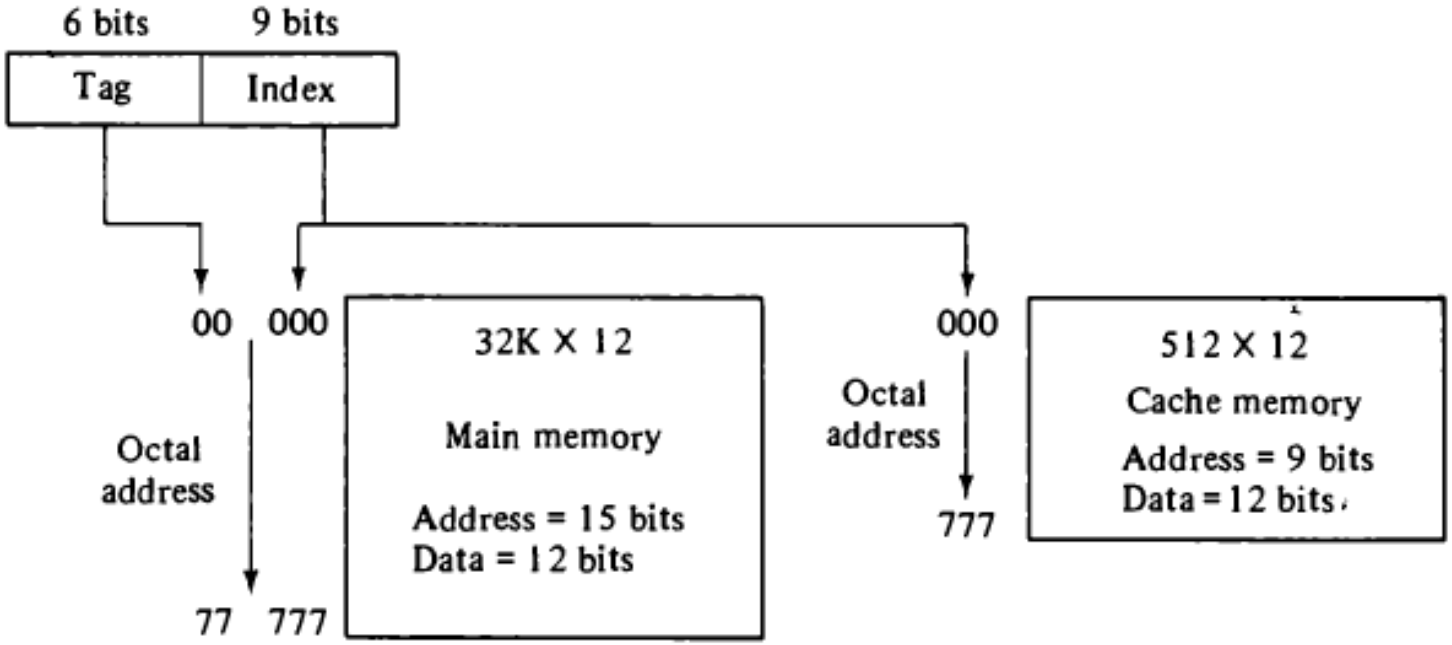
- Fastest and most flexible cache organization.
- Stores **both address and contents** of the memory word.
- Uses an associative memory (therefore expensive)

Direct Mapping

- Random access memory is used for cache.
- CPU address bits = tag field + index field
- Number of bits in index field = number of address bits of cache memory

Direct Mapping

Figure 12-12 Addressing relationships between main and cache memories.



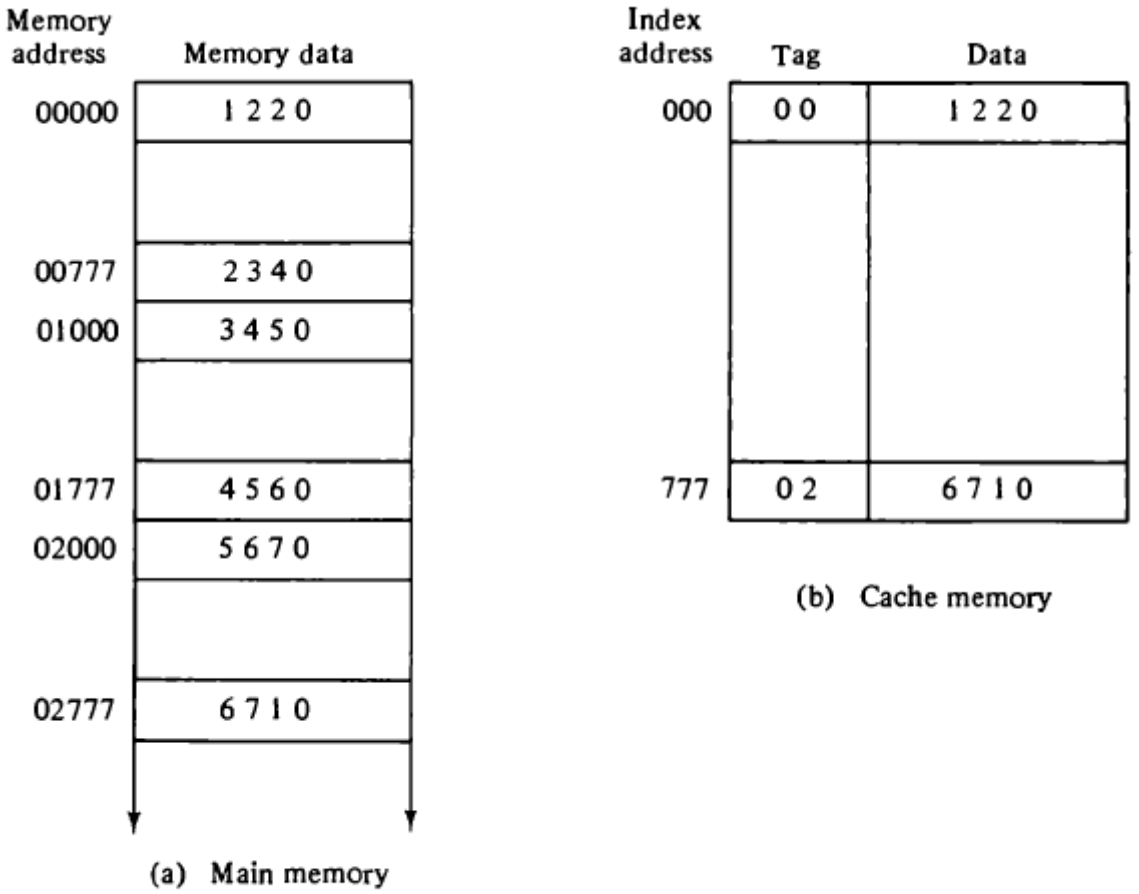


Figure 12-13 Direct mapping cache organization.

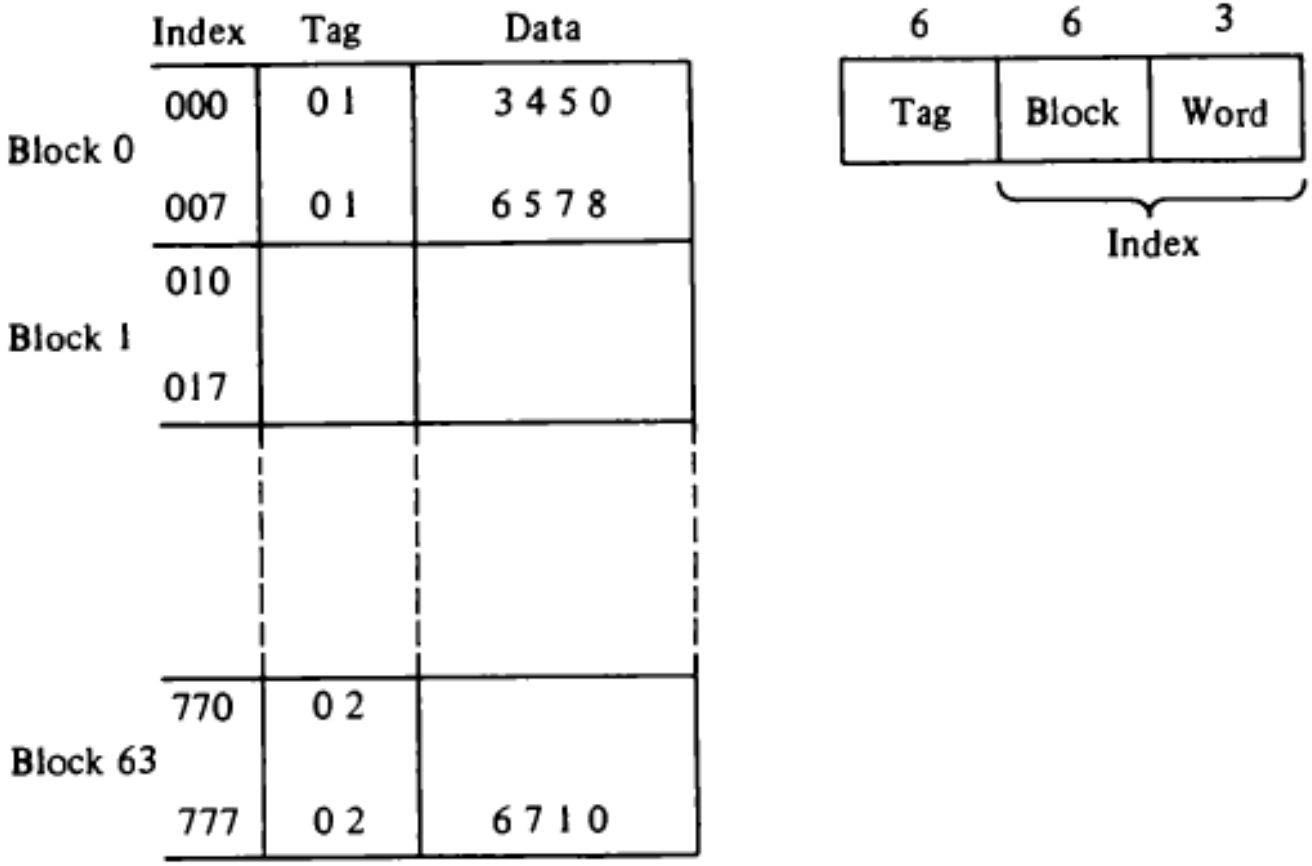


Figure 12-14 Direct mapping cache with block size of 8 words.

- Page replacement algorithm??
- Advantage: Simple and inexpensive to implement.

- Disadvantage : hit ratio drops considerably if two or more words whose addresses have the same index but different tags are accessed repeatedly.

Set-Associative Mapping

- Each word of cache can store two or more words of memory under the same index address.
- Each data word is stored together with tag.
- Set: number of tag-data items in a word form a set

Set-Associative Mapping

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 12-15 Two-way set-associative mapping cache.

Writing into cache

- When CPU finds a word in cache during a **read** operation, the main memory is **not** involved in the transfer.
- Write operation can be performed in two ways:
  - Write-through
  - Write-back

Replacement Algorithm

- Determines which block to be replaced if cache is full.
- Ex. FIFO, Least recently used (LRU)

Virtual Memory

- Virtual memory concept permits the user to construct programs larger than main memory size.
- Give programmers the illusion that they have very large memory at their disposal.
- Each address referenced by the CPU goes through an address mapping from virtual address to physical address.

Memory Management System

- Requirement of multiprogramming system:
  - Move programs and data around the memory
  - Vary the amount of memory in use by a given program
  - Prevent a program from changing other programs
- **MM System:** collection of hardware and software procedures for managing the various programs residing in memory
- MM software : part of overall operating system

Memory Management Unit

- Basic components of memory management unit:
  - A facility for dynamic storage relocation
  - 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 OS functions.