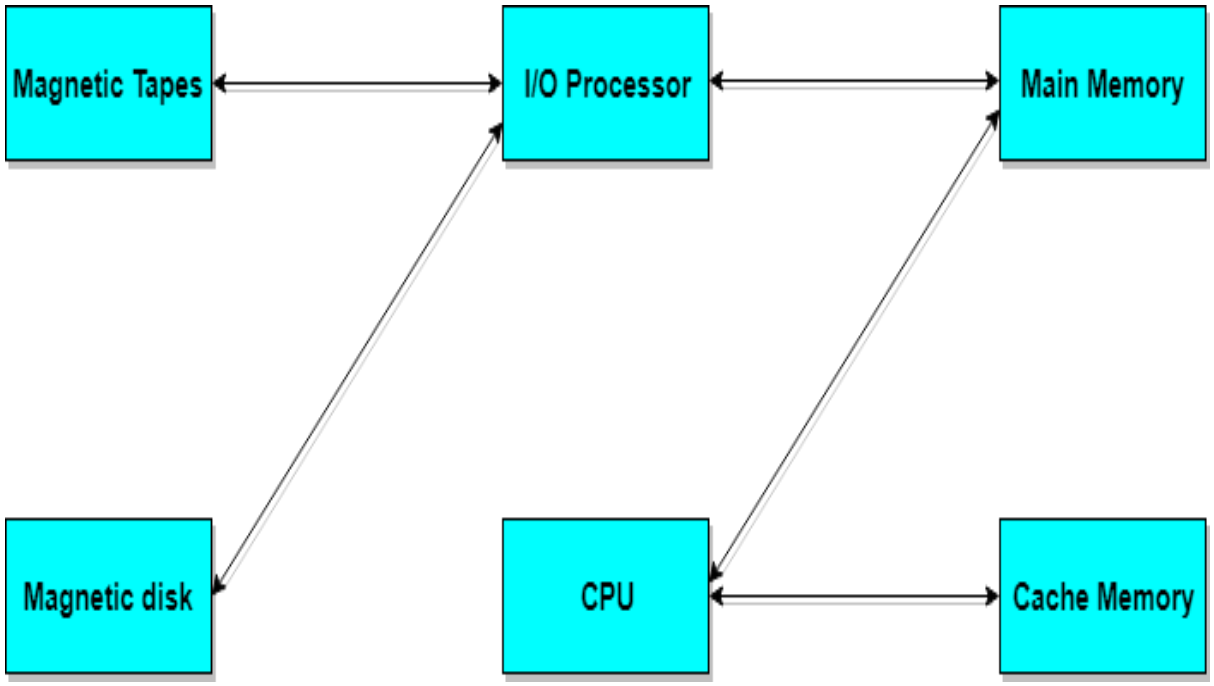
Memory Hierarchy

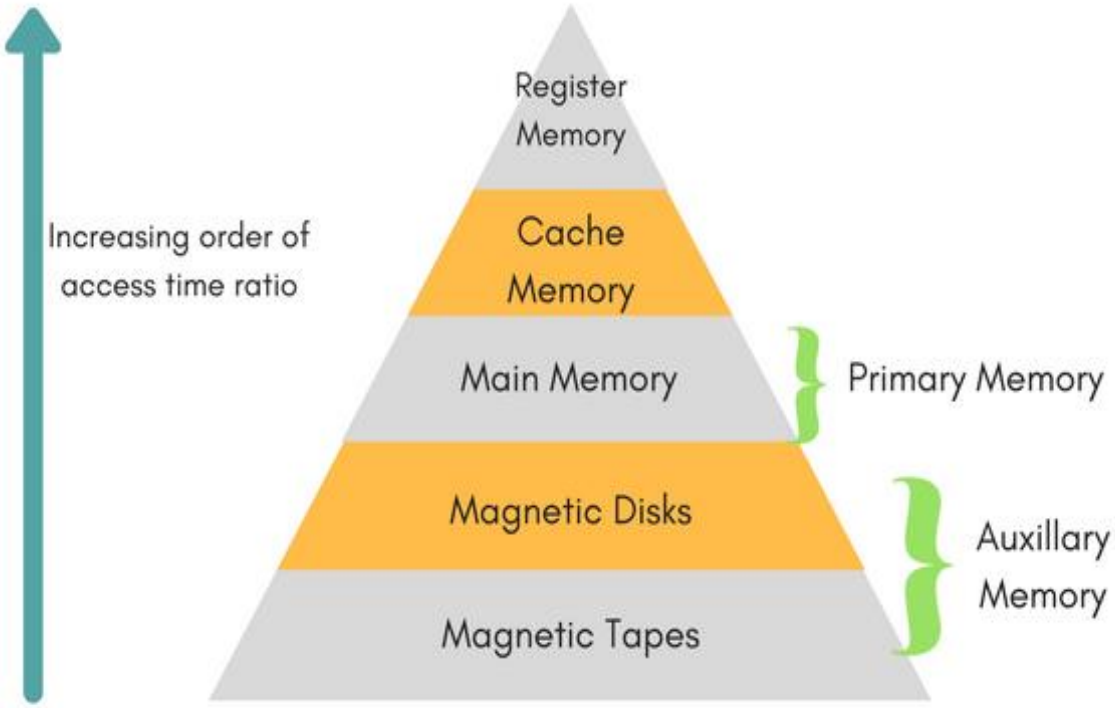
a. The total memory capacity of a computer can be visualized by hierarchy of components. The memory hierarchy system consists of all storage devices contained in a computer system from the slow Auxiliary Memory to fast Main Memory and to smaller Cache memory. **Auxillary memory** access time is generally **1000 times** that of the main memory, hence it is at the bottom of the hierarchy.The **main memory** occupies the central position because it is equipped to communicate directly with the CPU and with auxiliary memory devices through Input/output processor (I/O).

1. Memory Hierarchy

1. The Memory Hierarchy system consists of all storage devices employed in computer system.
2. The total memory capacity of a computer can visualized as being a hierarchy components.
3. The Memory Hierarchy Diagram as shown below



OR



4. From above Diagram, we can interpret the following points.
   From top i.e register memory to Down Auxiliary Memory
   a. Cost: The cost decreases from top to down, i.e. Register memory much cost than cache memory, cache memory more cost than main memory ...etc.
   b. Size: The Storage Capacity increases from top to down.
   c. Frequency of Access: The frequency of access time decreases from top to down. The CPU access more number of times cache memory. If requested data not found than CPU access the main memory.
      A very high speed memory called cache memory is sometimes used to increase speed of processing by making current programs and data available to the CPU at rapid rate.
      The Cache memory resolves the speed mismatch between the CPU and Main memory.
   d. The basic validity of condition (c) is known to be Locality of Reference. The main principle of cache memory is known to locality of reference.
5. The overall goal of using memory is to obtain the highest possible average access speed while minimizing the total cost of the entire system.

2. Main Memory

a. The main memory is the central storage unit in a computer system.
b. It is relatively larger and fast memory used to store programs and data during the computer operation.
c. The principle technology used for main memory is based on Semiconductor integrated chips.

d. The integrated RAM Chips are available in two possible operating modes.
   - Static RAM
     The Static RAM consists of essentially of Internal Flip-flops that store the binary information.
     The Stored Binary Information remains valid as long as power is applied to the unit.
   - Dynamic RAM
     The dynamic RAM stores the binary information in the form of electric charges that are applied to the capacitors.

   The dynamic RAM offers reduces power consumption and larger storage capacity in Single memory chip whereas static RAM is easier to use and has shorter read and write cycles.

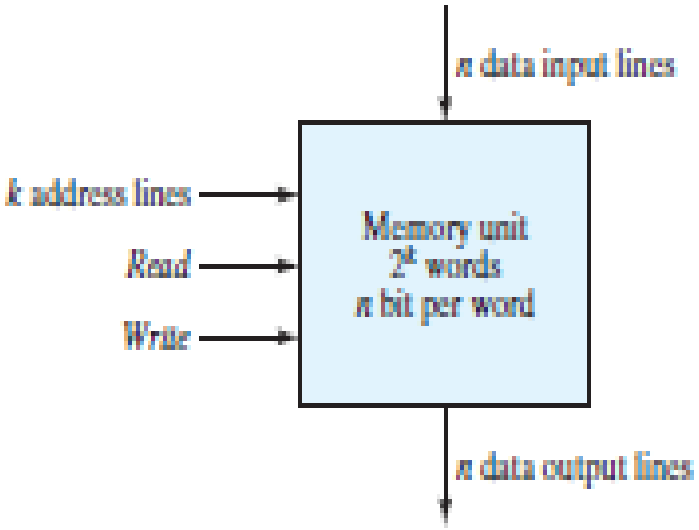
   The static RAM used for implementing cache memory where as dynamic RAM used for Main memory.

There are two types of memories that are used in digital systems

- *Random-access memory* (RAM):RAM stores new information for later use
- *Read-only memory* (ROM)

1. The process of storing new information into memory is referred to as a memory *write* operation.
2. The process of transferring the stored information out of memory is referred to as a memory *read* operation.
3. RAM can perform both write and read operations.
4. ROM can perform only the read operation.

The main memory block Diagram



Simple Problems Given to calculate address line and data lines for RAM chip and ROM Chip.

RAM Chip

Write and Read Operations

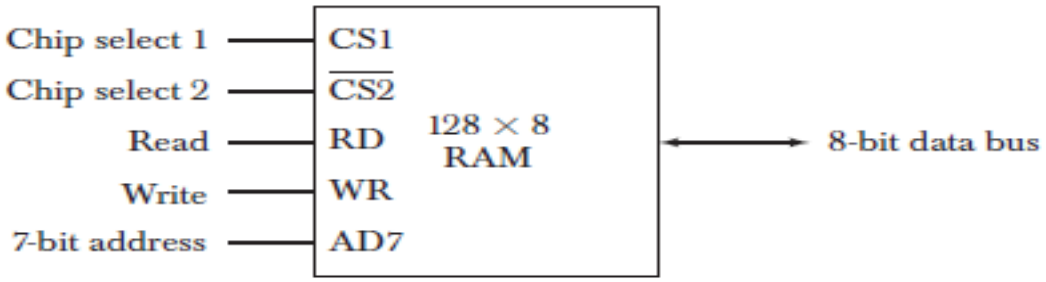
Write operation

a. Apply the binary address of the desired word to the address lines.
b. Apply the data bits that must be stored in memory to the data input lines.
c. Activate the *write* input.

Read operation

a. Apply the binary address of the desired word to the address lines.
b. Activate the *read* input.

RAM Chip Block Diagram.



(a) Block diagram

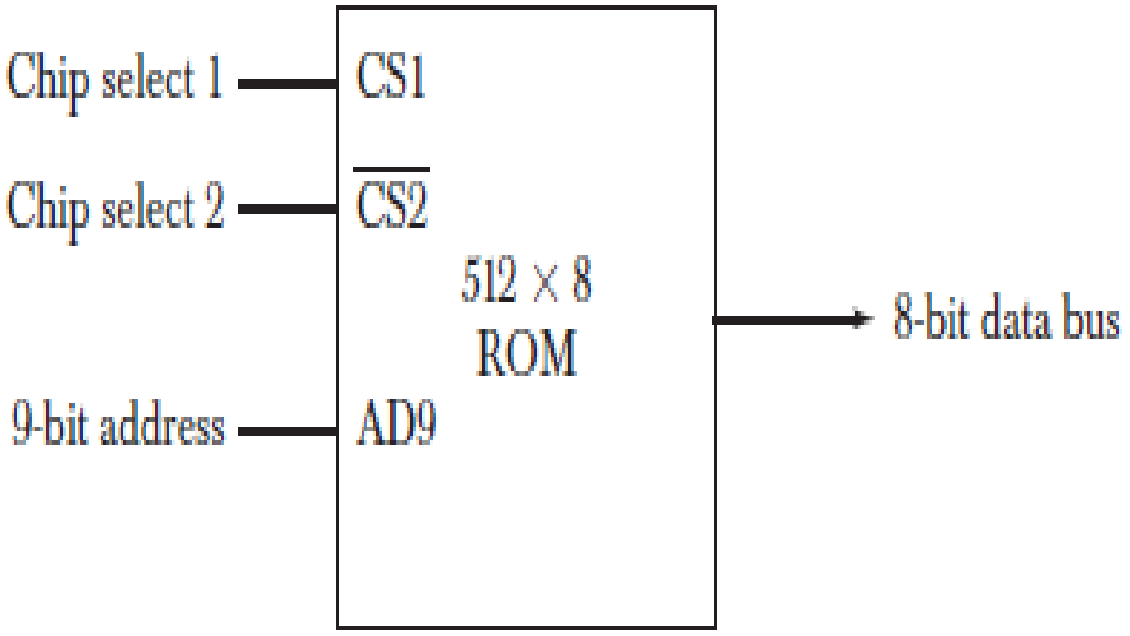
CS1	$\overline{CS2}$	RD	WR	Memory function	State of data bus
0	0	×	×	Inhibit	High-impedance
0	1	×	×	Inhibit	High-impedance
1	0	0	0	Inhibit	High-impedance
1	0	0	1	Write	Input data to RAM
1	0	1	×	Read	Output data from RAM
1	1	×	×	Inhibit	High-impedance

(b) Function table

Draw neat diagram for RAM Chip of Size 256×8 and 1024×16

ROM Chip

Read only memory: Once ROM Chip selects by means of Chip Select 1 and Chip Select 2, Data specified address location can transferred through datelines to destination unit. So there is no need read signal and write signal.



Draw neat diagram for RAM Chip of Size 256×8 and 1024×16

Memory Address Map:

The addressing of memory can be established by means of table that specifies the memory address assigned to each chip, This table is known to memory address map or Address resolution table.

For example,

Assume that computer system needs 512 bytes of RAM and 512 bytes of ROM. This means that RAM chip of size 512×8 and ROM of Size 512×8.

But RAM chip of Size 128×8 and ROM chip of Size 512×8 only available.

Total RAM Memory Size= 512×8

RAM memory Chip Size = 128×8

Number of 128×8 RAM chips Required = Total Memory Size /Unit RAM Chip size

No. Of Chips Required = 512×8/128×8 = 4

So, there are 4 RAM chips requited to construct total memory of 512×8 RAM chip and one 512×8 ROM chip.

Using above information, we can construct Address Memory map or Address Resolution Table.

The Address resolution table consist of Component Name , Hexadecimal Address and Address bus bits.

Address Memory bus table as shown below

TABLE 12-1 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

For RAM chip of Size 128×8, seven address line and 8 data lines.

ROM chip of Size 512×8, Nine address lines and 8 data lines.

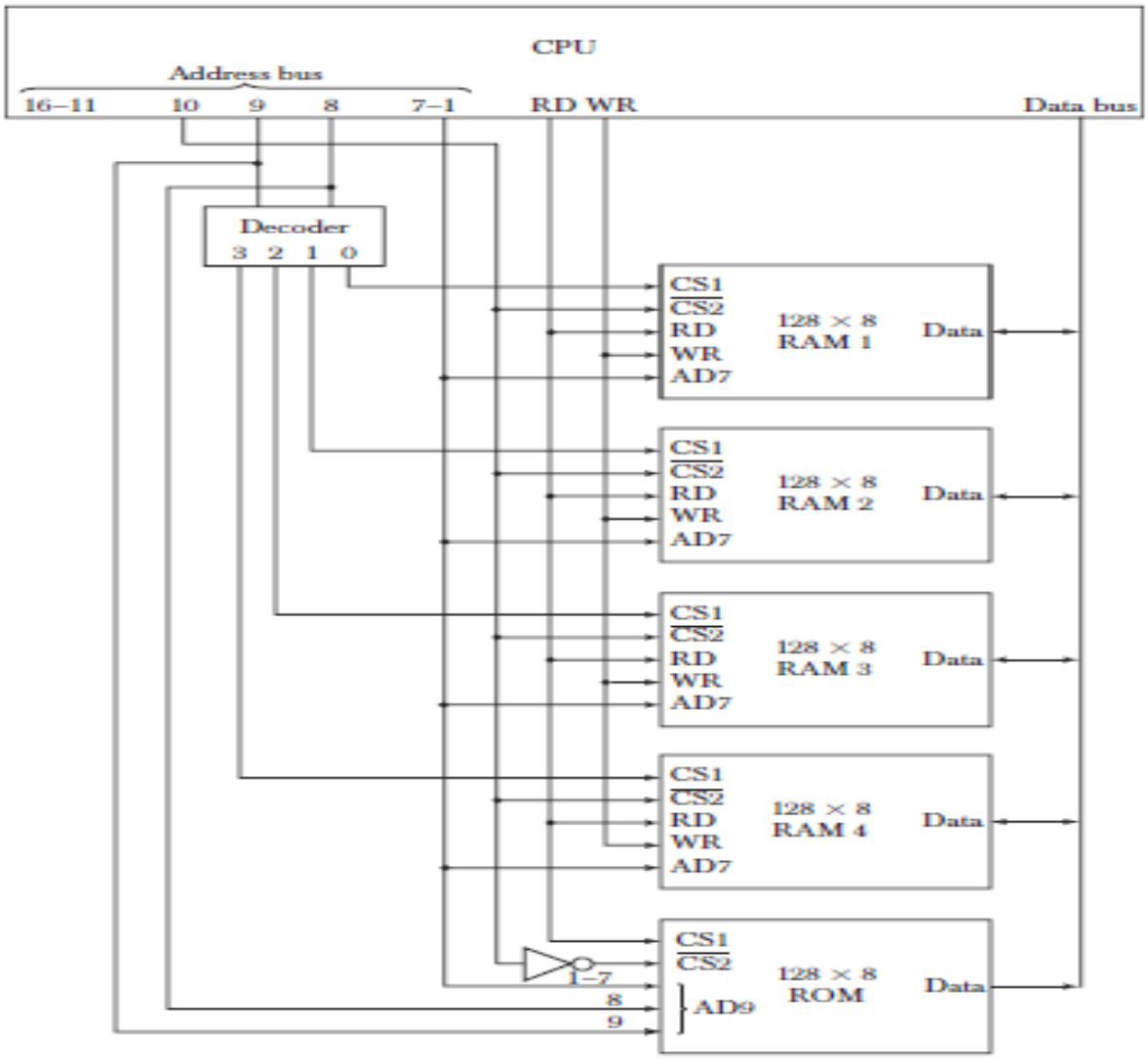
Note: Exercise: Construct Memory Address Map Table total memory of RAM Size 1024×8 and ROM Chip of 512×8 using 256×8 RAM chip and One ROM chip of Size 512×8.

Memory Connection to the CPU

Using above address map, we can memory modules are interfaced to CPU using decoder circuit.

There are Four RAM Chips are available, So, 2:4 decoder Circuit required.

The Circuit Diagram as shown below



Note: In diagram ROM chip of size 512×8 but diagram show 128×8 which is wrong. Modify accordingly.

Cache Memory

a. When the processor needs to read or write a location in main memory, it first checks for a corresponding entry in the cache.

b. If the processor finds that the memory location is in the cache, a **cache hit** has occurred and data is read from cache

c. If the processor **does not** find the memory location in the cache, a **cache miss** has occurred. For a cache miss, the cache allocates a new entry and copies in data from main memory, then the request is fulfilled from the contents of the cache.

d. The performance of cache memory is frequently measured in terms of a quantity called **Hit ratio.**

Hit ratio = hit / (hit + miss) = no. of hits/total accesses

The cache memory based on the principle of Locality of Reference in it reference to memory at any given interval of time tend to confined within few localized areas in the memory.

The performance of cache memory is frequently measured in terms of quantity called hit ratio.

When CPU refers to word and if it finds the word in the cache , it is called hit otherwise miss.

Memory Mapping Techniques:

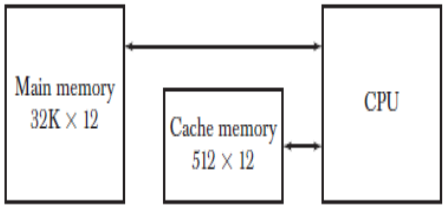
The Basic characteristic of cache memory is its fast access time. Therefore, very little or no time must wasted when searching for words in the cache. The transformation of data from main memory to cache memory is referred to as a mapping process.

Three types of mapping techniques as

a. Associative Mapping
b. Direct Mapping
c. Set-associative mapping

To demonstrate above mapping techniques, we use below example.

Cache Memory

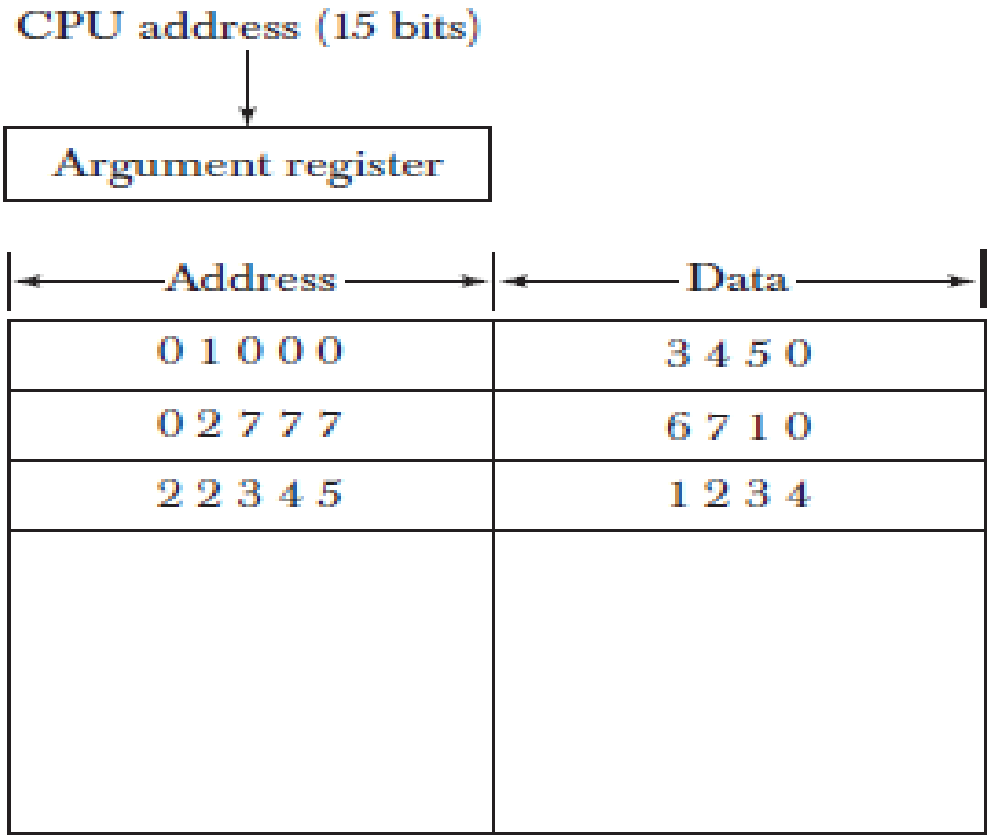


The main memory 32K location and each location of size 12bits and cache memory 512 locations and each location 12bits.The CPU communicates 15 bit address for both memory

Associative Mapping.

The fastest and most flexible cache organization uses an associative memory. In associative memory stores both address and content (data) of the memory word. This allows any location of cache memory stores any word from main memory.

An associative mapping for above example as shown below.



An CPU address of 15 bits placed in Augment Register as shown five digit octal number and 12 bit data word as shown 4 digit octal number.

If address is found 12 bit data read and send to CPU otherwise main memory accessed the word.

The main memory word transferred to cache memory location, if cache memory location full than address data must be displaced to make room for a pair that needed and not presently in cache. A simple procedure need to adopted which memory location of cache replace based on replacement algorithms (FCFS, LRU.....etc).

Disadvantage of Associative Memory:

1. More expensive as compared to Random Access Memory (RAM) due to added logic associative with each cell.
2. Complexity More due added logic.

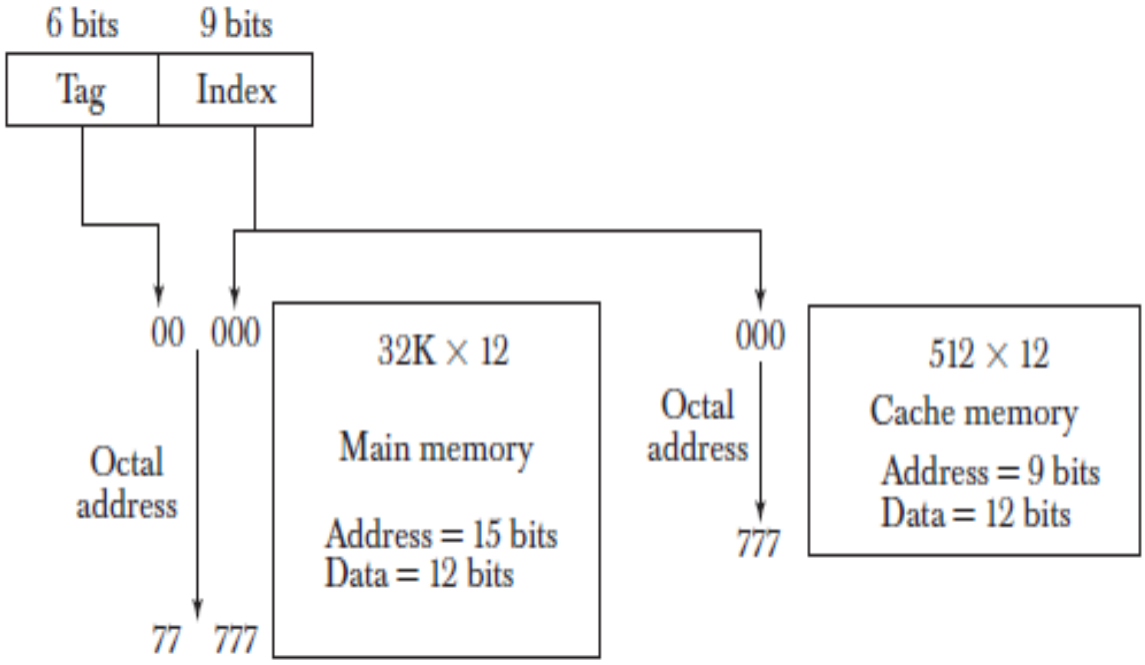
Advantage of Associative Memory:

1. Simplicity: The main memory word can be placed anywhere in the cache memory location.

Direct Memory: The CPU address divided into two fields, in our example 15bit address is divided into 9 least significant bits constitute the index field and remaining 6 bits tag

bits. The cache memory address lines 9 from its size (512×12) which is known to be cache address.

Addressing Relation between cache memory and main memory as give below



The Direct Memory Cache mapping table as

Memory address	Memory data
00000	1220
00777	2340
01000	3450
01777	4560
02000	5670
02777	6710

(a) Main Memory

Index address	Tag	Data
000	00	1220
777	02	6710

(b) Cache memory

Each word in the cache consists of data word and its associated tag .When new word brought to the cache the tag bits are stored alongside data bits.

When CPU generates memory request, the index field used for address to access the cache and tag field of both main memory and cache compared. If two tag fields are same then there is hit otherwise miss pentality.The required word read from memory and stored in the cache with new tag field.

Disadvantages of Direct Memory Mapping:

1. Two words of the same index in their address but different tag values cannot reside in the same memory at the same time.

Set-Associative Mapping:

To overcome the disadvantage of Direct Memory and Associate Mapping techniques, A third type cache memory organization called set associative mapping technique used.

This mapping organization in that each word of cache can store two or more of the memories under same index address.

Each data word is stored together with its tag –data items in one word of cache is said to form a set.

For our example list above, two ways set cache organization can be illustrate below.

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

See in above table, we can store two different data items with same index of cache with different tag values as shown, (for example Data values 6710 and 2340 and having same index with tag values 02 and 00).

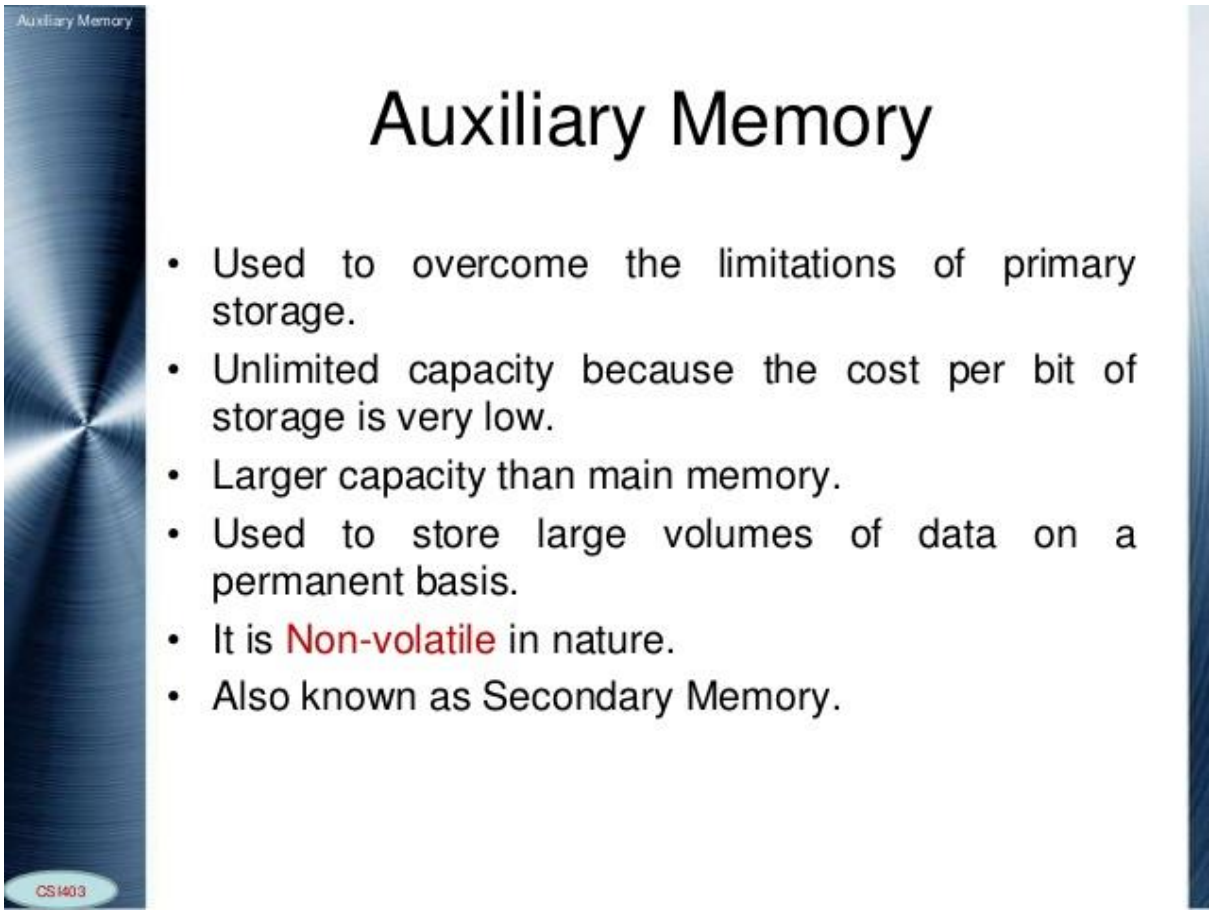
Writing into cache

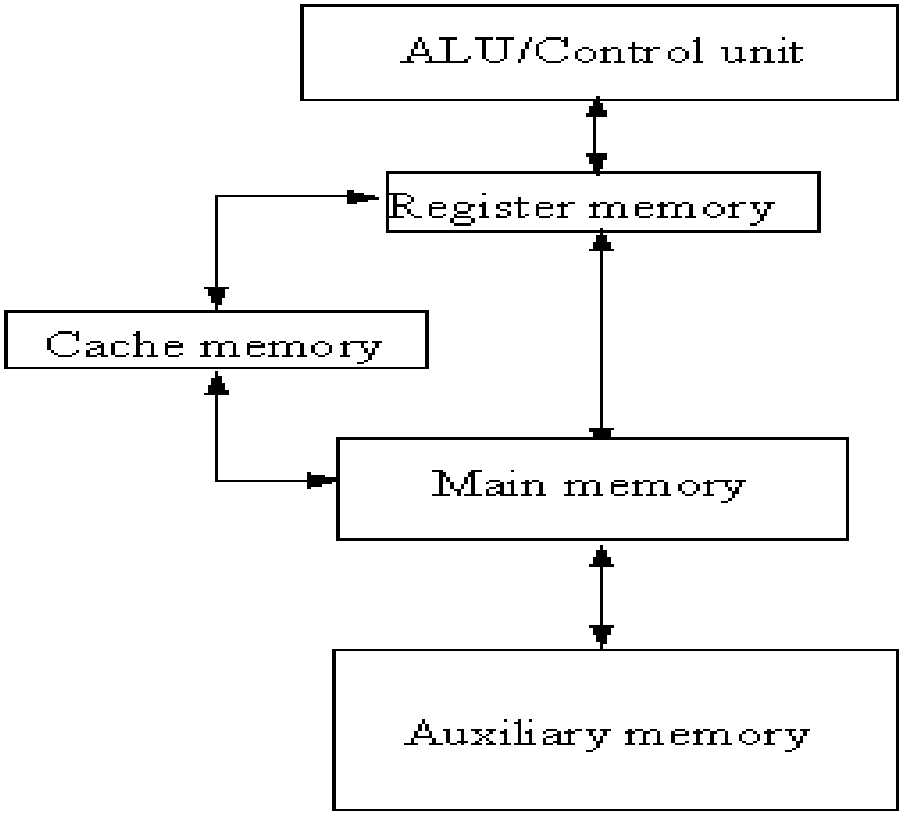
Whenever CPU reads the data from cache memory, the main memory is not involved in the transfer.

But whenever CPU writes the data into cache memory after computation, the main need to updates modified data. This operation write operation can be performed in two different ways.

1. Write through method.
   Both cache memory and main memory data updates in parallel when CPU performs write operation. The main memory always contains the same data as the cache.
2. Write Back Method.
   Initially, Cache memory location is updated during write operation and its location marked as a flag. The corresponding main memory location updates whenever it is removed from cache memory location. This known to be Write back method.

Auxiliary Memory



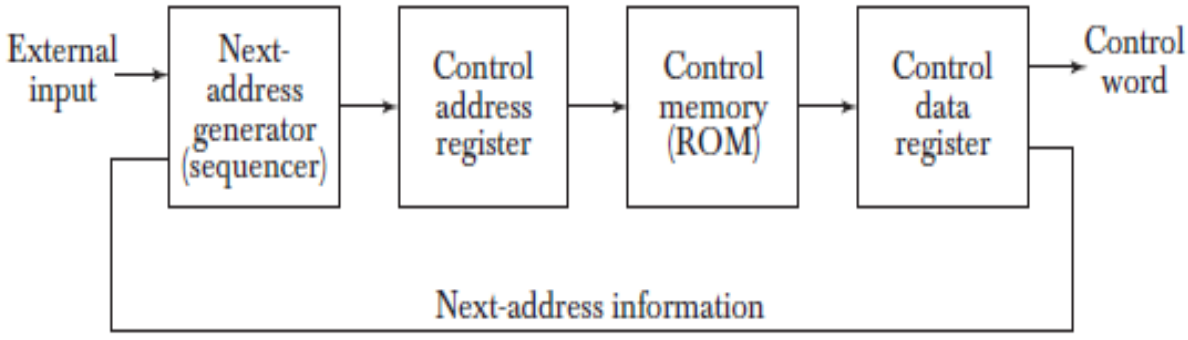

ALU/Control unit
Register memory
Cache memory
Main memory
Auxiliary memory

Control Memory

Control Unit

a. The main function of the control unit in a digital computer is to initiate sequences of micro-operations.
b. The number of different types of microoperations that are available in a given system is finite.
c. The complexity of the digital system is derived from the number of sequences of microoperations that are performed.
d. The control function that specifies a microoperation is a binary variable. When it is in one binary state, the corresponding microoperation is executed.
e. The control unit initiates a series of sequential steps of microoperations.The control variables at any given time can be represented by a string of 1's and 0's called a control word.
f. Control words can be programmed to perform various operations on the components of the system.
g. A control unit whose binary control variables are stored in memory is called a *microprogrammed control unit.*
h. Each word in control memory contains within it a *microinstruction.* The microinstruction specifies one or more microoperations for the system.
i. A sequence of microinstructions constitutes a *microprogram.* Since alterations of the microprogram are not needed once the control unit is in operation, the control memory can be a read-only memory (ROM).
j. The use of a microprogram involves placing all control variables in words of ROM for use by the control unit through successive read operations. The content of the word in ROM at a given address specifies a microinstruction.

A memory that is part of a control unit is referred to as a *control memory.*A computer that employs a microprogrammed control unit will have two separate memories: a main memory and a control memory.



The use of a microprogram involves placing all control variables in words of ROM for use by the control unit through successive read operations. The content of the word in ROM at a given address specifies a microinstruction.

The control data register holds the present microinstruction while the next address is computed and read from memory. The data register is sometimes called a *pipeline register.* It allows the execution of the micro operations specified by the control word simultaneously with the generation of the next microinstruction.

This configuration requires a two-phase clock, with one clock applied to the address register and the other to the data register.

There are two methods for designing control unit are

- Hardwired Control Unit
  The Control Signals are generated by hardware using conventional logic design techniques the control unit is said to be Hardwired Control unit.
- Microprogram Control Unit.
  The control Signals are generated by software for controlling microoperation sequence in digital system.

a. The key characteristics of hardwired control logic are high-speed operation, expensive, relatively complex, and no flexibility of adding new instructions.
b. Microprogramming low-speed operation, low cost, easy to debug and more flexible.

Address Sequencing

a. Microinstructions are stored in control memory in groups, with each address sequencing group specifying a *routine.*
b. Each computer instruction has its own microprogram routine in control memory to generate the microoperations that execute the instruction.
c. The hardware that controls the address sequencing of the control memory must be capable of sequencing the microinstructions within a routine and be able to branch from one routine to another.
d. To appreciate the address sequencing in a microprogram control unit, let us enumerate the steps that the control must undergo during the execution of a single computer instruction.
e. An initial address is loaded into the control address register when power is turned on in the computer. This address is usually the address of the first microinstruction that activates the instruction fetch routine.
f. The fetch routine may be sequenced by incrementing the control address register through the rest of its microinstructions.
g. At the end of the fetch routine, the instruction is in the instruction register of the computer.
h. The control memory next must go through the routine that determines the effective address of the operand.
i. A machine instruction may have bits that specify various addressing modes, such as indirect address and index registers.
j. The effective address computation routine in control memory can be reached through a branch microinstruction, which is conditioned on the status of the mode bits of the instruction.

k. When the effective address computation routine is completed, the address of the operand is available in the memory address register.
l. Each instruction has its own microprogram routine stored in a given location of control memory.
m. The transformation from the instruction code bits to an address in control memory where the routine is located is referred to as a *mapping process.*
n. When the execution of the instruction is completed, control must return to the fetch routine. This is accomplished by executing an unconditional branch microinstruction to the first address of the fetch routine.
o. In summary, the address sequencing capabilities required in a control memory are:
    - Incrementing of the control address register.
    - Unconditional branch or conditional branch, depending on status bit conditions.
    - A mapping process from the bits of the instruction to an address for control memory.
    - A facility for subroutine call and return.

p. A block diagram of a control memory and the associated hardware needed for selecting the next microinstruction address. The microinstruction in control memory contains a set of bits to initiate microoperations in computer registers and other bits to specify the method by which the next address is obtained.

q. The diagram shows four different paths from which the control address register (CAR) receives the address.

The diagram given below.