Computer Architecture module 3

Computer Organization & Architecture (Mahatma Gandhi University)

Module 3

Main Memory

Module 3: Main Memory: Organization of RAM, ROM, Auxiliary memory, Cache memory, Virtual Memory, Memory mapping Techniques.

MEMORY HIERARCHY

- The memory unit is needed for storing programs and data.
- The memory unit that communicate directly with the CPU is called the ***main memory*** and the devices that provide backup storage are called the ***auxiliary memory***.

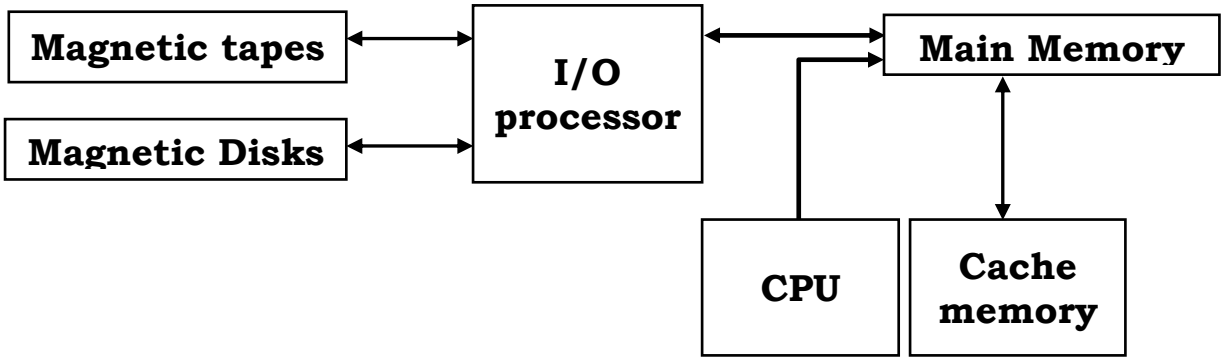


Figure: Memory Hierarchy

- At the ***bottom*** of the hierarchy are relativity ***slow magnetic tapes and discs used as backup storage.***
- The ***main memory*** occupies the ***central position*** to communicate directly with CPU.
- A special very ***high-speed memory called a cache*** is used to increase the speed of processing. Cache is used to compensate the speed difference between main memory and the processor.
  - **Main memory**: average capacity, high cost and low access speed.
  - **Auxiliary memory**: high storage capacity, cheap and low access speed.
  - **Cache memory**: very small, expensive and very high access speed.

- ***The overall goal of using a memory hierarchy is to obtain the highest possible average access speed while minimizing the total cost and maximizing the storage capacity*.**

MAIN MEMORY

- Main memory is the central storage unit in the computer system.
- It is relatively large and fast memory used to store programs and data.
- Main memory is divided into
  - **RAM**
  - **ROM**

RAM (Random Access Memory):-

- It is a volatile memory.
- RAM's job is to hold programs and data while they are in use.
- Physically, RAM consists of chips on a small circuit board.
- RAM is designed to be instantly accessible by the CPU or programs.
- The random in RAM implies that any portion of RAM can be accessed at any time. This helps make RAM very fast.
- RAM chips are available in two possible modes:
  1. **Static and**
  2. **Dynamic**

SI No	Static Random Access Memory **(SRAM)**	Dynamic Random Access Memory **(DRAM)**
1	***Flip-Flops*** are used in an integrated circuit.	***Capacitors*** are used in an integrated circuit.
2	Does not need to be periodically refreshed. The binary information stored in the SRAM remains valid as long as the power is applied	Binary information are stored in form of ***electrical charges***. Real capacitors leak charge so capacitors need to be refreshed periodically. The stored charge on

	to the unit.	the capacitor tends to discharge with time and the capacitor must be periodically recharged by refreshing the DRAM.
3	Expensive and have complex structure (6 transistors) so not used in high capacity applications.	Simple structure (1 transistor and 1 capacitor per bit) so it has very high density.
4	Faster and consumes low power.	It offers reduced power consumption and large storage capacity in single memory chip.

ROM (Read Only Memory):-

- It is the permanent storage area used to store the constant value that does not change in value once the production of the computer is completed.
- The contents of ROM remain unchanged after power is turned OFF and ON again.
- It is non-volatile. i.e., the chips hold data even when computer is unplugged.
- In fact, putting data permanently into this kind of memory is called ***"burning in the data"*** and it is usually done at the manufacturing time.
- The data in these chips is only read and used; therefore, the memory is called **read-only memory (ROM).**

Why do we need ROM?

- The RAM chips SRAM and SDRAM chips are volatile; i.e., Lose the contents when the power is turned off.

- Many applications need memory devices to retain contents after the power is turned off.
    - For example, computer is turned on, the operating system must be loaded from the disk into the memory.
    - Store instructions which would load the OS from the disk.
    - Need to store these instructions so that they will not be lost after the power is turned off.
- We need to store the instructions into a non-volatile memory.
- Non-volatile memory is read in the same manner as volatile memory.
- Separate writing process is needed to place information in this memory.
- Normal operation involves only reading of data, ***this type of memory is called Read-Only memory (ROM).***

➢ ROM contains two types of programs.

- **BIOS** (**Basic input output system (BIOS)-** When a computer is turned on, it must know how to start. ROM contains a set of start-up instructions called the BIOS for a computer.
- **POST** (**Power On Self-Test**) – It is a routine which ensure that the system is functioning properly and all expected hardware devices are present. This routine is called the **POST.**

➢ **Different types of ROM: -**

MROM (Mask ROM)

- It is a type of ROM whose contents are pre-programmed when the chip is manufactured, it can't be programmed by the user.
- It is inexpensive.
- Data can't be erasable.

Programmable ROM (PROM)

- It is a computer memory chip that can be ***programmed once after it is created.***
- Once the PROM is programmed, the information written is permanent and cannot be erased or deleted.
- When the PROM is created (manufactured), all bits read as "1." During the programming, any bit needing to be changed to a "0" is burned into the chip using a **gang programmer**.
- If a PROM is programmed with an error or needs updated, the chip is discarded and a new PROM is created, replacing the old chip.
- Example: CD-R, DVD-R

Erasable PROM (EPROM)

- Hardware manufacturers use EPROM when it may be needed that the data contained on the PROM needs to be changed.
- An EPROM can be reprogrammed by exposed to ultraviolet light.
- Example: CD-RW, DVD-RW.

Electrically EPROM (EEPROM)

- **EEPROM** is a PROM that can be erased and reprogrammed using electrical charges.
- It can be erased and reprogrammed about ten thousand times.
- Both erasing and programming take about 4 to 10 ms (millisecond).
- In EEPROM, any location can be selectively erased and programmed.
- EEPROMs can be erased one byte at a time, rather than erasing the entire chip. Hence, the process of re-programming is flexible but slow.

- Example: Pen drive

Flash memory:

- Has similar approach to EEPROM.
- Read the contents of a single cell, but write the contents of an entire block of cells.
- Flash devices have greater density.
- Higher capacity and low storage cost per bit.
- Power consumption of flash memory is very low, making it attractive for use in equipment that is battery-driven.
- Single flash chips are not sufficiently large, so
  - larger memory modules are implemented using
  - flash cards and flash drives.

RAM ORGANISATION: -

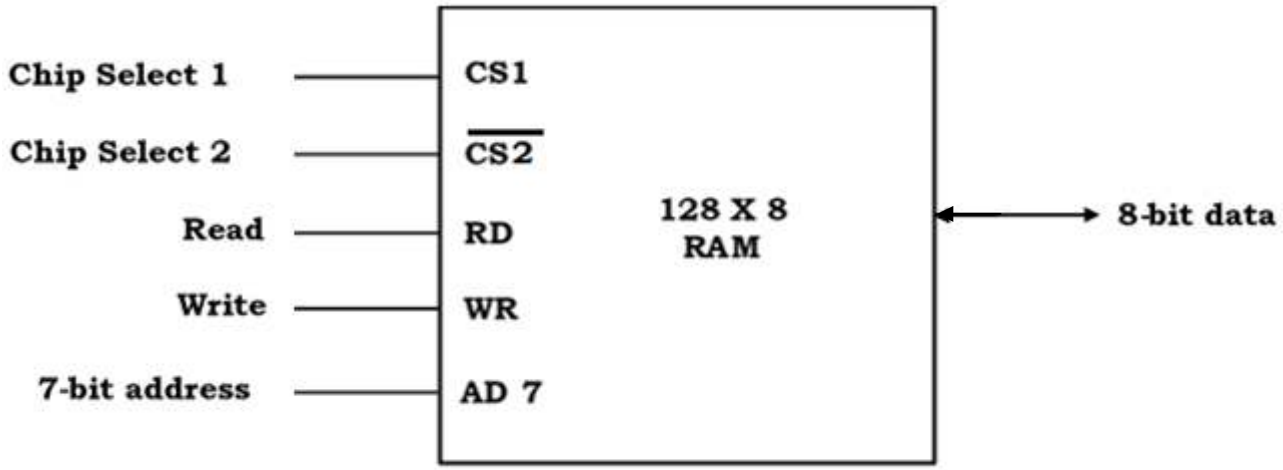


Figure: RAM Chip

- 128 X 8 represents the memory capacity means capacity is 128 words of 8 bits each.
- 8-bit bidirectional data bus allows the transfer of data from memory at the time of read operation or to the memory at the time of write operation.

- AD7 (7-bit address, since $128=2^7$ we need 7 bits to represent each location).
- RD and WR are the read and write inputs that we specify the memory operation.
- Two chips select controls (CS)- for enabling particular chip. The unit is in operation only when CS1=1 and CS2=0(complement input).

CS1	CS2	RD	WR	Memory Function	State of data bus
0	0	X	X	Inhibit	High impedance
0	1	X	X	Inhibit	High impedance
1	0	0	0	Inhibit	High impedance
1	0	0	1	Write	I/P data to RAM
1	0	1	0	Read	O/P data from RAM
1	0	1	1	Read (Priority for read)	O/P data from RAM
1	1	X	X	Inhibit	High impedance

Figure :Function Table

- When chip select inputs enabled i.e., CS1=1 and CS2=0,
    1. When write input is enabled the memory write operation occurs.
    2. When read input is enabled the memory read operation occurs.
    3. When read and write inputs are enabled, the memory read operation occurs since the read operation has high priority.

NOTE: Memory Read → byte information is transferred from memory to the data bus. **Memory write** → byte information is transferred from data bus to the memory location.

ROM ORGANISATION

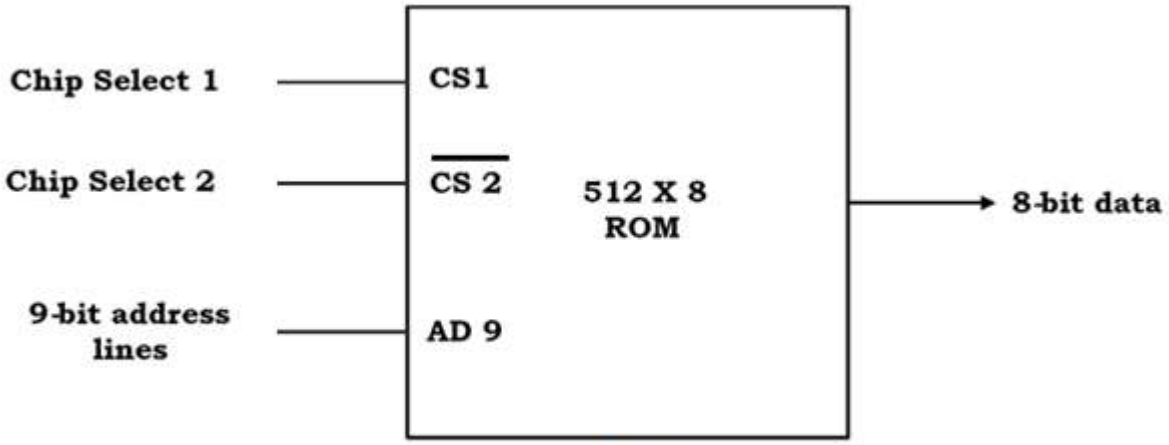


Figure: ROM Chip

- Since a ROM can only read, the ***data bus is unidirectional.***
- For the same size chip, it is possible to store more bits in ROM than in RAM, because the internal binary cells in ROM occupy less space than RAM.
- 512 X 8 capacity means 512 words of ROM 8 bits each.
- The two chip select inputs must be CS1 and CS2 =0 for the unit to operate. Otherwise memory is in an inhibit state.
- No need for RD and WR control lines because the unit can only read.
- AD9- 9-bit address (Since $512=2^9$ we need 9 bits to represent each location.)

Memory Address Map

Component	Hexadecimal address	Address Bus										Address in decimal
		10	9	8	7	6	5	4	3	2	1	
RAM 1	0000-007F	0	0	0	X	X	X	X	X	X	X	0-127
RAM 2	0080-00FF	0	0	1	X	X	X	X	X	X	X	128-255
RAM 3	0100-017F	0	1	0	X	X	X	X	X	X	X	256-386
RAM 4	0180-01FF	0	1	1	X	X	X	X	X	X	X	984-511
ROM	0200-03FF	1	X	X	X	X	X	X	X	X	X	512-1023

- Suppose a computer system needs 512 bytes of RAM and 512 bytes of ROM.
- The addressing of the memory can be established by means of a table that specifies the memory address assign to each chip. ***The table called a memory address map is a pictorial representation of assigned address space for each chip in the system.***
- The component column specifies whether a RAM or ROM chip is used.
- Hexadecimal address column gives a range of hexadecimal address for each chip.
- The address bus lines are listed in the third column.
- X's are always assigned to low order bus lines (line 1-7 for RAM and 1-9 for ROM).
- Bus lines 8 & 9 are used to distinguish between 4 RAM chips. It represents 4 distinct binary combinations.
- i.e., 00-RAM1, 00-RAM2,10-RAM3 ,11-RAM4
- Address bus line 10 is used to distinguish between RAM and ROM. When line 10 is zero, CPU selects a RAM and when line 10 is one CPU selects a ROM.

AUXILIARY MEMORY: -

- Auxiliary memory is used to supplement the main memory (backup storage is another name of auxiliary memory).
- The most common auxiliary memory devices are magnetic disk and tapes.

Characteristics.

Characteristic		Description
Access mode	-	Sequential access, Random access.
Access time	-	Average time required to reach a storage location in memory.

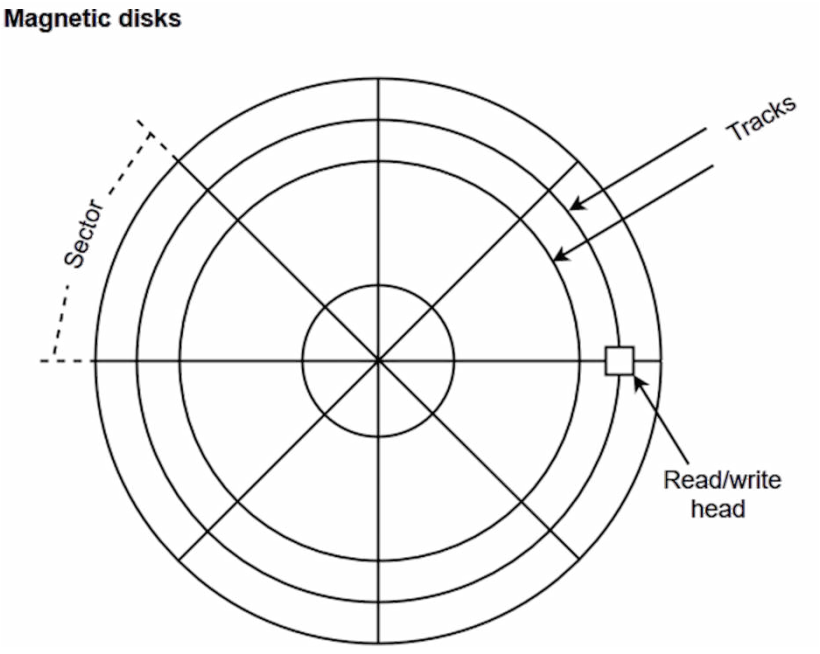
Transfer rate - Number of characters or word that the device can transfer per second after it has been positioned at the beginning of the record.

Capacity - Secondary storage can store large volumes of data in sets of multiple disks

Cost - It is much lesser expensive to store data on a tape or disk than primary memory.

Magnetic Disks: -

- Magnetic disk is a circular plate, constructed of metal or plastic, coated with magnetized material.
- Often both of the disks are used.
- It is possible to stack several disks on one spindle.
- Bits are stored in spots along a concentric circle called ***tracks***.
- Tracks are divided into sections called ***sectors***.

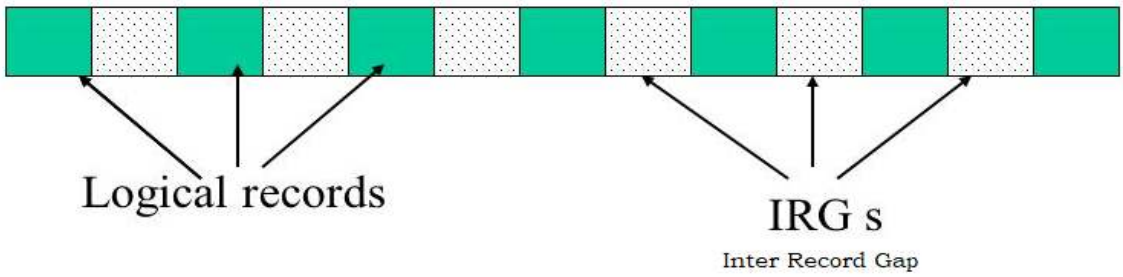


- In most system, the minimum quantity of information which can be transferred is a sector.
- Some system uses single Read/Write head for each disk surface.
- It uses a mechanical assembly to position Read/Write head.

- In the case of system using separate read write head, the address bits are selected with the decoder circuit.
- A single disc system is addressed by address bits that specifies the disk number, disc surface, sector number, and track within the sector.
- A track near the circumference is longer than a track near the centre.
- Disk that are permanently attached to the unit and cannot be removed by the occasional uses are called ***hard disks.***
- A hard drive with removable disk is called a ***floppy disk***.
- There are 2 sizes for floppy, with diameter of 5.25 and 3.5 inches.
- The 3.5-inch disks are smaller and can be store more data.

Magnetic tape:-

- Magnetic tape is a ***plastic strip coated with the magnetic recording medium.***
- Bits are recorded as magnetic spot on the tape along several tracks.
- Usually 7 or 9 bits are recorded simultaneously to form a character together with the 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, move forward, reversed or rewound.
- But they cannot be started or stopped fast enough between individual characters.
- Information is stored as record separated by gaps(IRG's-Inter Record Gaps).
- Each record has an identification bit pattern at the beginning and end.
- By reading the bit pattern at the beginning the tape control recognises beginning of a record and by reading the bit pattern at end, it recognizes beginning of a gap.
- Record may be fixed length or variable length.



CACHE MEMORY

- Cache mechanism is based on the property of computer program called the ***locality of reference.***
- ***This property states that over a short interval of time, the address generated by the typical program refers to a new localised area of memory repeatedly, while the remainders of memory are addressed relatively infrequently.***
- ***Cache memory is a fast-small memory placed between the CPU and the main memory.***
- It is used to compensate the speed difference between the main memory and the CPU.
- The ***basic operations 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 read from the cache.
    - If the word is not in the cache, then the main memory is read and the content is transferred from main memory to the cache.

Locality of references: -

- Cache mechanism is based on the property of computer program called the ***locality of reference.***
- ***This property states that over a short interval of time, the address generated by the typical program refers to a new localised area of memory repeatedly, while the remainders of memory are addressed relatively infrequently.***
- It manifests itself in 2 ways.

- **Temporal and**
- **Spatial.**

- ***Temporal means*** that the recently executed instruction is likely to be executed again, very soon. ***The spatial means*** that the instruction in close proximity are also likely to be executed soon.
- When a program loop is executed, the CPU repeatedly refers to a set of instructions in memory.
- Every time a subroutine is called; its set of instructions are fetched.
- Thus, ***loops and subroutine related to locality of reference***.

✓ **Example**

```
for (i=1;i<=5;i++)
      {
      for (j=1;j<=i; j++)
        {
          cout<<j;
        }
          cout<<'\n'
      }
```

O/P

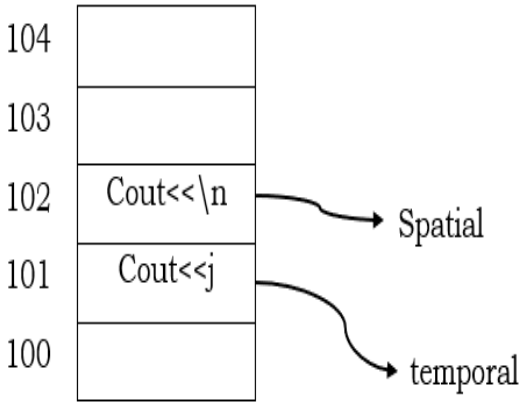
1

12

123

1234

12345



Hit Ratio

- ***The performance of cache memory is measured in terms of quantity called hit ratio.***
- When the CPU refers to the memory and find the word in cache, it is a ***hit.*** If the word is not found in the cache, it is in the main memory and it count as ***miss.***

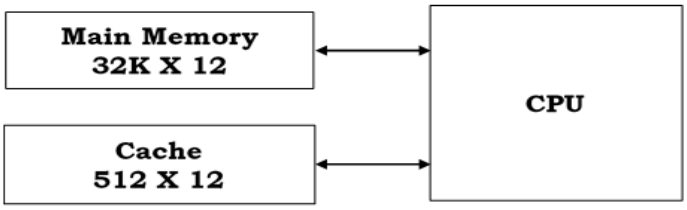
$$\text{Hit Ratio} = \frac{\text{Number of hits}}{\text{Total number of references}}$$

$$\text{(number of hits + number of misses)}$$

Mapping

- The basic property of the cache memory is its ***fast access time.***

- The transformation of data from main memory to cache memory is referred to as the ***mapping process.***
- Three types of mapping procedures are
  - ***1. Associative mapping***
  - ***2. Direct mapping***
  - ***3. Set associative mapping***
- Consider an Example



- Main memory capacity → 32K X 12

  i.e., 32K words of 12 bits each

  [1K=1024, 32K=32 X 1024,2^5 X 2^{10}=2^{15}]
- Cache capacity → (512 X 12)

  i.e., 512 words of 12 bits each

  [512=2^9]
- The CPU sends 15-bit address to the cache.
  - If there is a hit CPU get 12-bit data from cache.
  - Otherwise CPU reads the data from the main memory and is transferred to cache i.e., mapping.

1. Associative Mapping: -

Augmented Register	
Address	Data
01000	3450
21777	1234
…….	…….
…….	…….

00000	1220
	…….
	…….
00777	2340
01000	3450

…….	…….
…….	…….
…….	…….
…….	…….

Cache Memory

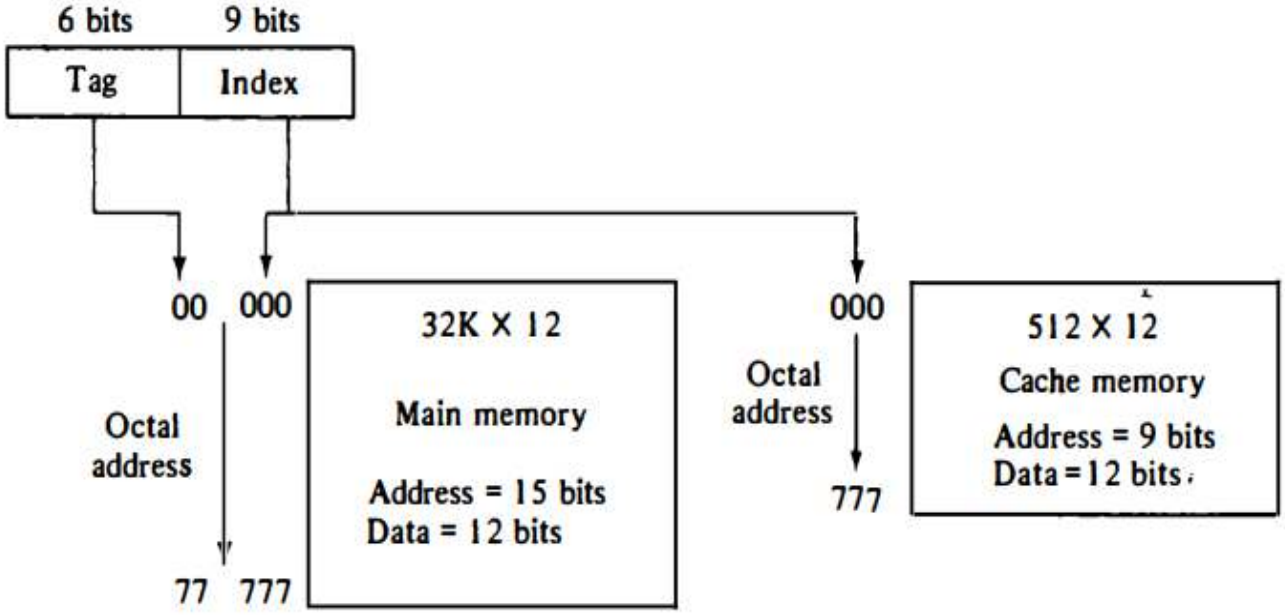
01777	4560
02000	5670
02777	6710

Main Memory

- The fastest and most flexible cache organisation is an associative memory.
- It stores both the address and the data of the memory word.
- The address value of 15 bits is shows as 5-digit octal number and data of 12 bits as 4-digit octal number.
- The CPU address of 15 bit is placed in augmented register (AR) and associative memory, is searched for a matching address.
    - If a match occurs, the corresponding 12-bit data is read and sent to the CPU.
    - Otherwise the main memory is accessed and the address data pair is transferred to the cache.
- If the cache is full, an old address data pair is replaced for new pair.
- The simple algorithm is First In First Out (FIFO).

2. Direct Mapping: -

Figure: Addressing relationships between main and cache memories.



- The CPU address of 15 bits is divided into 2 fields.
- The 9 least significant bits form the ***index*** field and the remaining 6 bits form the ***tag*** field.
- ***The number of bits in the index field = number of address bit required to access cache memory.***
- The internal organisation of the word in cache memory is as shown in the figure.

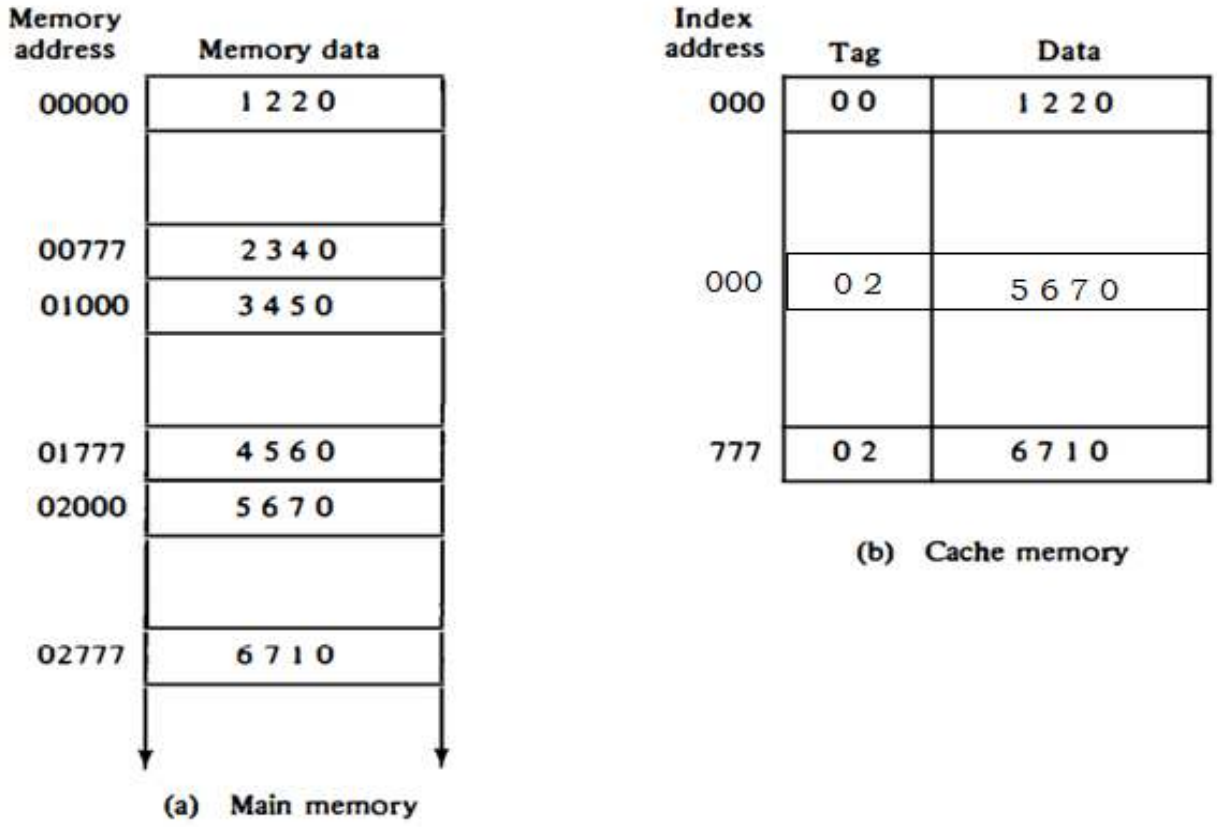


Figure: Direct mapping cache organization.

- Each word in the cache consists of data word and its associated tag.
- When the CPU generate a memory request, the index field is used to access the cache.
    - The tag field of the CPU address is then compared with the tag in the cache.
    - If the two tag matches, there is a hit and the required data word is in cache.
    - If no match occurs, then there is a miss and the required data word read from the memory.
- The ***disadvantage*** of direct mapping is that the hit ratio can drop considerably. i.e., if two or more words whose addresses have the same index and different tags are access repeatedly.

3. Set associative mapping: -

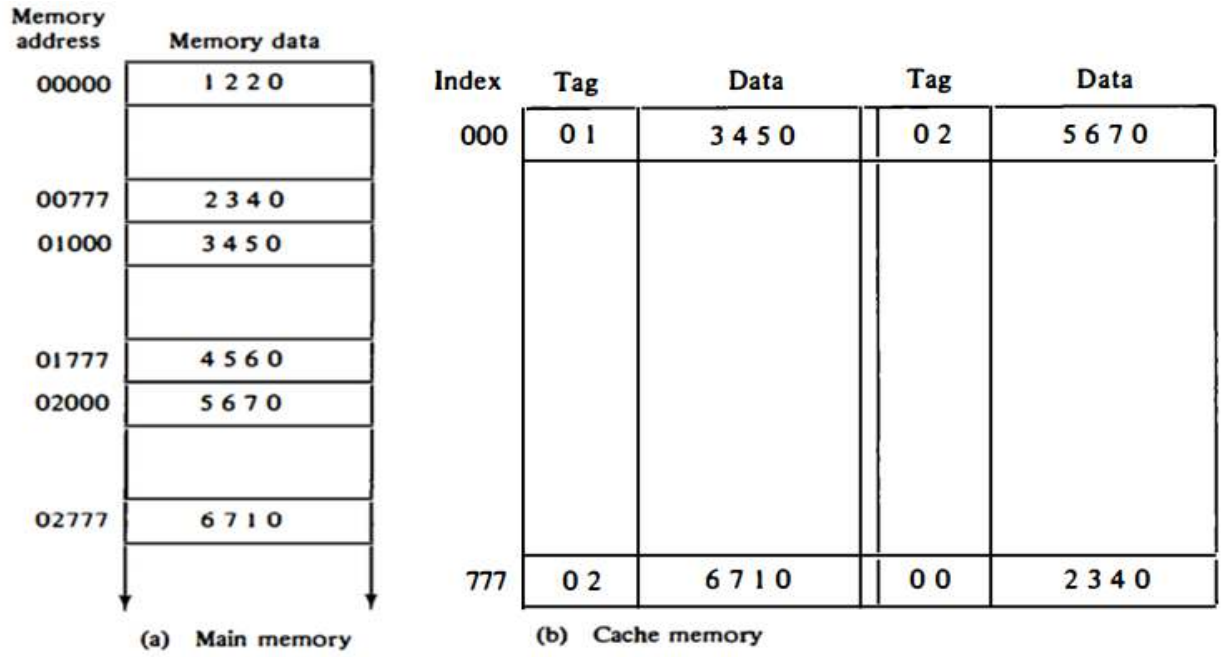


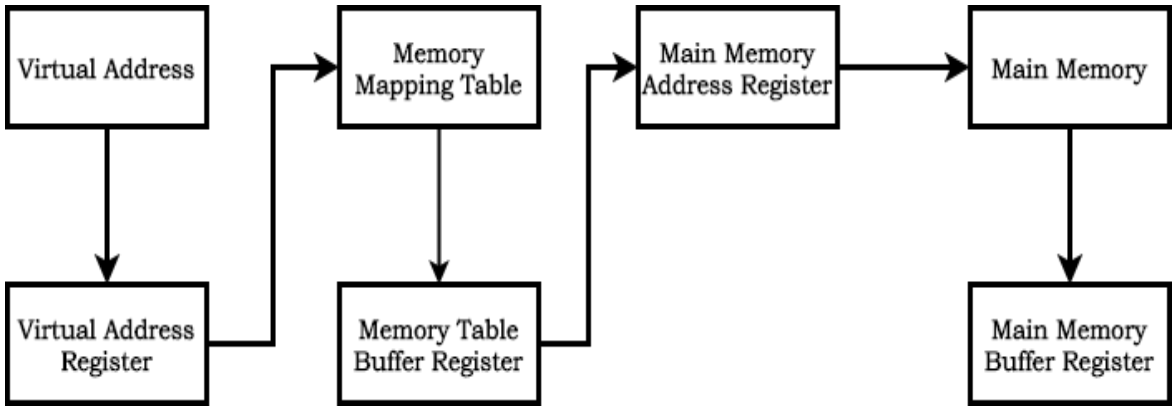
Figure: Two-way set-associative mapping cache.

- Here each word of cache can store 2 or more words of address under the same index address.
- Each data word is stored together with its tag and number of tag data item in one word of cache is said to form a ***set.***
- Here each index address refers to 2 data word and their associated tags.
    - The words stored at the address 10000 and 02000 of the main memory stored in cache at index address 000.
- When the CPU generates a memory request, the index value is used to access cache.
    - The tag field is then compared with both tags in the cache to determine if a match occur or not.
- When a miss occurs in a set associative cache and the set is full, it is necessary to replace one of the tag-data pair with a new value.
- The most common replacement algorithms are:
    1. First In First Out (FIFO)
    2. Random Replacement

3. Least Recently Used (LRU)

VIRTUAL MEMORY

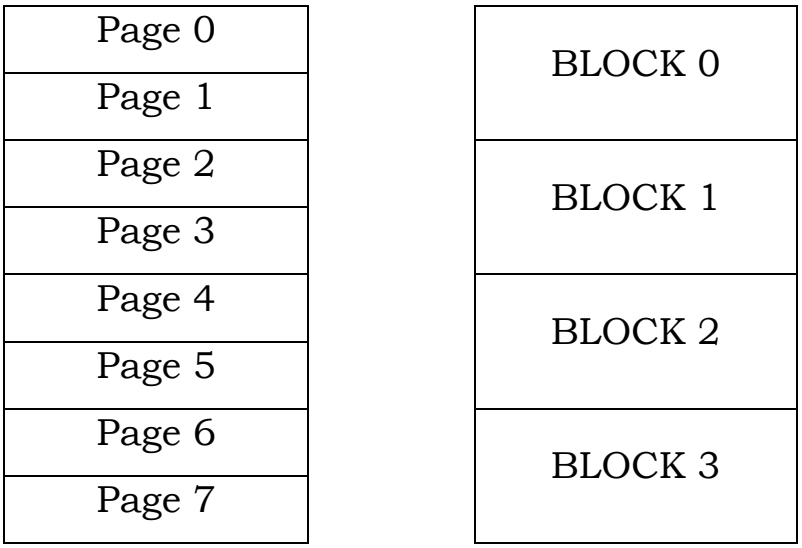
- Virtual Memory is used to give programmer an illusion that they have a very large memory equal to the totality of auxiliary memory.
- A virtual memory system provides a mechanism for translating program generated address into correct main memory locations.
- The translation or mapping is handled by means of a table called ***mapping table.***
- The address used by programmer is called ***virtual address*** and set of such address is called ***address space.***
- An address in main memory is called a location or ***physical address*** and set of such address is called ***memory space.***
- The mapping is a dynamic operation which means that every virtual address is translated immediately as a word referenced by the CPU (Physical address).
- Memory table for mapping a virtual address us shown in the figure.



Address mapping using pages:

- The physical memory is broken down into group of equal size called ***blocks***.
- The term ***page*** refers to group of address space of the same size.
- ***A page refers to the organization of address space while a block refers to the organization of memory space***.

- The term ***page frame*** is sometimes used to denote a block.
- Consider a computer with address space of 8K and Memory space of 4K. If we split each into a group of 1K words. W obtain 8 pages and 4 blocks as shown in figure.



Address Space N=8K=2^{13} Memory Space M=4K=2^{12}

Figure: Address Space and Memory Space into group of 1K words

- Each virtual address is represented by 2 numbers – a page number address and a line number.
- In a computer with $2^{\mathbf{p}}$ words/page, p bits for line address and remaining higher order bits for page number.

Example: Here Virtual Memory Capacity 8K.

- 1K/page
  2^{10}/page
  Here p=10 (line number bits)
- 8K=8 X 1024=2^{13}
  Total 13 bits
  10 bits for line number
  ∴ remaining 3 bits (higher order) for page no.
- In memory,4K (2^{12}). 1K/Block

Here also line number,10 bits

Block number =2bits

∴ Mapping is required from a page number to a block number (3 bits to 2 bits)

- The organization of memory mapping table in a page system is shown in figure.

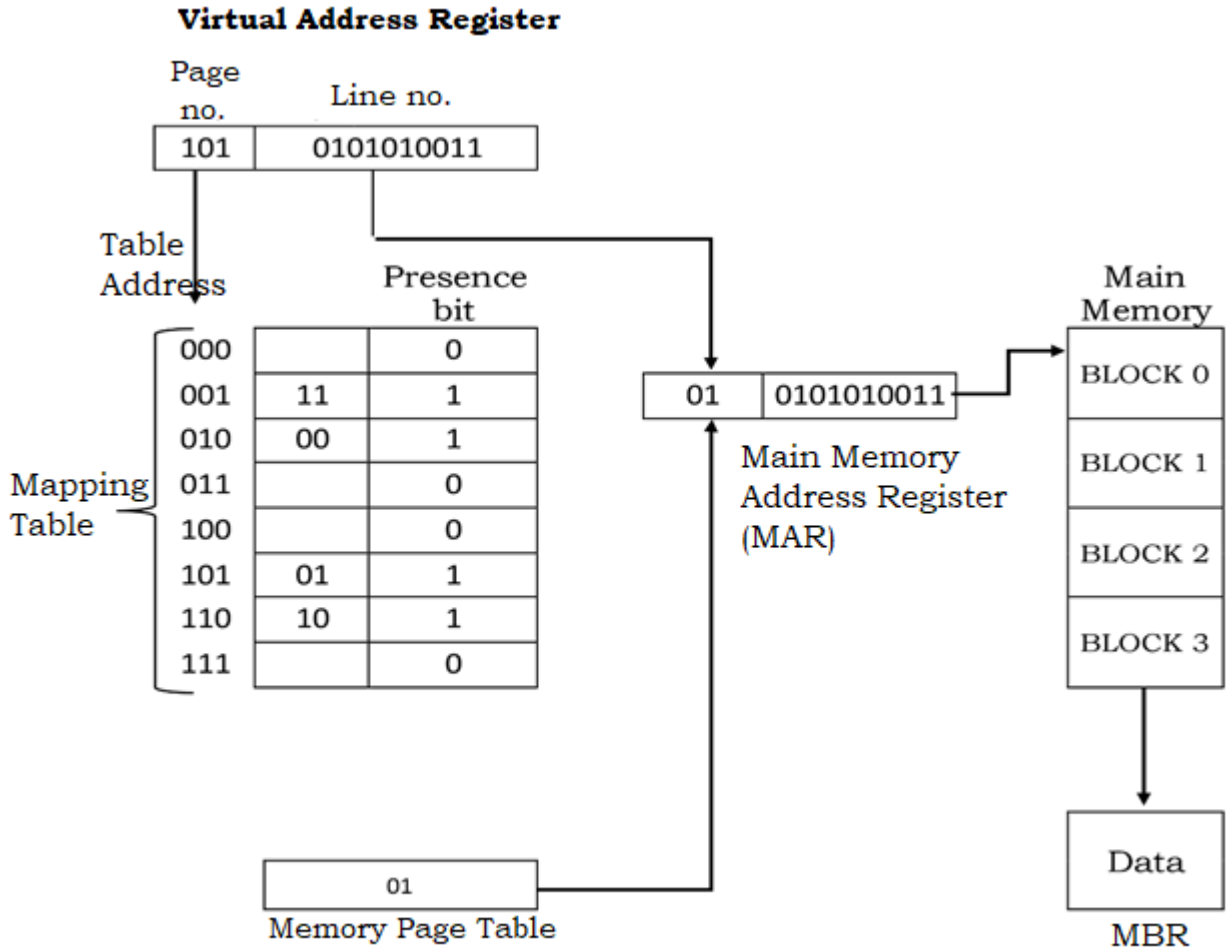


Figure:Memory Table in a Paged System

- The memory page table consist of 8 words for each page (1K/page).
- The address in the page table denotes the page number and the content of the word given the block number where the page is stored in main memory.
- The table shows that pages 1,2,5 and 6 are now available in main memory in block 3,0,1,2 respectively.

- A presence bit in each location indicate whether the page has been transferred from auxiliary memory to main memory. A zero in the presence bit indicate that the corresponding page is not available in the main memory.
- The CPU refers a word in memory with a virtual address of 13 bits.
- The 3 higher order bits of the virtual address specify a page number and also an address for the memory page table.
- If the presence bit is 1, then the corresponding block number is transferred to the 2 higher order bits of MAR. If presence bit is '0' it signifies that the word does not reside in main memory.
- So, the required page is fetched from auxiliary memory and place it into the main memory.

Page Replacement: -

- A virtual memory system is a combination of hardware & software techniques.
- The software system handles all the operation for the efficient utilization of memory space.
- It must include
    a) Which page is to be removed to make space?
    b) Where the page is to be placed and when?
- If the referenced page is not the main memory, it is still in auxiliary memory. This condition is called ***page fault.***
- If the main memory is full, it could be necessary to remove a page from a memory block to make space for the new page.
- For this purpose, replacement algorithms are used.
- Two of them are
    1. **First In First Out (FIFO)**
    2. **Least Recently Used (LRU)**

FIFO

- This algorithm first selects the page first gain each time a page is loaded into memory. Its id number is pushed in a FIFO stack.
- When a new page must be loaded, the page to be removed is easily determined.
- **Advantage**: Easy to implement.
- **Disadvantage:** Sometimes pages are removed and loaded from memory too frequently.

LRU

- This algorithm can be implemented by associating a ***counter*** with every page that is in main memory.
- When a page is referenced, its associated counter is set to zero.
- At fixed interval the counters are incremented by one.
- The least recently used page is the page with highest count.
- The counters are called ***aging registers***.
- **Advantage**: Better algorithm than FIFO.
- **Disadvantage**: Difficulty to implement.

.....END.....