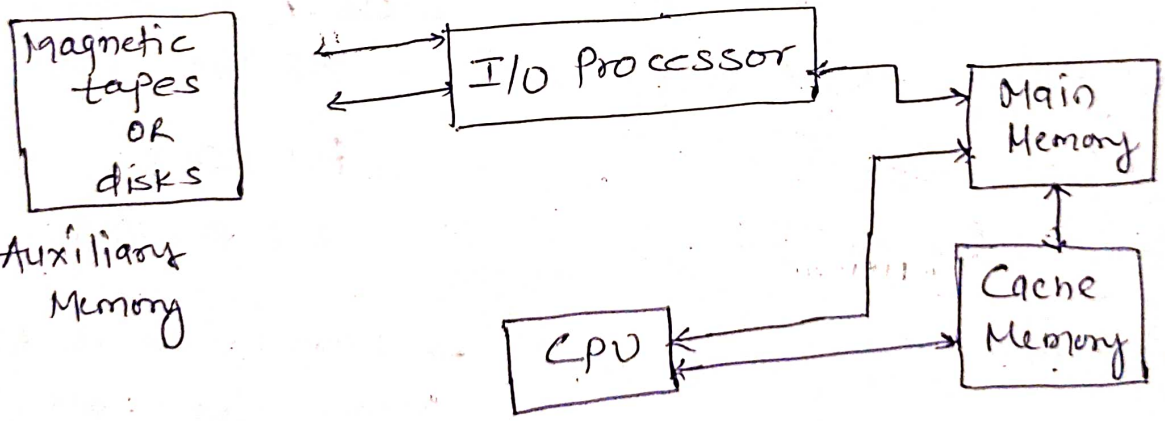
UNIT-5

Memory Hierarchy

- The memory Unit that directly communicates with ~~memory~~ CPU is called main memory / Primary memory.
- Devices that provide backup storage are called secondary memory / Auxiliary memory.
- Magnetic Disk, or taps are most common auxiliary memory.



Cache memory: A special very high speed memory called a cache.

- It is B/w the CPU & Main memory
- This is used to compensate the speed difference B/w the main memory & CPU.

→ Cache Memory hold the address of the frequently used data in Main memory

→ The purpose of memory hierarchy is to obtain the highest possible average access speed while minimizing the total cost of the entire memory system.

Main memory

- It is the central storage in the computer system.
- main memory is based on semiconductor integrated circuits

main memory

RAM

Random Access memory

STATIC RAM - Made up of flop flops, that store Binary information
- volatile

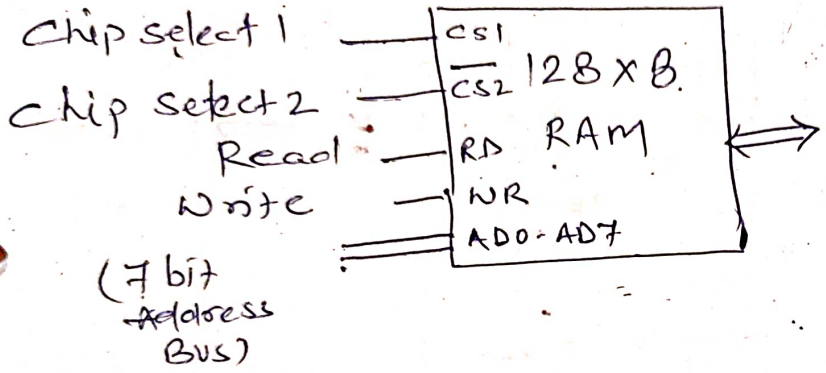
Dynamic RAM - Made up of Capacitors that Accumulates charge to store information
- volatile
- Need to refresh after Particular interval

ROM

Read only memory

Types
- PROM
- EPROM
- EEPROM

→ ROM is need to store the initial Program called Boot strap loader
→ to start computer POST is run to check the Hardware
→ POST → Program on self test.



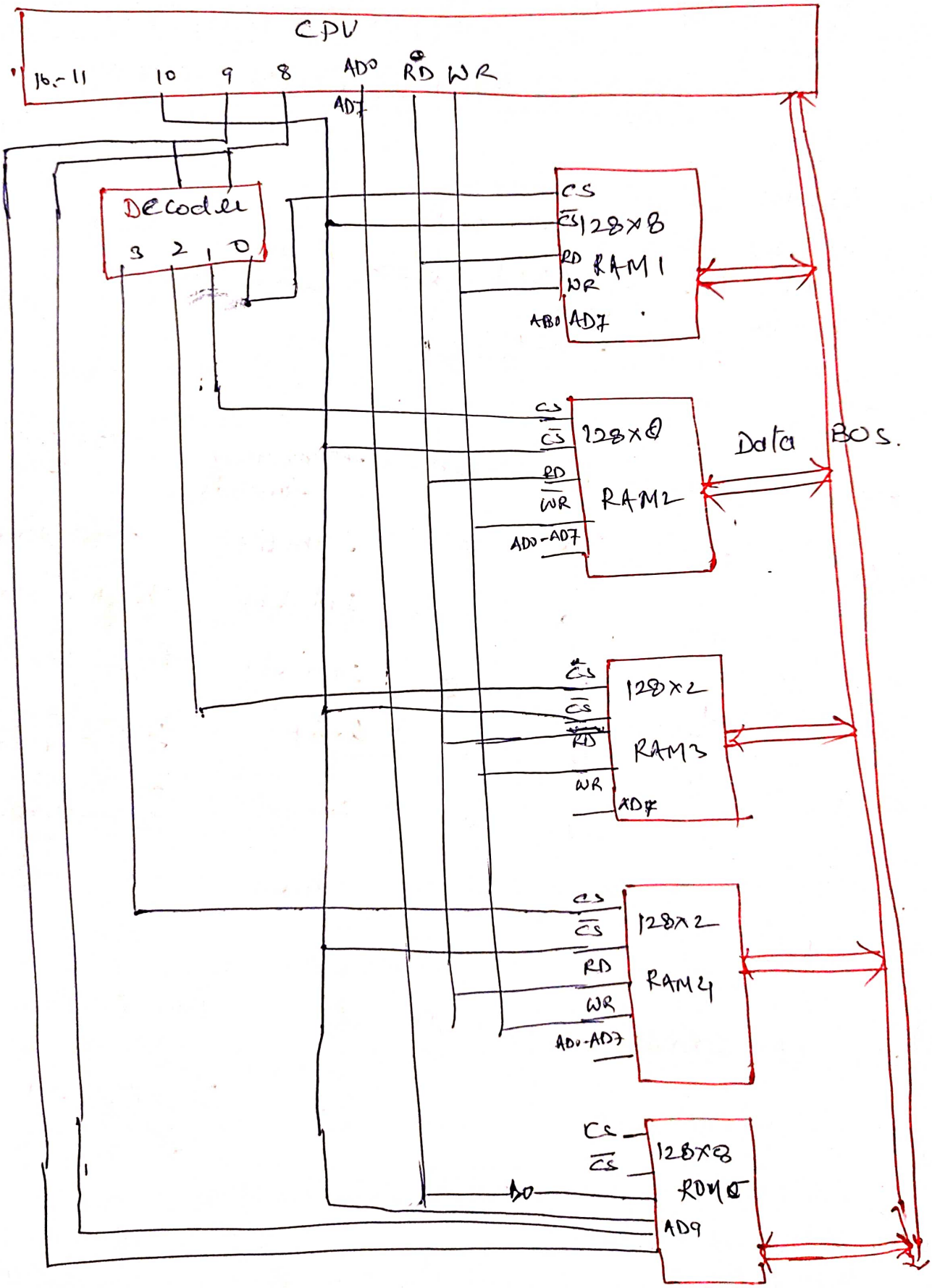
CS1	CS2	RD	WR	memory function	State of Data bus
0	0	x	x	Inhibit	high impedance
0	1	x	x	Inhibit	high impedance
1	0	0	0	Inhibit	high impedance
1	0	0	1	Write	Input Data to RAM
1	0	1	x	Read	output data from RAM
1	1	x	x	Inhibit	high impedance

No. of Addline is calculate $\log_2(\text{size})$

$= \log_2(128) = 7.$

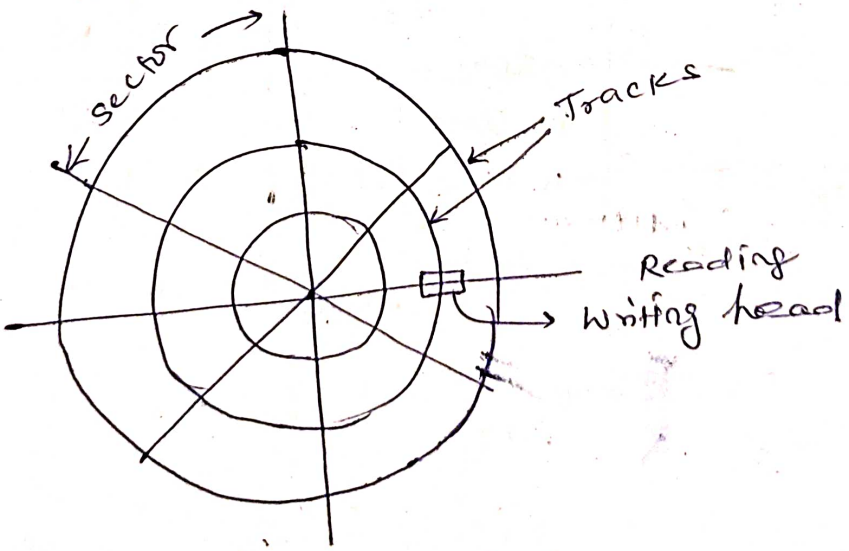
~~Auxiliary Memory~~

Memory Connection to CPU



Auxiliary Memory:-

- Most common memory used in computer system is magnetic disk & tapes
- The Average time required to reach a storage location in memory & obtain its contents is called the acess time.
- Bits are recorded as magnetic spot on the surface as it passes a stationary mechanism called a write head.

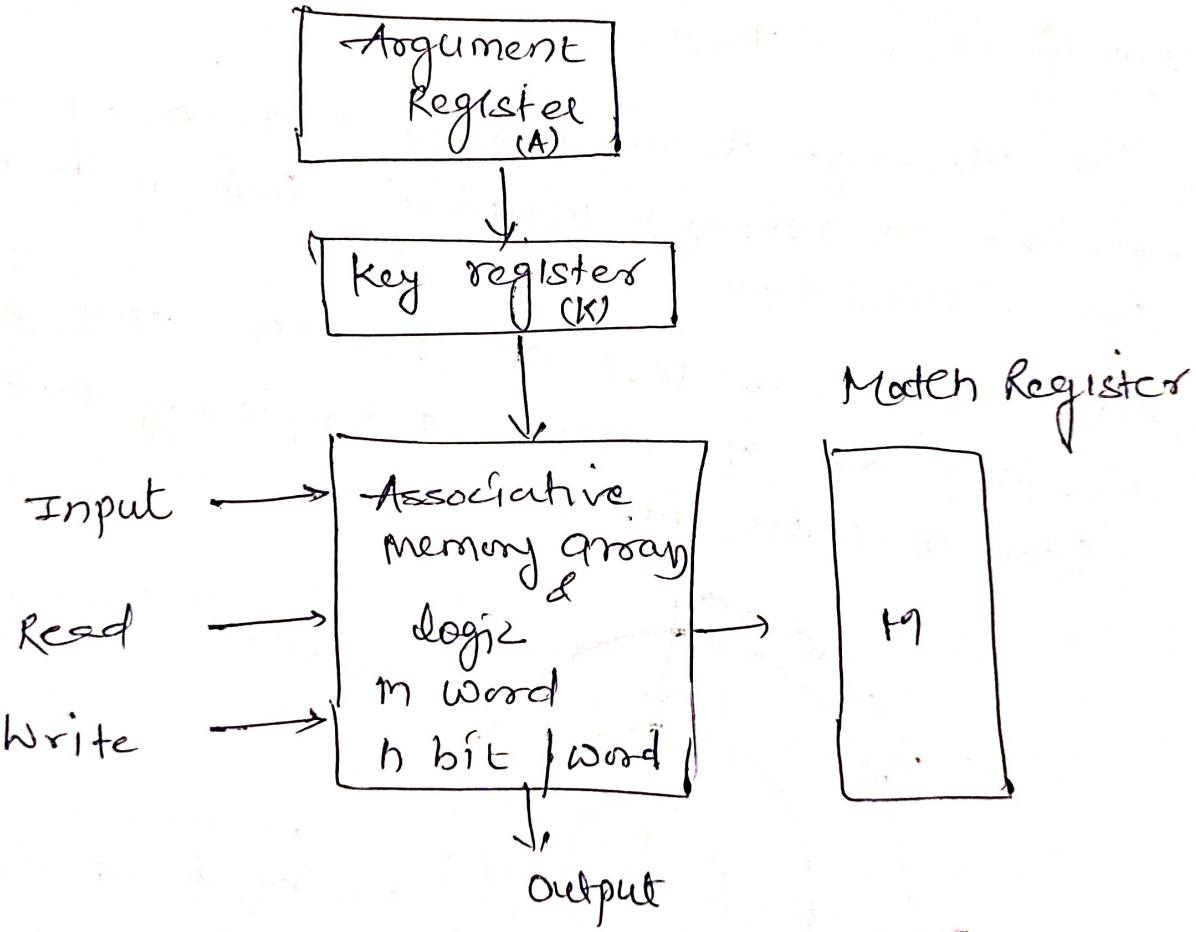


Magnetic disk.

- Disks paormanently attached to the unit assembly and cannot be removed by occasional user are called Hard disk. & Removeable is called floppy disk

Associative Memory

- A memory unit accessed by content is called the Associative memory or content addressable memory
- This memory is accessed simultaneously and in parallel on the basis of data content rather than specific address or location.



A	101	1111 00	
K	111	000000	
Word 1	100	1111 00	no match
Word 2	101	000001	Match.

PAGE NO.

Match Logic:

The Match logic for each word can be derived from the comparision algorithm for two binary numbers.

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

Cache Memory:-

- Analysis of a large number of typical programs has shown that the references to memory at any given interval of time tend to be ~~confused~~ confined with in a few localized memory Area. This is known as locality of Reference
- The performance of cache memory is frequently measured in terms of a quantity of called hit ratio
- When CPU refers to memory and find the word in cache, it is said to be hit. if the word is not found in the cache it is in main memory. it is counted as miss
- The ratio of the number of hits devided by the total CPU references to memory is the hit ratio.

The Transferring of data from main memory to Cache memory is refered to as mapping process.

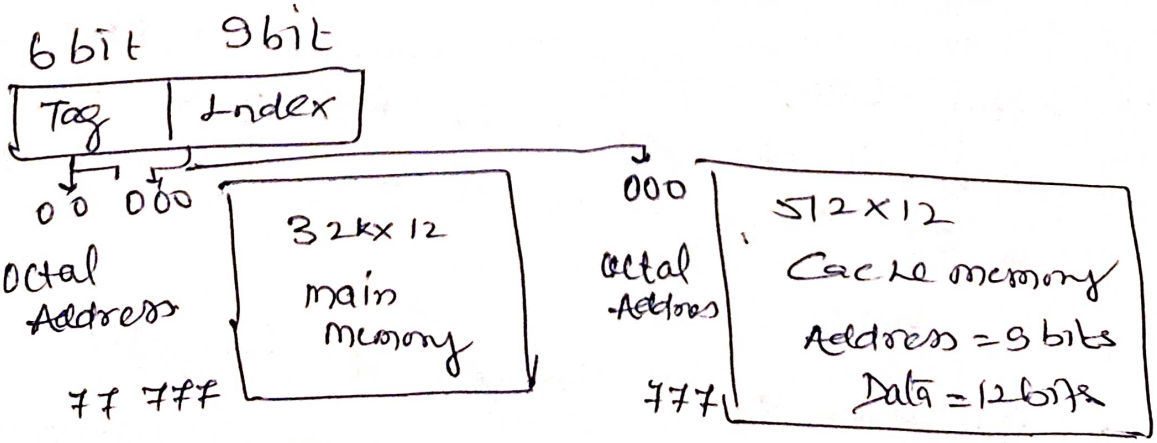
1. Associative mapping
2. Direct mapping
3. Set Associative mapping.

Associative mapping

- The fastest & flexible cache organization uses an associative memory.
- The Associative memory stores the address & Content of the memory word.

Direct mapping

- The Direct memory mapping cache organization uses the n bit address to access the main memory and the k bit to Index to access the Cache
- When CPU generates the ~~tag~~ ~~bit~~ memory request the index field is used to for the address to access the Cache
-



Set Associative Mapping

It is an improvement over the Direct memory organization in that each word of cache can store two or more words of memory under the same Index address

- In general a set Associative cache of size K will accomandate K words of main memory in each word of cache

- Common Replacement Algorithms are
  1. first in first out
  2. least recently used
  3. Random replacement
  4. Most recently used

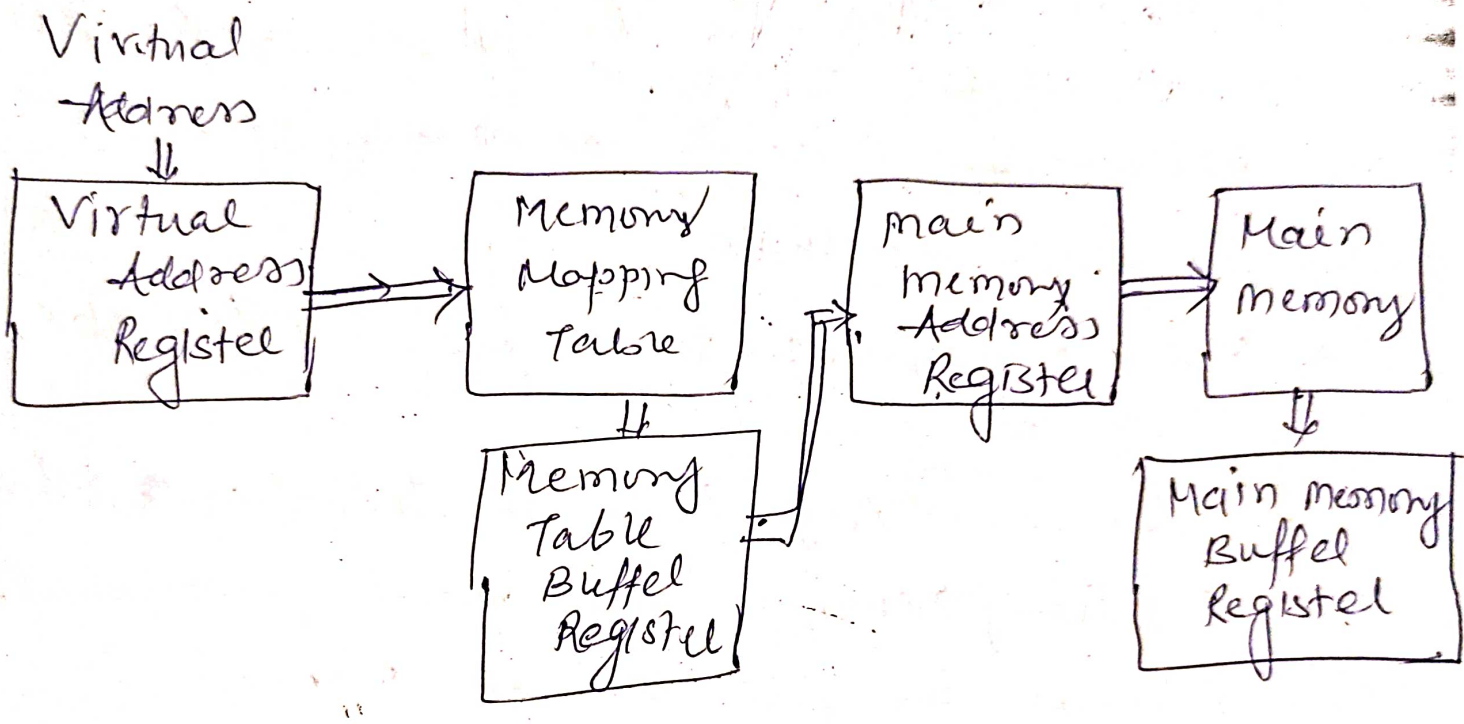
- The cache is initilised by clearing all the bits to 0. The valid bit of a perticular cache word is set to be 1. the first time this word is loaded from main memory and stays unless the cache has to be initilised again.

Virtual Memory :→

- Virtual Memory is a concept used in some large Computer system that permits the user to construct Programs Through a large memory space were available, equal to the total of Auxiliary memory.
- Each address that is used to give programmers the illusion that they have a very large memory at their disposal. even though the computer Actually has a small memory
- Each Address that is reffered by the CPU Goes ~~to the~~ through an address mapping from so called Virtual Address

Address space & Memory space

- The Address used by the Programmer will be called Virtual Address and the set of such addresses the address space
- An address in memory is called a location or physical address
- the set of such location is called memory space
- Address space N and the Memory space by M, We have than have for this example N = 1024K and M = 32K.



Address Mapping Using Pages :-

- Physical Memory is Booken into group of equal size called Block.
- The term Page refer to the group of address space of the Same size
- The page frame is some time used to denote a block.

Associative Memory Page table

- a system with n Pages and m block would require a memory page table of n locations of which upto m blocks will be marked with Block

numbers and all other will be empty.

Page Replacement

- A virtual memory system is a combination of both S/W & H/W.
- The memory management system software system handles all the soft ware operations for the efficient utilisation of memory space
  - It must include
    - * Main memory ought to be removed to make room for a new page
    - * When new page is to be transferred from auxiliary memory to main memory
    - * Where the page is to be placed in the main memory

PAGE NO.

Multiprocessors

Characteristics of multiprocessors

- The MIMD systems are known as multiprocessors
- Multiprocessors can mean either a central processing unit or an input-output processor.
- The Benefit of multiprocessor organisation is an improved system performance
- Multiple independent jobs can be made to operate parallel.
- A single job can be partioned into multiple parallel task.
- A multiprocessor system with common shared memory as a shared or tightly coupled multiprocessor
- Each system has it's own private local memory. It is known as distributed or loosly coupled system.

Interconnection Structures

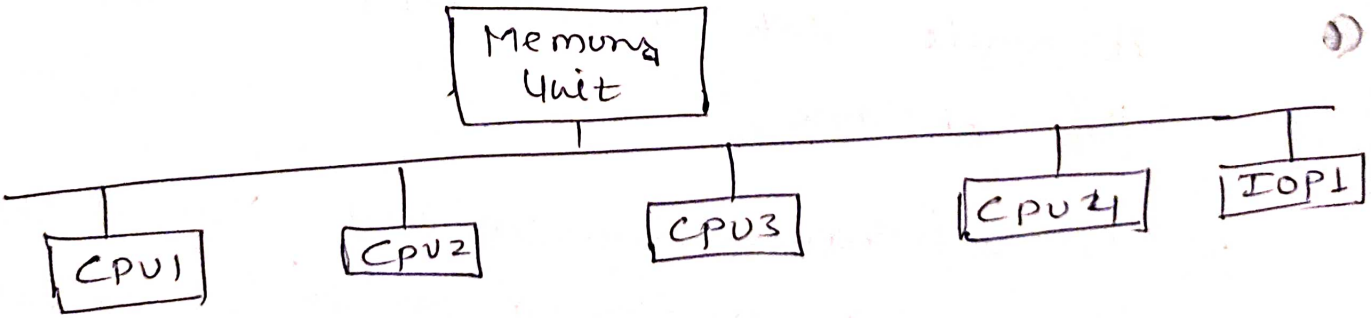
The components that form a multiprocessor system are CPUs. IOP's connected to input output devices and memory unit that may be partioned into a number of seperate modules.

- Several physical forms available for establishing an interconnection network.

1. Time shared common bus.
2. Multiport memory
3. Crossbar Switch.
4. Multistage Switching network.
5. Hypel Cube system.

① Time shared Common Bus

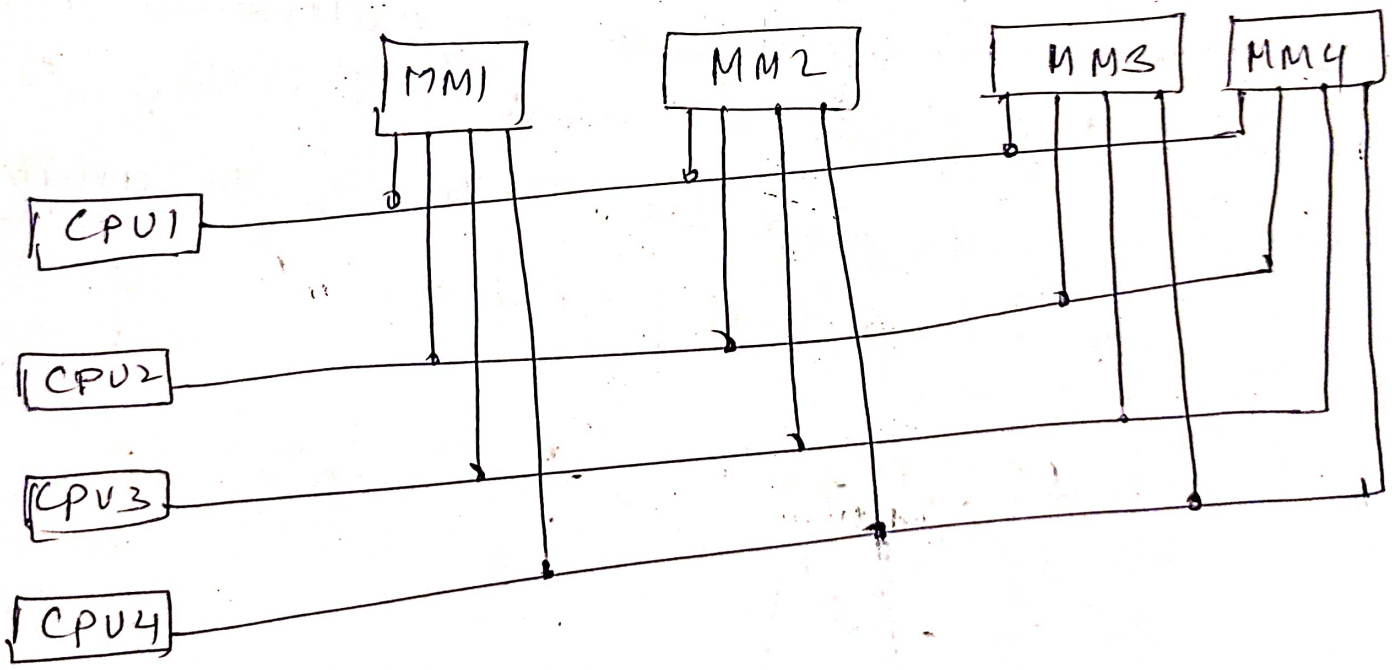
- Number of processors connected through a common bus.



- Only one processor can communicate with the shared memory and other common resources through the system bus at any given time.

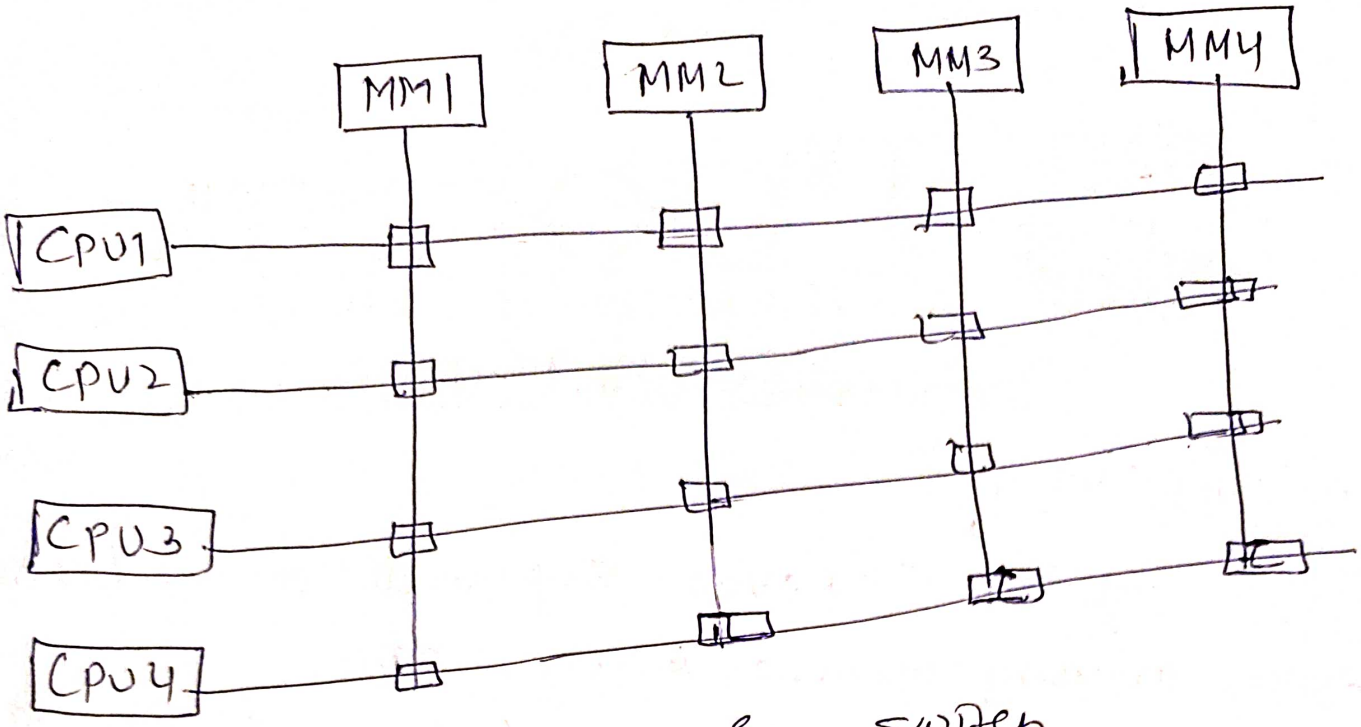
Multiport Memory :-

- This system employes seperate buses between each memory module & each CPU.
- A processor Bus consist of the address, data & control bus to communicate with the memory.



Crossbar switch

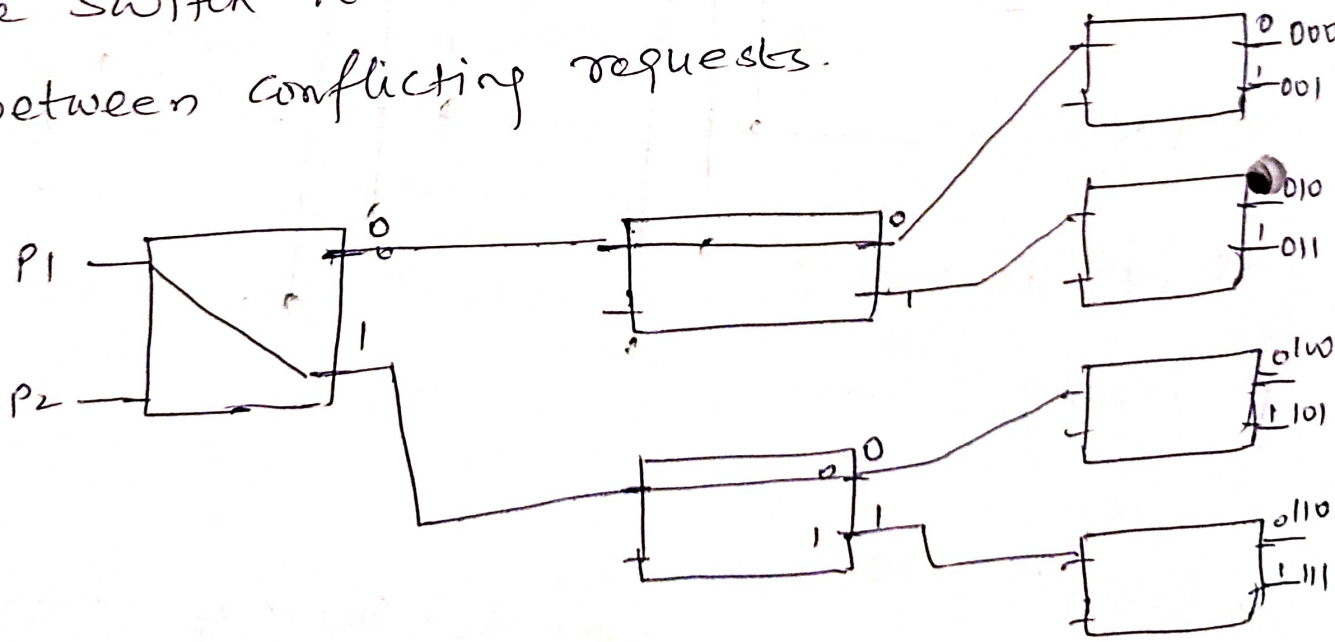
The crossbar switch organization consits of a number of crosss points that are placed at intersections b/w the processor Buses & memory module paths.



Cross Bar switch

Multistage switching Network –

- A Basic Component of a multistage network is a two 2input, 2 output interchange switch.
- The switch has the capability to arbitrate between conflicting requests.



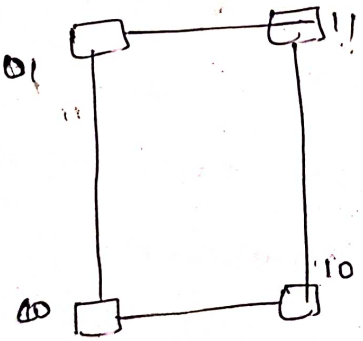
Binary Tree with 2x2 switches

PAGE NO.

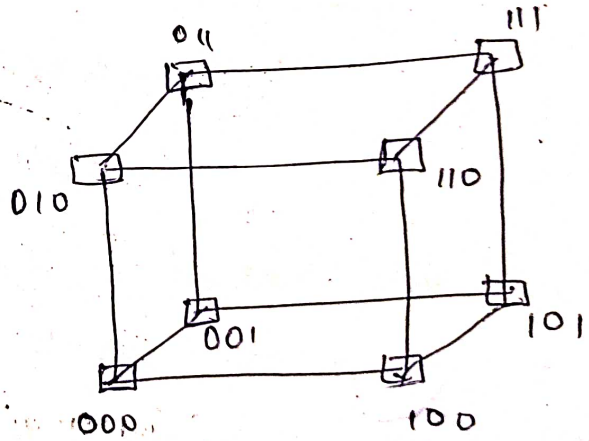
Hypercube Interconnection

- The hypercube or binary n cube multiprocessor structure is a loosely coupled system composed of $N = 2^n$ processors interconnected in a n-dimensional binary cube.
- Each Processor form a Node of a cube.

One Cube



Two Cube



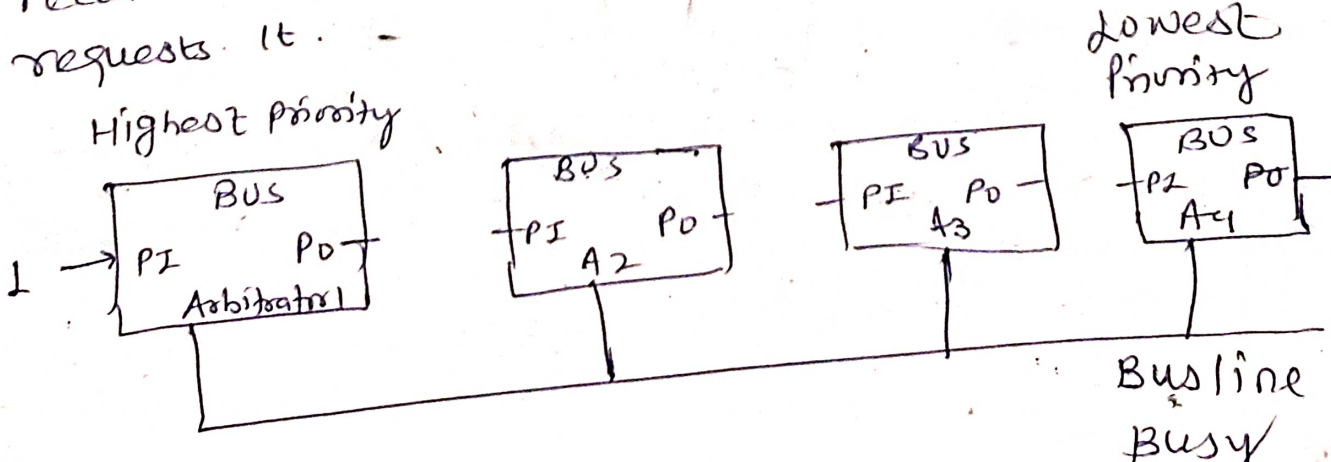
Three Cube

Inter processor ~~Abbri~~ Arbitration:-

- Computer system contains a number of buses at various levels to facilitate the Transfer of Information between components.
- A bus that connects major components in the multiprocessor system such as CPU, IOP's and Memory is called a system bus.
- A system bus typically consist of 100 signal lines. These are devide into three functional groups.
  - ~~Address~~ Bus
  - data bus
  - control signal.
- Data Transfer over the system bus may be synchronous & Asynchronous.
- In synchronous bus, each data item is transferred during the time slice known in advance to ~~the~~ both ~~the~~ sources & destination unit. it is acheived by driving both units from common clock source
- In an Asynchronous bus each data item being Transferred is accompanied by handshaking control signals to indicate when the data are transferred from the source and received by destination

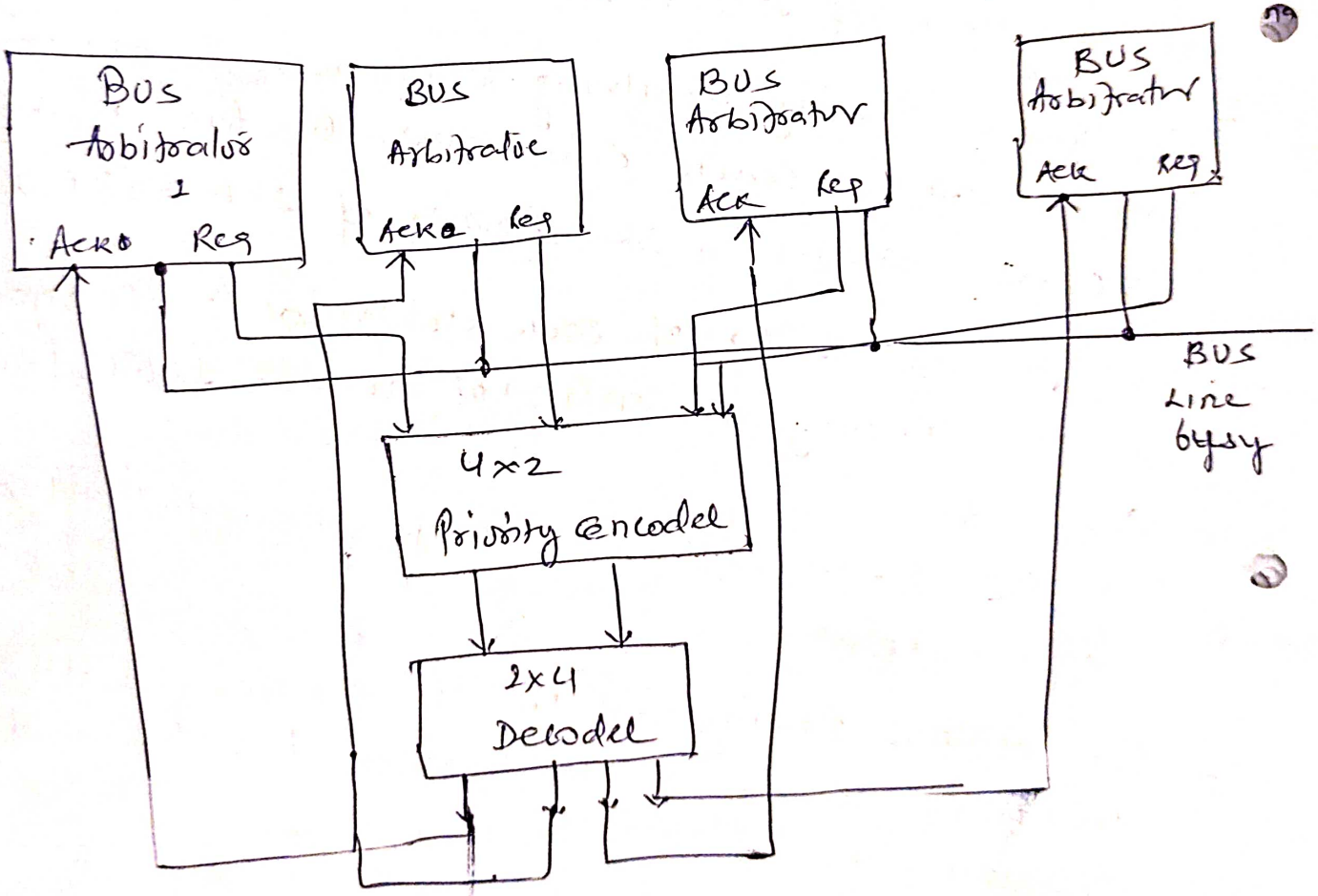
Serial Arbitration Procedure :→

- Arbitration procedures services all processors requests on the basis of established priorities.
- The Serial priority resolving Technique is obtained from a daisy chain connection of bus arbitration circuits similar to the priority of interrupt logic.
- The Priority out (PO) of each arbitrator is connected to the Priority in (PI) of the next lowest priority arbitrator.
- The PI of the highest priority is maintained at a logic 1 value.
- The highest priority unit in the system will always receives access to the system but when it requests it.



Parallel Arbitration logic:

- The Parallel Bus Arbitration technique uses an external priority encoder and a decoder
- Each bus Arbiter in the parallel scheme has a bus request output line and a bus acknowledge input line.
- Each arbiter enables the request line when it's processor is requesting access to the system bus.



Parallel Arbitration

Dynamic Arbitration Algorithm

The two bus Arbitration procedures just described use a static priority algorithm since the priority of each device is fixed by the way it is connected to the bus.

- In the dynamic priority algorithm gives the system the capability for changing the priority of the devices while the system is in operation.
- The time slice Algorithm allocates a fixed length time slice of bus time that is offered sequentially to each processor in Round Robin fashion.
- The Least Recently used (LRU) algorithm gives the highest priority to the requesting device that has not used the bus for the longest interval
- The rotating daisy chain procedure is a dynamic extension of the daisy chain algorithm. In this

There is no central bus controller and the priority line is connected from the priority out of the last device back to the priority in the first device in a closed loop.

Interprocessor Communication & synchronization

- Multiprocessor system must be provided the facility for communicating with each other
  - A communication path can be established through common input output channels.
  - In a shared memory multiprocessor system the most common procedure to set aside a portion of memory to that is accessible to all processors.
  - In addition a shared memory, a multiprocessor system may have other shared resources.
  - for example, a magnetic disk storage unit connected to an IOP may be available to All CPU's.

- Interprocessor synchronization

The instruction set of a multiprocessor contains basic instructions that are used to implement communication and synchronization between co-operating processes.

Cache Coherence

In a shared memory system, all the processors share a common memory. In addition, each processor may have a local memory, part or all of which may be a cache. The compelling region for having seperate caches for each processors is to reduce the average access time in each processor.

The same information may reside in a number of copies in some caches and main memory. To ensure the ability of the system to execute memory operation correctly the mutiple copies must be kept identical. This requirement imposes a cache coherence problem. A memory scheme is coherent if the value returned on a load instruction is always the value given by the least store instruction with same address.

Condition for Incoherence

Cache coherence problem exist in multi processor with private caches because of the need of share writeable data. Read only data can safely be replicated without cache coherence enforcement mechanism.

PAGE NO.

- Communication refers to the exchange of data between different processes.
- Synchronization refers to the special case where the data used to communicate b/w processors is control information.

Synchronization is needed to enforce the correct sequence of processes and to ensure mutually exclusive access to shared writable data.

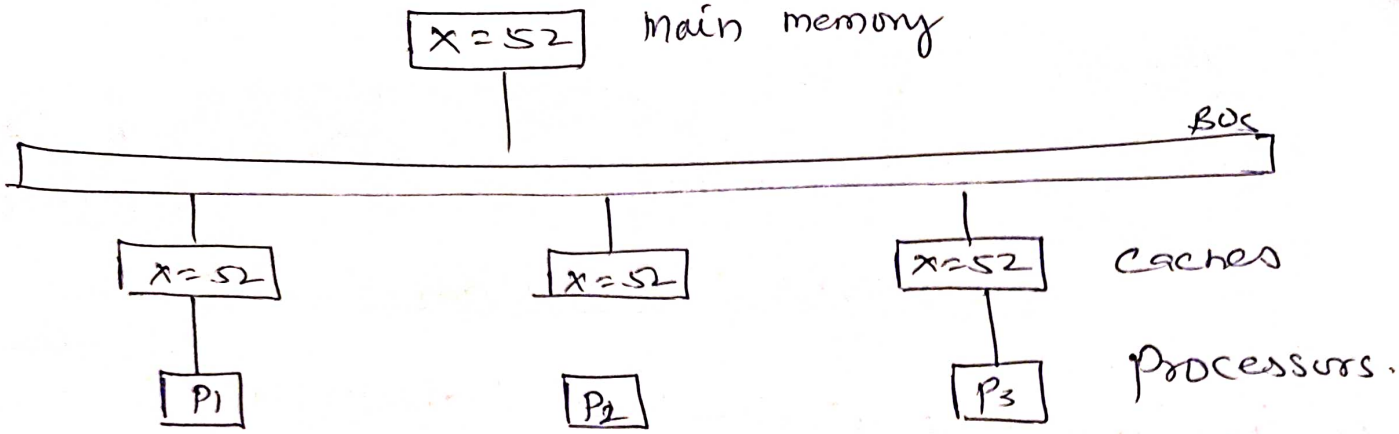
Mutual exclusion with a semaphore

- A properly functioning multiprocessor system must provide a mechanism that will guarantee orderly access to shared memory and other shared resources.
- It is necessary to protect data from being changed simultaneously by two or more processors. This mechanism has been termed as mutual exclusion

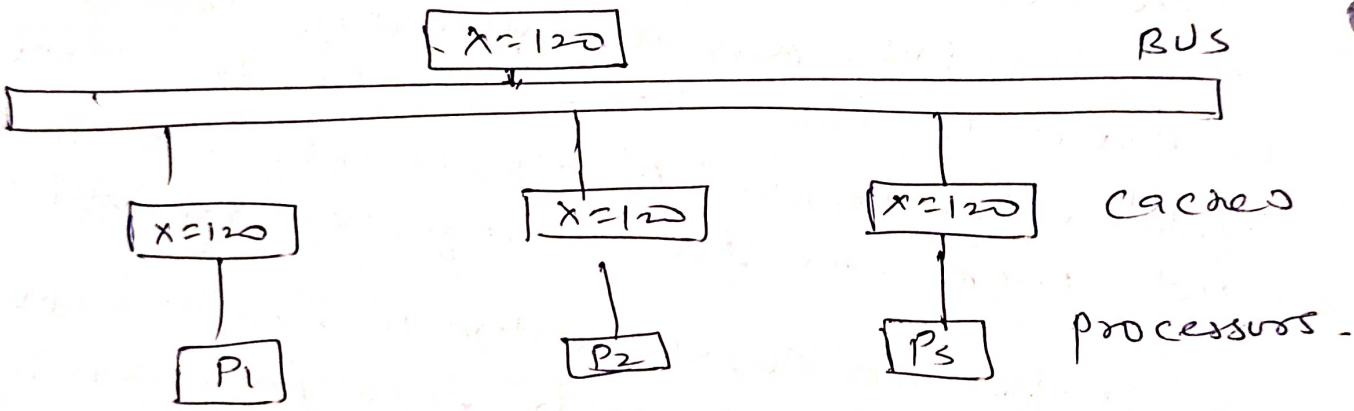
Solution to Cache Coherence Problem:→

* Simple scheme is to disallow private caches for each processor and have a shared cache memory associated with main memory. Every data Access is made to the shared cache. This method violates the principal of closeness of CPU to cache and increases the average memory Access time

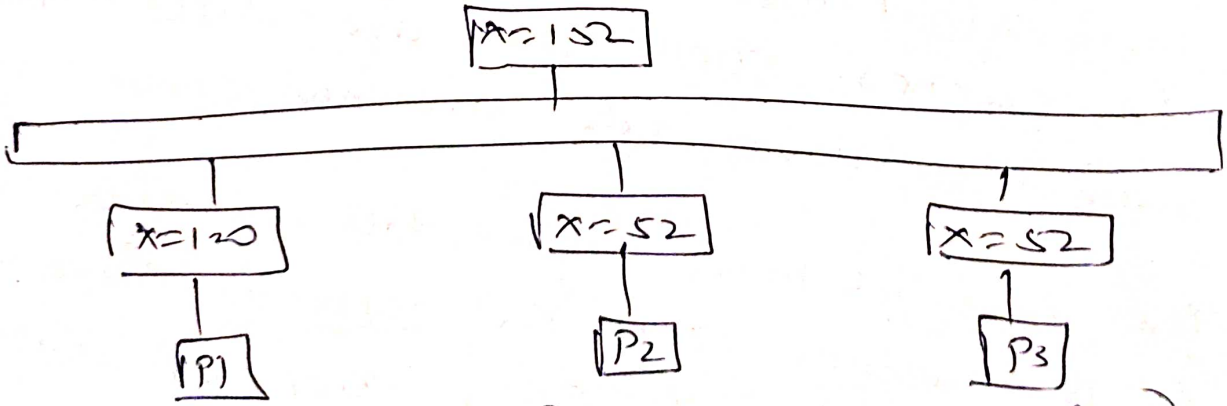
- for performace consideration it is desirable to attach a private cache to each processor. One scheme that has been used to allows only non shared and read only data to be stored in caches. Such items are called Cachable
- Shared Writable data are non cacheable.
- The Compiler must tag a data as either cachable or non cachable and the system hardware makes sure that only cachable data are stored in caches. The non cachable data remain in main memory.



Cache configuration after a load on X.



With write through cache policy



With write back cache policy

End