Interconnection Network

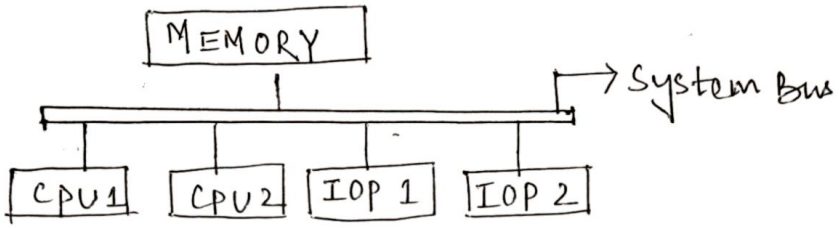
It describes different physical configuration among the components (CPU, IOP, memory) depending on no.oif transmission path available between processor and memory of a sharable memory system (tightly coupled system) or between processor & memory of a distributed system (loosely coupled system)

Types of interconnection network

1. Time-shared common bus
2. Multiport memory
3. Crossbar switch
4. Multistage switching network
5. Hypercube interconnection

1. Time-shared common bus

- A common bus multiprocessor system consists of no.of processors connected through a common path to memory unit.

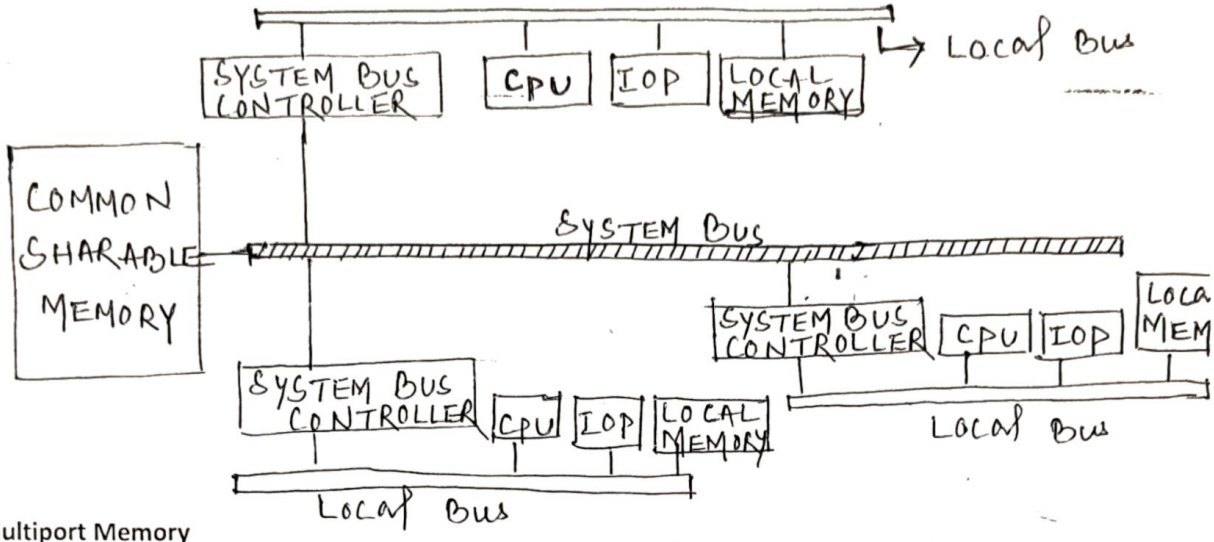


- At a time only one processor has the control of the bus. Therefore at a time only one processor can communicate either with the memory or with other processor.
- To transfer data a command or control signal is issued by CPU to inform the destination unit what operation is to be performed. Then the receiving unit recognizes the address in the bus & responds to the control signals from sender.
- In time-shared bus interconnection system there exist transfer conflicts since one common bus is shared by all processor. So to resolve this problem, incorporate a bus controller that establishes priorities among the requesting units.
- A single common bus system restricted to only one transmission at a time i:e: when one processor is busy in communicating with memory, other processors wait or they may busy with internal operation or may be idle which means waiting for bus. Therefore overall transmission rate of system is limited by speed of single path.

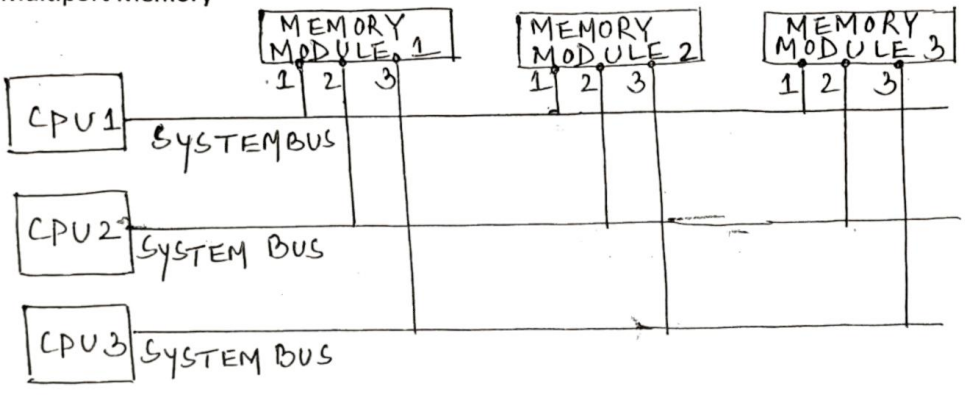
Remedy-

- An economical implementation of a single common bus structure is use no.of local bus & each bus connected to its own local memory, one or more processing unit & a local IOP.
- There is a system bus controller for each local bus that interconnects the local bus with the system bus. And the IO devices connected to the local IOP.

- The memory connected to system bus is a sharable one that is shared by all the processor. At a time only one processor can communicate with the shared memory & other common resources through the system bus. In local memory, a portion or part of it may be designed as cache memory.

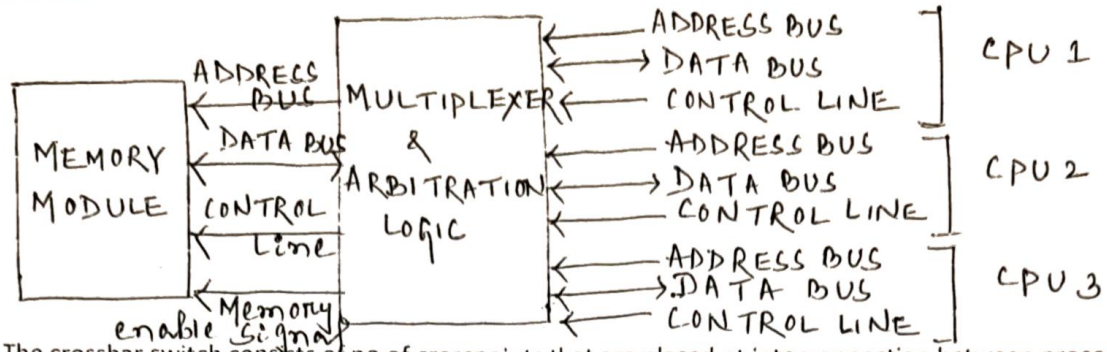


2. Multiport Memory



- A multiport memory system employs separate buses between each memory module & each CPU.
- Each processor bus connected to each memory module & a processor bus consists of the address bus, data bus & control line.
- No.of ports associated with memory module is equals to no.of processor in the system. With each port a single bus is connected to the memory or with each port a single processor is connected to memory.
- Each memory module must have internal control logic to determine which port or processor has currently accessed the memory module at a time.
- A fixed priority is assigned to each port to determine which processor has accessed the memory module *i* at a time.
- Advantage- high data transmission rate because of multiple paths between processor & memory.
- Drawback-very expensive because of control logic in each memory, more connectors, no.of cables. Hence it is suitable for small system only.

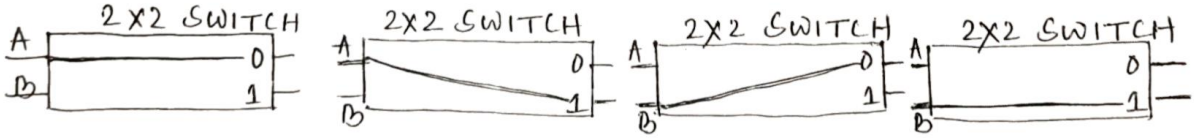
3. Crossbar Switch



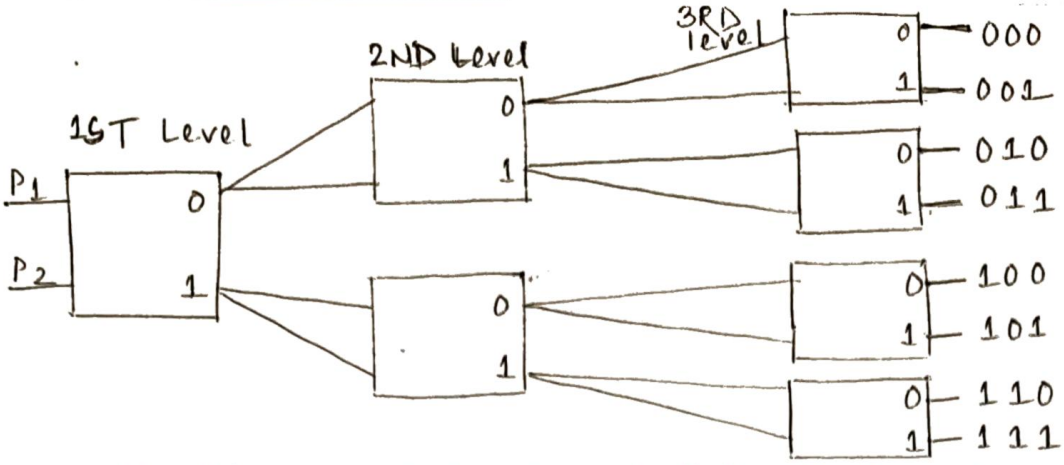
- The crossbar switch consists of no.of crosspoints that are placed at interconnection between processor buses & memory module path.
- The intersection point or crosspoint is a switch that determines the path from a processor to a memory module. Each switch point has control logic to establish the transfer path between a processor & memory.
- Function of control logic-
  1. It examines the address flow over the bus is whether for its connected memory module or not.
  2. It also resolves the problem of access conflicts for the same memory module by using predetermined priorities i: e: only one CPU or processor is selected to access memory when two or more CPU attempts to access the same memory.
- The multiplexer decides or selects the data, address & control lines from any one CPU for communication with the memory module.
- The arbitration logic establishes priority levels to select any one CPU to access the memory.
- The multiplexer is controlled by the binary code that is generated by a priority encoder present within the arbitration logic.
- Advantage- crossbar switch supports simultaneous transmission from all memory modules due to separate path associated with each memory modules.
- Drawback- design complexity is more.

4. Multistage Switching Network

- The basic component of a multistage switching network is a two-input, two-output interchange switch known as 2×2 switch. A 2×2 switch is the basic building block of multistage network that has no.of source & destination.
- No.of output lines in a stage= 2^n, where n= stage number.



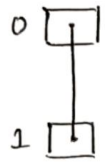
- The switch has the capability of connecting any one of the input terminal to either of the output. The switch also has the capability to arbitrate between conflicting requests.
- E.g. – two processors p1 & p2 connected to 8 memory modules and the memory modules are marked in binary from 000 to 111. Mem.mod0- 000, Mem.mod1-001, Mem.mod2-010 ------- Mem.mod7-111. The two processors are identified as 0 & 1, where P1 as 0 & P2 as 1.



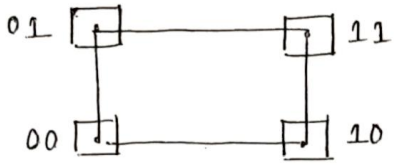
- The path from source to destination is determined by binary bits of destination number
  E.g. destination number- 001
  1st bit (=0) of destination number determine the switch o/p in 1st level
  2nd bit (=0) of destination number determine the switch o/p in 2nd level
  3rd bit (=1) of destination number determine the switch o/p in 3rd level
- Drawback- some request can't be satisfied simultaneously i:e: in this example p1 can be connected to any one of the destination from 000 to 011 and p2 can be connected to any one of the destination from 100 to 111.

5. Hypercube interconnection

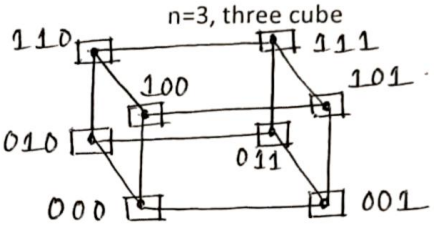
- It is also known as binary n-cube multiprocessor structure.
- E.g. n=1, one cube



n=2, two-cube

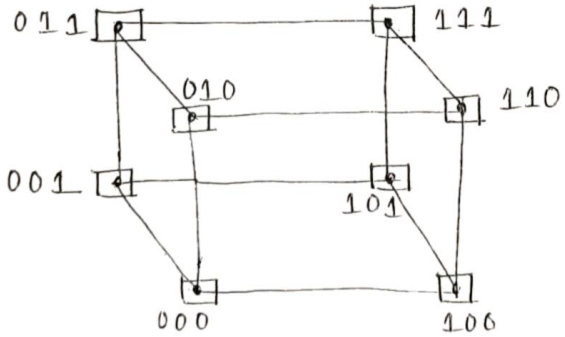


n=3, three cube



- It is a loosely coupled system composed of N=2^n processors interconnected in an n-dimensional binary cube. Each processor is represented as a node.
  E.g. n=1, one-node, 2^n=2^1=2 processor
  n=2, two-node, 2^n=2^2=4 processor
- Each node contains a CPU, local memory & IO interface. Each processor has direct communication path to 'n' other remaining other neighbour processor. These paths are edge of cube.
- Using n-bit binary address, 2^n no.of processor are identified by 2^n distinct address.
- Each processor address differs from the address of its n-neighbour processors by exactly one bit positions.

e.g.



Neighbour of P0 (000) are P1 (001), P2 (010) & P4 (100).
Difference between
P0 (000) & P1 (001) = 1
P0 (000) & P2 (010) = 1
P0 (000) & P4 (100) = 1

- Routing messages through an n-cube structure may take one to 'n' links (minimum) from a source node to its neighbor destination nodes.

  e.g. from node address 000 to node address 001 – one link
  from node address 000 to node address 011 – two links
  from node address 000 to node address 111 – three links

  or, how many minimum no.of links required to route messages from source to destination can be evaluated by XORing source node address with destination node address and no.of 1's in the result of XOR operation gives the value of no.of links.

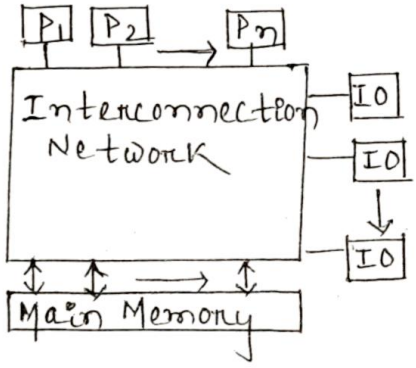
  e.g. 000 (source node address)
  (XOR) 011 (destination node address)

  011→ no.of 1's in 011 is 2, so no.of link=2.

Shared Memory Multiprocessor

SMP

- SMP refers to computer hardware arch. & also operating system behaviour that reflects the arch. An SMP can be defined as a standalone computer system with the following characteristics:-
  1. There can be two or more similar processors of comparable capability.
  2. These processors share main memory & IO devices through bus in such a way that access time of memory or memory access time is approximately same for each processor.
  3. IO devices are also shared by all processors through same channel or different channels.
  4. All processors can perform same function.
  5. SMP based computer system is controlled by integrated operating system that provides interaction between processors & their programs. i:e: the operating system of an SMP schedules processes or threads across all of the processors & there exists a high degree coordination between processors.
- Advantages of SMP over uniprocessing :-
  1. Performance- If the work to be performed by a computer can be organized so that some portion of the work can be done in parallel then a system with multiprocessor will yield great performance than uniprocessor.
  2. Availability- In a SMP because all the processors can perform the same function so the failure of a single processor doesn't halt the machine rather the machine continue to perform at a slow performance or speed.
  3. Incremental growth- An user can enhance the performance of a system by adding additional processor.
  4. Scaling- vendors offer range of products with different price & performance where performance based on characteristics of no.of processors.
- Organization:-

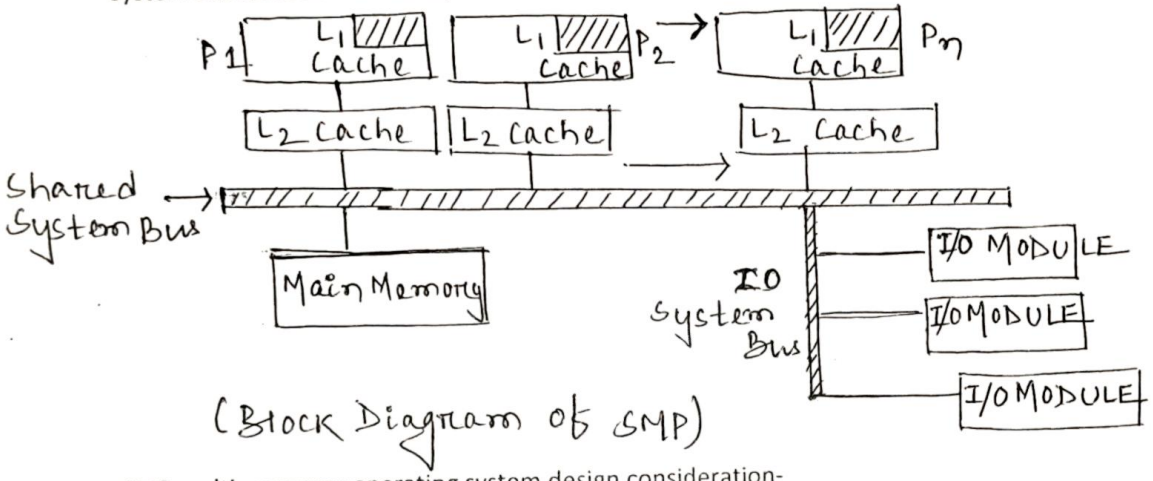


  - There are two or more processors in SMP & each processor is self contained by including CU, ALU, registers & cache memory.
  - Each processor has access to shared main memory & IO devices through interconnection network of any type. The processors can communicate with each other directly or through memory.
  - The main memory of SMP is organized in such a way that multiple simultaneous access to different blocks of main memory can be possible.

➢ In some SMP system there are private main memory & private IO channels for each processor along with shared resources. Interconnection network organization is mostly time-shared bus.

- DMA data transfer in SMP provides the following features:-
  1. Addressing- Addressing bits are required to distinguish or identify modules on the bus to determine the source & destination of data.
  2. Arbitration- The IO modules or DMA controller is provided with arbitrate competing request for bus control using some priority scheme.
  3. Time sharing- Control of system bus is assigned to only one IO module at a time & the other IO modules are locked out or suspended their operation until system bus control is achieved.
- Features of system bus organization:-

  1. Simplicity- addressing, arbitration & time sharing logic of each processor is same as a single processor system.
  2. Flexibility- SMP system can be easily expanded by attaching more processors to the bus.
  3. Reliability- Failure of any attached device never causes failure of the whole system.
  4. Apart from the above 3 features, the performance degradation due to passing of all references through system bus can be improved by attaching each processor with a cache memory that reduces system bus access dramatically.



(Block Diagram of SMP)

- SMP multiprocessor operating system design consideration-

- Drawback of SMP-

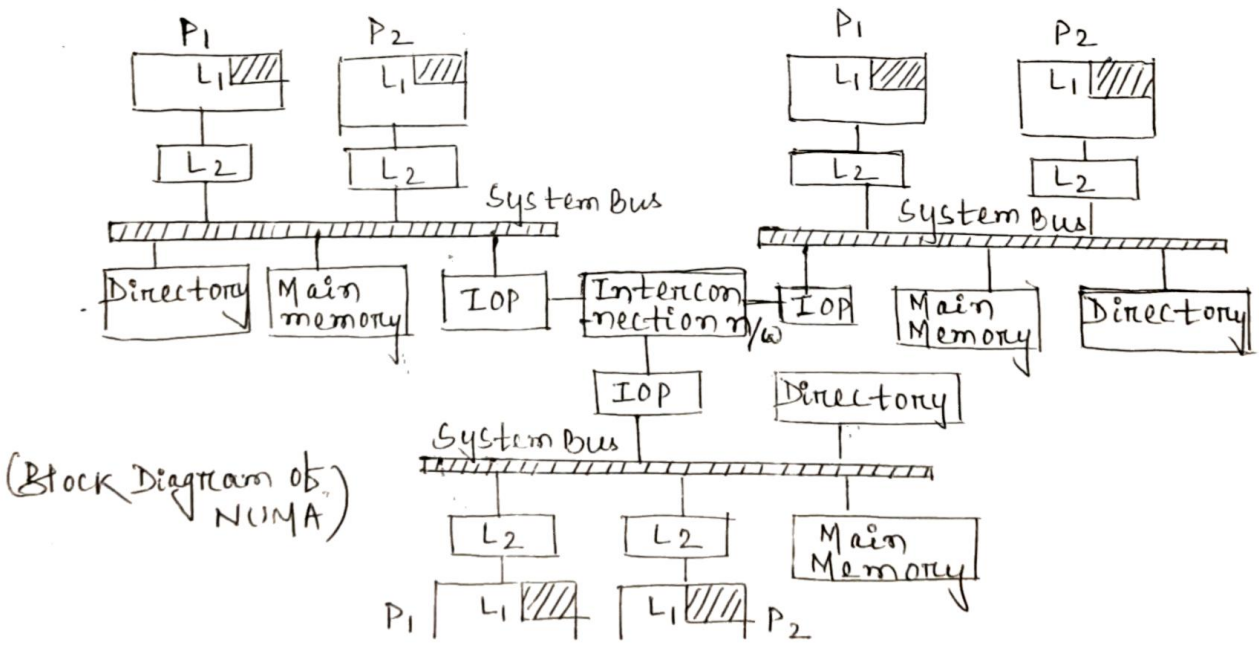
  SMP without cache memory increase bus traffic so use cache memory to reduce bus traffic between processor & main memory. But cache memory of different processor should maintain consistency or to maintain cache coherence among no.of cache memories of various processors by flowing some signals over system bus. So not only references of different processors but also signals for cache coherence are also flowing over system bus. Hence the performance degraded if no.of processors connected in SMP are not limited (16 to 64 processors in SMP). Therefore SMP approaches to small-scale multiprocessing.

NUMA

- Uniform memory access (UMA)- all processors have access to all parts of memory & the memory access time to all the regions of main memory for all the processor is same.
- Non-uniform memory access (NUMA)- all the processors have access to all the regions of main memory but the memory access time of all the processor differs depending on which region of main memory is accessed.
- Cache coherence NUMA (CC-NUMA) - it is a NUMA system where there is coherence or consistency maintained among the caches of various processors.
- NUMA with cache coherency is known as CC-NUMA & NUMA without cache coherence is known as cluster.
- NUMA helps in achieving large scale multiprocessing while using the same function & features of SMP.
- NUMA maintain a transparent system with wide memory while permitting multiple processor nodes, each with its own system bus or internal interconnection system.

Organization-

- There are multiple independent SMP organizations that get interconnected by a interconnection network where each independent SMP organization is known as a node. a node (SMP organization) is basic building block of CC-NUMA.
- Each node (SMP organization) contains multiple processors, each with its own L1 & L2 cache memory (multiple cache memories) along with its own main memory.



- When a processor initiates a memory access, if the requested memory location is not in the processor cache (not in L1 & L2) then the L2 cache memory initiates a fetch operation. And if the desired line is in the local portion of main memory then the line is fetched across local bus. But if the desired line is in a remote portion of main memory then an automatic request is sent out to fetch that line across the interconnection network by delivering the desired line over local bus & then to the requesting cache on that bus. This activity is automatic.
- Each node (SMP) must maintain a directory that gives an indication of the location of various portion of main memory & also gives information about cache memory status.
- Example-
  Content retrieval from a location by a processor of a node (SMP). Suppose a processor P0 of node2 (SMP2) requests for a memory location 225 & this memory location belongs to memory of node1 (SMP1). The following steps are followed-
  1. P0 of node2 send a read request on snoopy bus of node2 for location 225.
  2. The directory on node 2 sees the request & recognizes that the location is in node1 not in node2.
  3. Node2's directory send request to node1's directory. Node1's directory acts as a surrogate of P0 (of node2) & send request to its local main memory for location 225.
  4. Now node1's main memory respond to the request by placing the requested data on the bus. Then node1's directory picks up the requested data from the bus & send it to node2's directory.
  5. Node2's directory now place the data of location 225 on bus that is received & placed in the cache memory of P0 (of node2). Then from cache memory it is delivered to processor P0.

Cache coherence in CC-NUMA-

- If a search line is not present in the local cache then the local directory keeps a record that some remote cache has a copy of the line containing search word or search location.
- There should be a cooperative protocol to take care of modifications, i :e: if a modification is done in a cache then this fact can be broadcasted to other nodes (SMPS). Each node's directory that receives the broadcast message can then determine if any local cache has that line or not. If any local cache has that line then that node's directory needs to maintain entry indicating that line of memory is invalid & it will remain as invalid until a write back occurs.

  If another processor (local or remote) requests the invalid line then the local directory performs write back operation to update main memory before providing data.

CC-NUMA pros & cons-

- Greater or better performance at higher level parallelism than SMP.
- Bus traffic is handled by the bus for a node but if memory accesses to remote nodes are more then the performance of CC-NUMA degrades. So one of the remedy for this is use of L1 & L2 cache memories to minimize memory access including remote ones.

Distributed Memory Multiprocessor

Cluster

- Cluster can be defined as a group of interconnected, whole computer working together as a unified computing resource that can create the illusion of being one machine. Each computer in a cluster is known as a node.

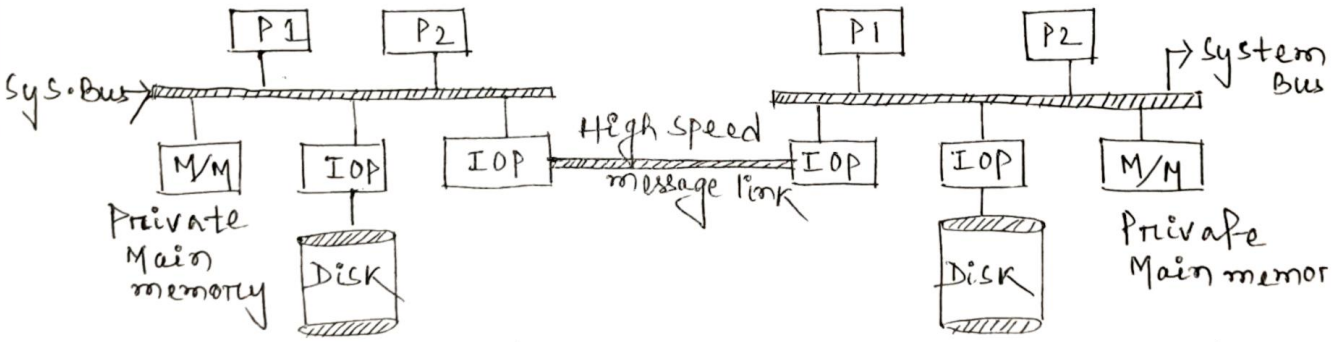
Benefits Of Cluster:-

1. Absolute scalability- a cluster can have hundred or thousand or more machines & each machine is a multiprocessor. So it is possible to create large cluster that act as a powerful standalone machine.
2. Incremental scalability- if an existing small system need an upgrade or expand then it is possible to add new systems to the cluster for small increment.
3. High availability- each node in a cluster is a standalone computer so the failure of one node does not mean loss of service. This may be handled by software in some system.
4. Superior price/performance- it is possible to place together a cluster with equal or greater computing power in low cost than a single large machine.

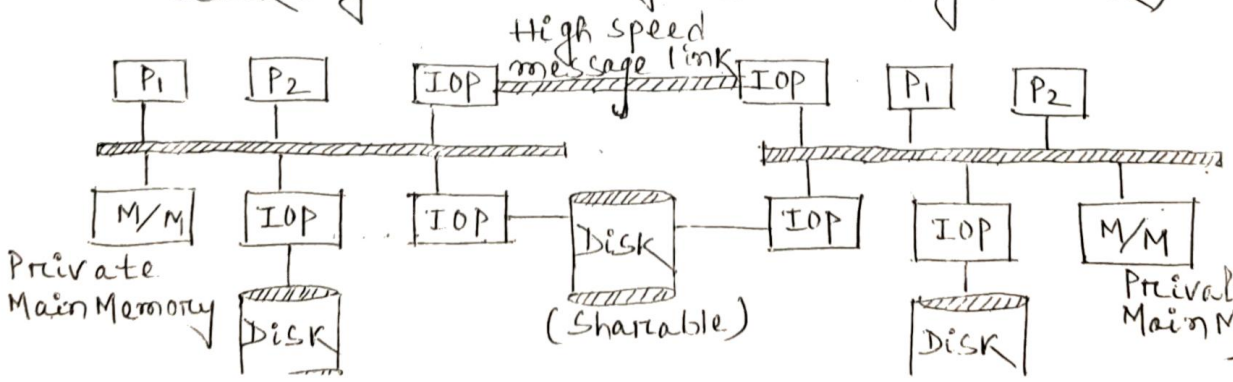
Configuration:-

- Cluster classification is based on whether the computers in a cluster share access to the same disk or not. Two different configuration approaches are-
  1. A high speed message link in between two node cluster to exchange message or to coordinate cluster activity. The link can be a LAN or can be a dedicated interconnection network.
  2. Two or more nodes will have a link or a LAN or a WAN so that there is a connection between server cluster & remote clients.

(Block Diagram of standby server with no shared disk)



(Block Diagram of standby server sharing a disk)



Clustering Method:-

1. Passive standby-

- In this method one of the computer in a cluster known as primary can handle all the processing load while other computer remains inactive, standing by to take over the primary in the event of failure.
- To coordinate among active or primary & other machines, the primary send a "heartbeat" message to the standby machines. When the "heartbeat" message stop arriving then the standby machines assumes that primary server has failed down & then one of the standby machines take over the primary.
- Drawback- The only information exchanged between two system is a heartbeat message & they don't share common disk, so no access to database of primary. Hence a standby machine provides only functional back up. Therefore it is not referred to as cluster.

2. Active secondary-

- In this method multiple interconnected computers that are all actively doing processing while maintaining the image of single system to the outside world.
- It is of 3 types-
  a) Separate server
  b) Server connected to a single disk
  c) Server shares disk

a) Separate server-

- Each computer is a separate server with its own disk & there is no sharing disk in the system. It provides high performance & high availability.
- Management software or scheduling software is there to assign incoming client request to server to balance load & to support high utilization.
- If a computer fails while executing an application, another computer in the cluster can take over & complete the execution of an application. Therefore it is necessary to constantly copy data among systems so that each system can access current data of other system. This leads to high availability but at the cost of performance means performance degrade.

b) Server connected to a single disk-

- Here no.of servers connected to a common disk and the sharable or common disk are partitioned into volumes & each volume is dedicated to a single computer. Use of a sharable common disk is to reduce the communication overhead.
- The cluster is configured in such a way that if any one computer fails then some other computer has ownership of the failed computer.

c) Server shares disk-

- In this approach multiple computers shares the same disk at the same time i:e: each computer has access to all the volumes of all disk at the same time. Therefore this approach requires locking facility so that data can be accessed by only one computer at a time.

Operating System Design Issue In Cluster:-

- The cluster hardware design get exploited if the OS is enhanced .the issues or the exploitations are-
  a) Failure management
  b) Load balancing
  c) Parallelizing computation

a) Failure management- Two different approaches are there to deal with failure. They are-
   1. Highly available cluster- A highly available cluster offers a high probability that all resources will be in service. If any kind of failure occurs in a computer then the queries processed by the system are also lost or if the queries are retrieved then it will be handled by another computer in cluster. But any partial execution transactions are not handled.
   2. Fault-tolerant cluster- Cluster ensures that all the resources are always available & due to use of redundant shared disk the uncommitted transaction are committed if there is any failure in the system. There are two different states of a system-

      Failover- The process of switching application & data resources from a failed system to an alternative system in a cluster is known as failover.

      Failback- The process of restoring of data & application resources to the original system once it has been fixed is known as failback.

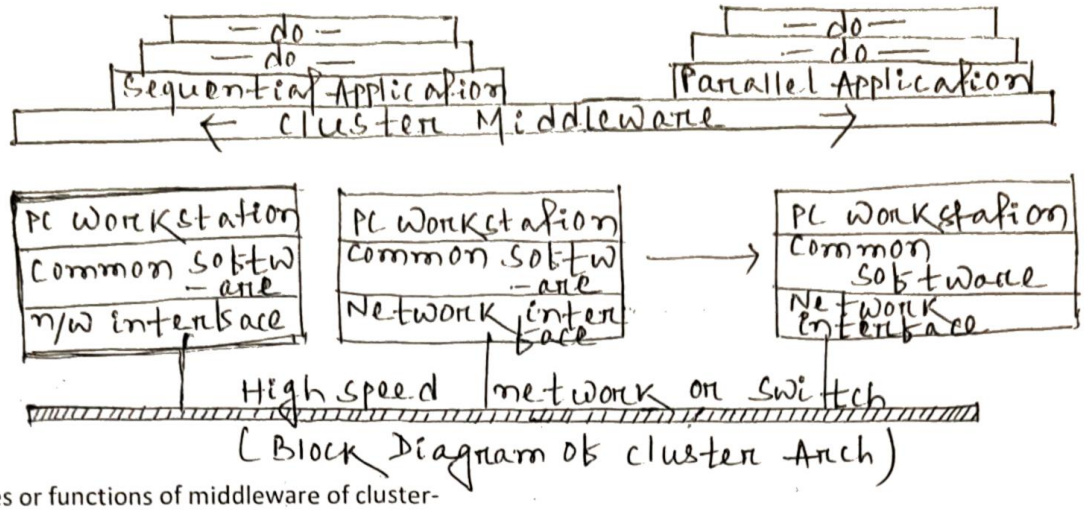
b) Load management- To balance the load among no.of computers of a cluster it is required that cluster should be incrementally scalable i:e: when a new computer added to the cluster , the load balancing facility should automatically include this computer in scheduling application.

c) Parallelizing computation- To take advantage of cluster it is required to execute software from a single application in parallel. But this parallelism computation or execution has 3 different problems. These are-
   1. Parallelism compiler- A parallelizing compiler at compile time can determine which part of an application can be executed in parallel. Then the application gets split off & each part assigned to a different computer in the cluster.

      Performance of parallelizing compiler depends on nature of problem & design of compiler. And it is quiet difficult to design such a compiler.
   2. Parallelized application- In this approach the programmer writes application program to run on a cluster & use message passing to move data between cluster nodes if required. It creates burden on programmers to writes such independent programs in application software.
   3. Parametric computing-in this approach the application program executed no.of times on different set of parameters like simulation model. But parameter processing tools need to organize, manage & run the jobs in an effective manner

Cluster Computer Architecture:-

- Individual multiprocessing computers or nodes are connected by some high-speed LAN or switch. Each computer is capable of operating independently.
- To enable cluster operation each computer is installed with a middleware software. A middleware software provides a unified single system image to the user. Middleware software also provides high availability by load balancing & handling nodes in the event of failure.



(Block Diagram of cluster Arch)

- Services or functions of middleware of cluster-

1. Single entry point- A user log onto a cluster rather than to an individual computer.
2. Single file hierarchy- The users sees a single hierarchy of file directories under same root directory.
3. Single control point- There is a single or default workstation used for cluster management & control.
4. Single virtual network- Even if the actual cluster configuration consists of multiple interconnected networks, any node can access any point (node) in the cluster.
5. Single memory space-In distributed shared memory space, the programs can share variable by using middleware.
6. Single job management- In cluster there is a job scheduler to assign jobs to the host. So a user can simply submit the job without specifying the host computer to execute the job.
7. Single user interface- There is a common graphics interface for all users irrespective of the workstation from which they enter the cluster.
8. Single IO space- Any node in a cluster can access any peripheral devices irrespective of its physical location.
9. Single process space- A process on any node can communicate with any process on a remote node because there is a uniform process-identification scheme.
10. Checkpoint- Checkpoint function in middleware periodically saves the process state & intermediate computing result for a rollback recovery in case of failure event.
11. Process migration- Process assigned to one node by a job scheduler can be assigned to or migrated to another node in a cluster for balancing load.

SYNCHRONIZATION

- In a multiprocessing system, the instruction set contains some basic instructions to implement communication as well as synchronization among cooperating processes. Synchronization refers to the special case where the data or control information used to communicate between processor.
- Synchronization is needed in multiprocessing system to enforce correct sequence of processes & to ensure mutually exclusive access to sharable write data. Hence synchronization enforces mutual exclusion & the most popular way of implementing mutual exclusion is binary semaphore.
- Mutual exclusion using semaphore-
  - Synchronization is a mechanism of providing guaranteed orderly access to shared memory & other sharable resources. Therefore synchronization protects data from being changed simultaneously by two or more processor. This mechanism where sharable data is being protected from simultaneous change by two or more processor is known as mutual exclusion.
  - Mutual exclusion in a multiprocessing system enables one processor to exclude or lock out access to shared resource by other processor when it is in critical section.
  - A critical section is a program sequence that once begun or started, must complete execution before another processor access the same shared resources.
  - A binary variable known as semaphore is used to indicate whether a processor is executing critical section or not. Hence a semaphore is a software–controlled flag that is stored in a memory location so that all the processors can access it.
    - If semaphore=1, it means a processor is executing critical section. So shared memory can't be available to other processors.
    - If semaphore= 0, it means shared memory is available & it can be access by any requesting processor.
  - Therefore when a processor is executing critical section then it set semaphore=1 & when execution of critical section is completed, semaphore is set to 0.
  - A semaphore can be initialized by means of test & set instruction. This test & set instruction must be single & indivisible operation, otherwise two or more processor may test the semaphore simultaneously & then each processor set the value and then consequently each processor enter into critical section at the same time which results in erroneous or loss of information. Therefore a semaphore is initialized by means of a test & set instruction in addition with a hardware-lock mechanism.
    - A hardware-lock is a processor generated signal that prevents other processors from being using the system bus as long as the signal is active.
    - During the execution of test & set instruction, it not only test & set the semaphore but also activates hardware-lock. This prevents other processor from changing the semaphore in between the time of testing & setting semaphore by a processor.
  - Example:-

    Implement of semaphore using instruction, TSL & SEM.

    TSL- TEST, SET & LOCK instruction mnemonic.

    SEM- semaphore instruction mnemonic.

TSL ; semaphore is tested initially by processor

R ← M [SEM] ; value of semaphore variable stored in memory is transferred to R.

If R=1, it means already some processor has set semaphore & entered into critical section. So the processor that has tested the semaphore can't access shared memory.

If R=0, it means noone has set semaphore & noone has entered into critical section. So the processor that has tested the semaphore can now set M [SEM] = 1 & go for execution of critical section.

M [SEM] ← 1

After execution of critical section semaphore set to 0 to release shared resource.

M [SEM] ← 0