Large Computer System

Large Computer System:-

Forms of parallel processing, Array Processor, The Structure of General-purpose, Inter Connection Networks.

Forms of Parallel Processing:-

Many oppurtunities are available for parts of a given Computational task to be excecuted. in parallel. For ex:- in handling I/O operations, most Computer Systems have hardware for that performs direct memory access (DMA) between an I/O device and main memory. The transfer of data in either direction between the main memory and a magnetic disk under the direction of DMA Controller that Operates in parallel with the processor.

When a block of data is to be transferred from disk to main memory, the processor indicates the transfer by sending instructions to the DMA Controlle While the Controller transfers the required data using cycle stealing, the processor Continues to perform Some Computation that is unrelated to the data transfer. When the controller Completes the transfer it Sends an interrupt request to the processor to

Signal that the requested data are available in the St main memory. In response, the processor switches to a Computation that uses the data.

Classification of Parallel Structures:-

A general Classification of parallel processing has been proposed by "Flynn". In this Classification. a Single-processor Computer System is called 'Single Instruction Stream', Single Data Stream [SISD] System. A program excecuted by the processor Constitutes the. Single instruction Stream. In the second Scheme, a Single Stream of instruction is broad cast to a no.of processors.

Each processor Operates on its own data. This Scheme in which all processors excecute the Same prograi but operates on different data is called "Single Instruction Stream Multiple Data Stream" [SIMD] System. The multiple data streams are the sequences of data items accessed by the individual processor in their own memories. The third Scheme involves a no.of independent processors, each excecuting a different program and accessing in Own sequence of data items. Such machines are Called "Multiple Instruction

Stream ,Single Data Stream"[MISD]· System. In such a System, a Common data structure is manipulated by Seperate processors, each excecuting a different program.

Array Processors :-

The SIMD form of parallel processing also Called "array processing". was the first form of parallel processing to be studied and Implemented. In early 1970's a System named ILLIAC-IV was designed at the university of Illinois using the approach and was later built by Bro Burroughs Corporation. A two-dimensional grid of processing elements excecutes an instruction stream that is broad Cast from central Processor· Control processor· As each instruction is broad cast, all elements excecute it Simultaneously. Each processing element is Connected to its four nearest neighbours for purpose of exchanging data.

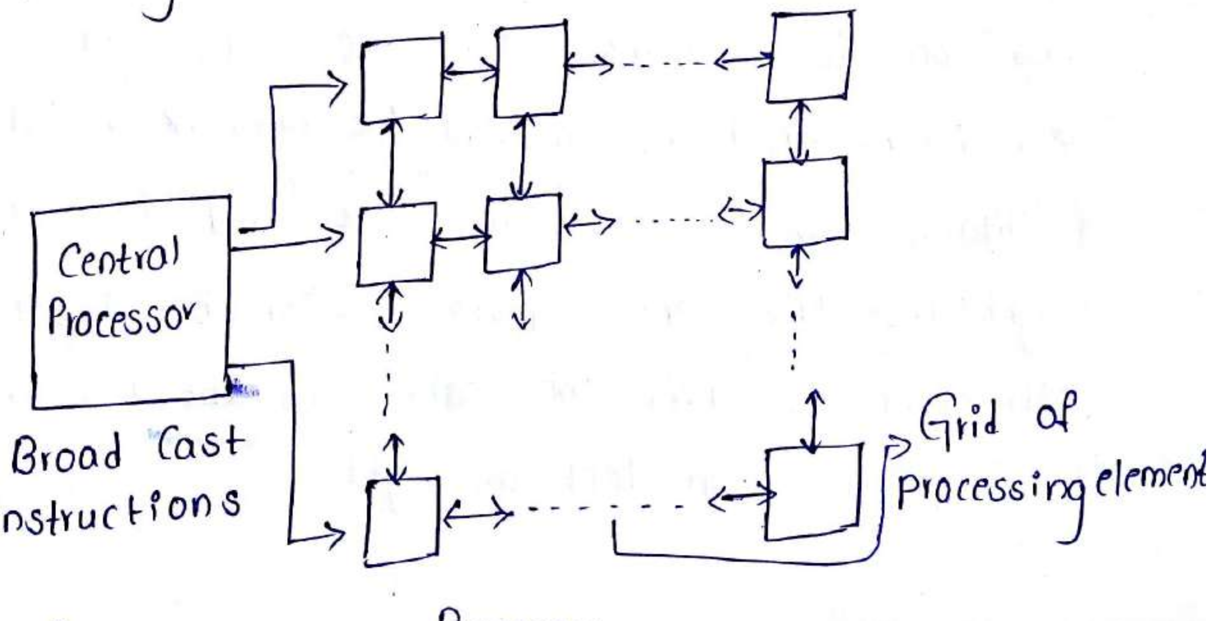


fig:- An array Processor

End-arround Connections may be provided in both rows and columns. The grid of processing elements can be used to Solve two-dimensional problems. Assume that the edges of the plane are held at Some fixed temperatures. The outer edges are initialized to the Specified temperatures. All interior points are initialized to arbitory Values, not necessorily the same. Iterations are then excecuted in parallel at each element. Each iteration Consists of Calculating an improved estimate of the temperature at a point by averaging the Current values of its four nearest neighbours. The process stops, when changes in the estimate during Successive iterations are less than Some predefined Small quantity.

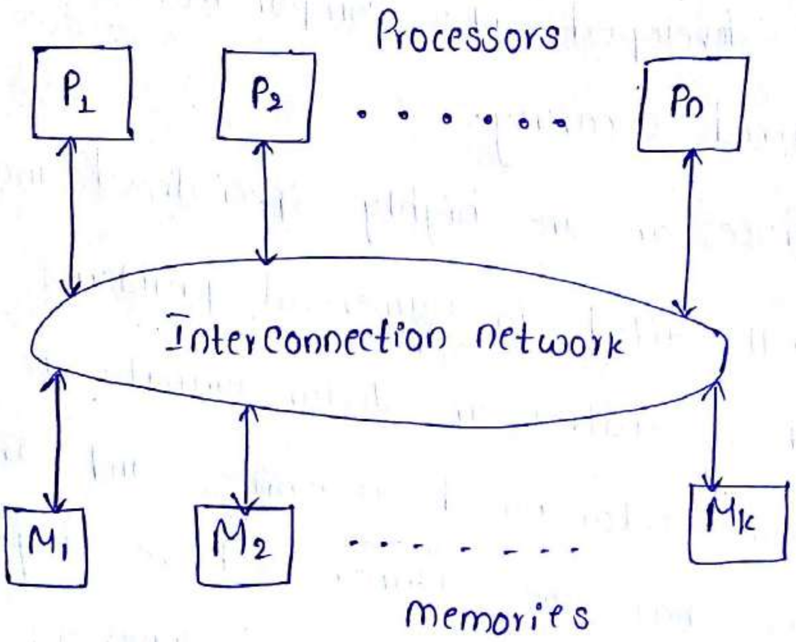
In array processor each element must be able to exchange values with each of its neighbours Over the parts. Each processing element has a few registers and some local memory to store data. It also has a register, which we can call the network register, that facilitates movement of values to and from its neighbours. The Central processor can broadcast an instruction to shift the values in the network register one step up, down, left or right.

Each processing element also contains an ALU to excecute arithmetic instructions broad cast by the Control processor. The Control processor must be able to determine which when each of the processing elements has developed its Component of the temperature to the required accuracy.

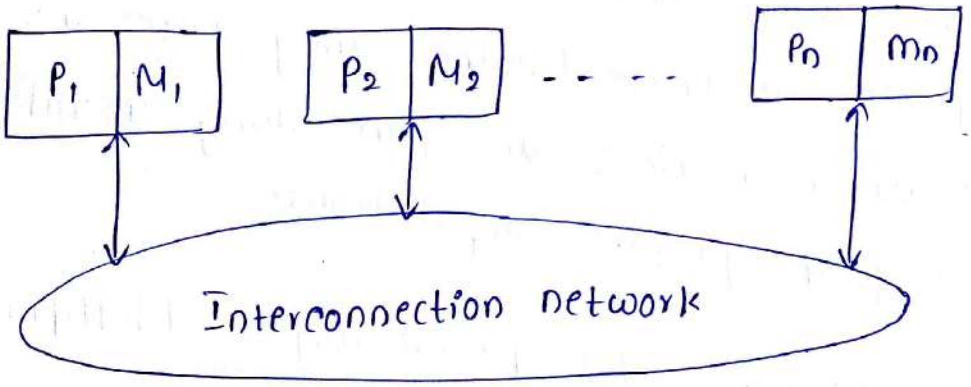
Array processor are highly Specialized machines. They are well-suited to numerical problems that can be expressed in matrix or Vector format. A key difference b/w Vector-based machines and array processors is that the former achieve high performance through a heavy use of pipelining, whereas the latter provide extensive parallelism by replication of Computing modules. Neither array processors nor Vector-based machines are particularly useful in Speeding up general Computations.

The Structure of General-Purpose Multiprocessors:-

The array processor architecture described in the peceding Section is a design for a Computer System that Corresponds directly to a class of Computational problems that exhibit an Obvious form of data parallelism. In more general Cases in which parallelism is not so obvious, it is useful to have an MIMD architecture, which involves a no.of processors Capable of independently excecuting differen routines in parallel.

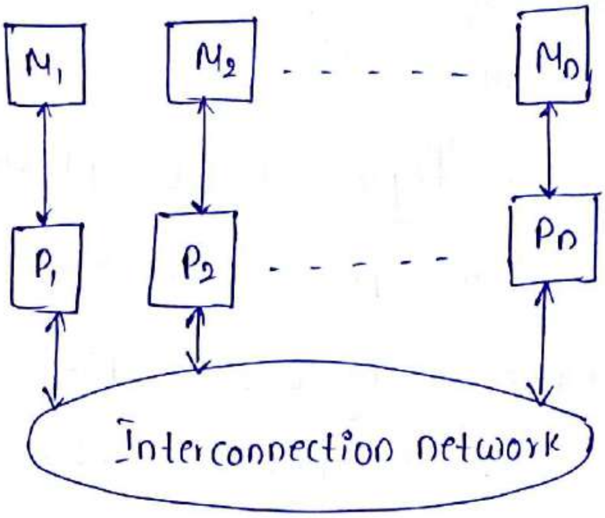


(a) A UMA multiprocessor



(b) A NUMA multiprocessor

An interconnection network permits n processor to access k memories so that any of the processors can access any time of memories. The inter Connection network are may introduce Considerable delay b/w a processor and memory. If this delays is same for all access to memory which is common for this organization, then such machine is called a "Uniform Memory Access" (UMA) multiprocessor.



(c) Distributed memory System

Because of extremely short instruction excecution times achievable by processors, the network delay in fetching instructions and data from the memories in unacceptable if it is too long. Unfortunately interconnection network with very short delays are costly and complex to implement.

An attractive alternative, which alows a high computation in all processors, is to attach the memory modules directly to the processor. In addition to accessing its local memory, each processor can also access, other memories over network. Since, the remote access process through the network, the accesses take Considerably longer than access to the local memory. Because of this difference in access times, such multiprocessors are Called "Non-Uniform Memory Access" (NMUA) multiprocessors.

The Organization of figures a & b provides a "global memory" where any processor can access any memory module without intervention by another processor. Here all memory modules serve as private memories for the Processors that are directly Connected to them. A Processor Cannot access a remote memory without the Cooperation of the remote processor. This Cooperation takes place in the form of messages exchanged by the processor. Such Systems are Called "distributed-memory System" with a message-passing protocol.

Figures a,b,c depict a high-level View of Possible multiprocessor Organizations. The performance and Cost of these machines depend greatly on on implementation details.

Interconnection Networks:-

In general, the network must allow information transfer between any pair of modules in the system. The network may also be used to broadcast information from one module to many other modules. The traffic in the network Consist of requests (such as R/W) data transfers and Various Commands.

The suitability of a particular network is judged in terms of Cost bandwidth, effective pr throughput and ease of implementation. The term "band width" refers to the Capacity of a transmission link to transfer data and is expressed in bits or bytes per second. Then "effective throughput" is the actual rate of data transfer. This rate is less than the available bandwidth because a ginven link usually doesnot carry data all of the time.

Information transfer through the network usually takes place in the form of packet of fixed length and specified format. Longer messages may require many packets.

Ideally, a Complete packet would be handled in parallel in one Clock Cycle at any node or Switch in the network. However, to reduce cost and Complexity, the links are often Considerably narrower. In Such cases, a packet must be divided into Smaller piece each of which can be transmitted in one clock cycle.

Single Bus :-

The Simplest and most economical means for interconnecting a no.of modules is to use a Single bus The Since, Several modules are Connected to the bus and any module Can request a data transfer any time, it is essential to have an efficient bus arbitration scheme.

In a Simple mode of Operation, the bus is dedicated to a particular Source-destination pair for the full duration of the requested transfer. The memory module needs a certain amount t of time to access the data, the bus will be idle until the memory is ready to respond with the data.

Suppose the bus transer takes T times units, and the memory access time is 4T units. It then takes 6T units to Complete a read request. Thus the bus is idle for two thirds of time. A scheme known as "split-transaction protocol" makes it possible to use the bus during the idle period to serve another request.

Consider the following method of handling a series of read requests. possibly from different processors. After transferring the address involved in the first request, the bus may be reassigned to transfer the address for the second request. Assuming that this request is to a different memory module. we now have two modules Proceeding with read access Cycles in parallel. If neither module has finished with its access, the bus may reasigned to a third request and so.on. Address and data transfers for different requests represent independent use of the bus that can be interleaved in any order.

The split-transaction protocol allows the bus and the available bandwidth to be used more efficiently. The performance improvement achieved with this protocol depends on the relationship b/w the bus transfer time and the memory access time. Performance is improved at the cost of increased bus Complexity. There are two reasons why Complexity increases. Since a memory module needs to know which source

initiated a given read request, a Source identificati-
tag must be attached to the request. Complexity also increases because all modules, not just the processors, must be able to act as bus masters.

Multiprocessor that use the split transaction bus vary in size from 4-32 processors. In larger sizes, the bandwidth of the bus can become a problem. The bandwidth can be increased if a wider bus, i.e., a bus that has more wires is used. Most of the data transferre between the processor and memory modules consists of a cache blocks. where a block consists of no.of words. The Challonge multiprocessor from Silicon Graphics Corporation uses a bus that allows parallel transfer of 256 bits of data.

The main limitation of single bus that is no.of modules that can be connected to the bus is not large. An ordinary bus functions well if no more than 10 to 15 modules are connected to it. Using a wider bus to increase the bandwidth allows the no.of modules to be doubled. Networks that allow multiple independent transfer operations to proceed in parallel can provide significantly increased data transfer rates.

Crossbar Networks:-

A Versatile Switching arrangement (is Shown in fig). It is known as "Crossbar switch" which was Originally developed for use in telephone networks. Any module Qi Can be Connected to any other module Qj. by Closing an appropriate Switch. Such networks, where there is a direct link between all pairs of nodes. are Called "fully Connected networks". many Simultaneous transfers are possible. If n Sources need to send data to n distinct destinations, then all of these transfers Can take place Concurrently. Since no transfer is prevented by the lack of a Communication path, the Crossbar is called "nonblocking Switch".

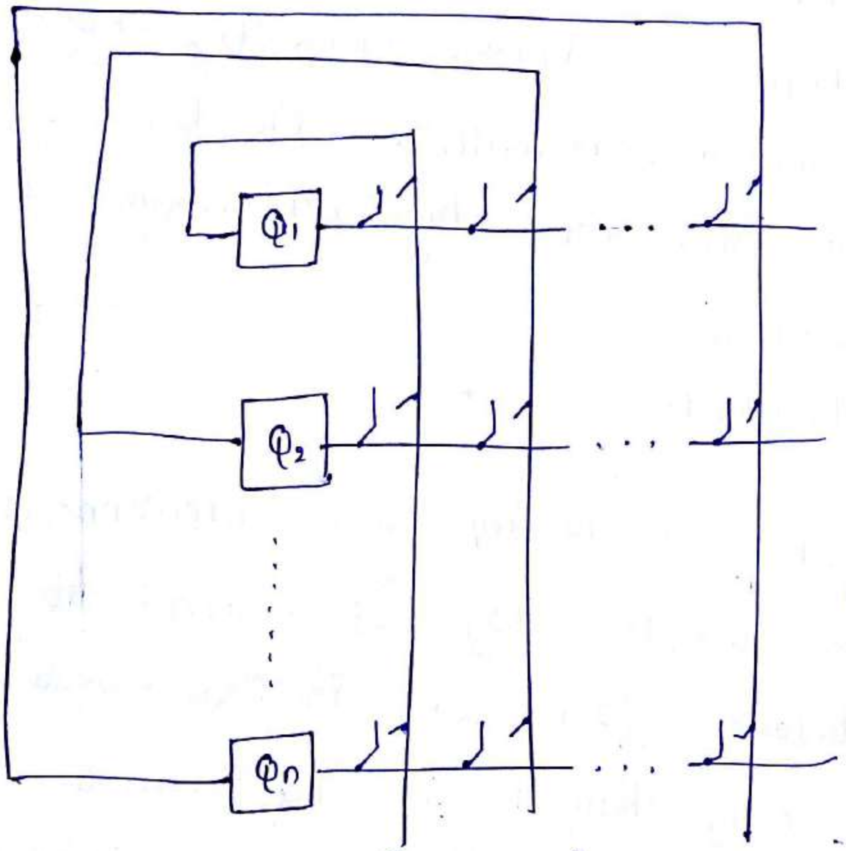


fig: Crossbar interconnection network.

If fig. we show just a single switch at each crosspoint. In actual multiprocessor. however, the paths through the Crossbar network are much wider. The no. of Crosspoints is n^2 in a network used to interconnect n modules, the total no. of switches becomes large as n increases. This results in high cost and cumbersome implementation.

One of the large crossbar switches is found in Sun's E10000 System in which 16 four-processor nodes are connected by a 16×16 crossbar switch. It is also possible to use a multilevel crossbar switch, where a crossbar switch at level 1 connects to a crossbar swith at level 2 and so on. Such schemes are found in Fujits'u's VPP5000, Hitachi's SR8000, and NEC's SX-5 machines. A multilevel crossbar has become a popular choice for a high performance interconnection medium.

Multistage Networks:-

It is also possible to implement interconnection network that use multiple stages of switches to set up paths between sources and destination. Such netwoks are less costly than the crossbar structure. Yet, they provide a reasonable large no. of parallel paths b/w Sources and destinations.

The (fig) shows a three-stage network called a "Shuffle network that interconnects eight modules.

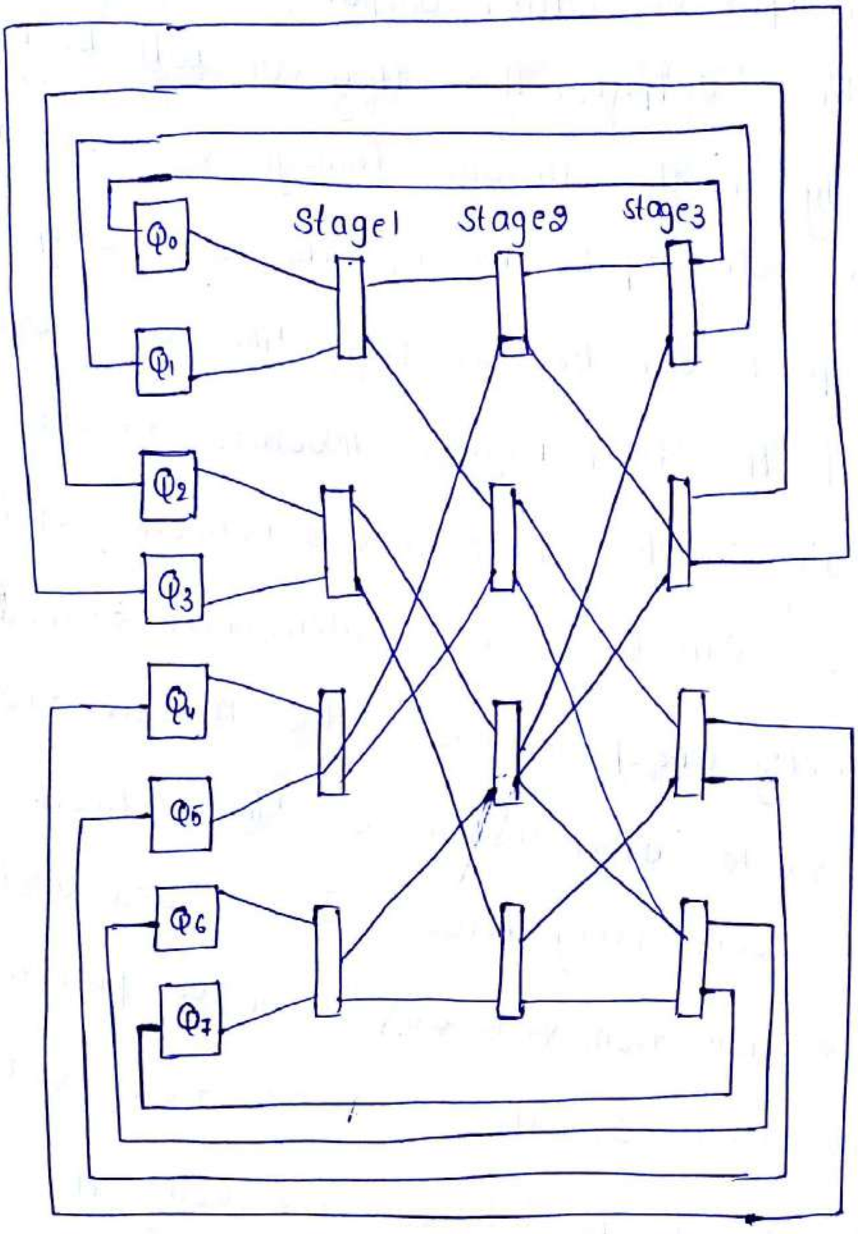


fig: multistage Shuffle network

The term shuffle describes the pattern of Connections from the outputs of one stage to the inputs of next stage. This pattern is identical to the repositioning of playing Cards in a deck that is shuffled by

Splitting the deck into two halves and interleaving the Cards in each half.

Each Switchbox in the fig is a 2x2 Switch that can route either input or either Output. If the inputs request distinct Outputs, then they Can both be routed Simultaneously in the Straight through Or Crossed pattern. If both inputs request the Same Output. Only One request Can be Satisfied. The Other one is blocked until the first request involves finishes using the switch. It Can be shown that a network Consisting of 'S' Stages can be used to interconnect 2^s modules. There is exactly One path through the network from any module Q_i to other module Q_j. This network provides full Connectivity between Sources and destinations

Eg: The Connection from Q_0 to Q_4 Cannot be provided at the Same time as the Connection from Q_1 to Q_5

If n nodes are to be interconnected using the Scheme (in fig) then we must use $s = \log_2 n$ Stages with $n/2$ Switch boxes per stage. Since each Switchbox Contains four switches, the total no.of Switches is

$$4 \times \frac{n}{2} \times \log_2 n = 2n \times \log_2 n$$

which for large networks, is Considerably less than n^2 switches needed in a Crossbar network.

Multistage networks are less Capable of providing Concurrent Connections than Crossbar Switches, but they are also less Costly to implement.

Hypercube Networks:-

The interconnection network imposes the Same delay for paths connecting any two modules. Such Schemes can be used to implement UMA multiprocessors. The network topologies that are suitable only for NUMA multiprocessor. The first Such scheme that gained popular it uses the topology of n-dimensional cube, Called a "hypercube". to implement a network that interconnects 2^n nodes.

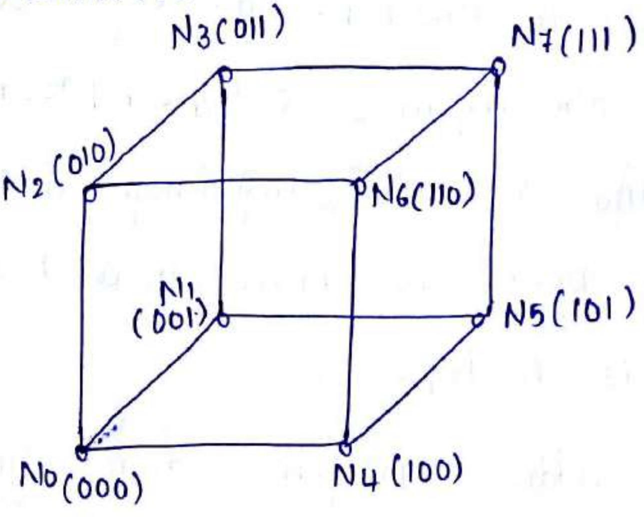


fig: A 3-dimensional hypercube network

The small circles represent the communication circuits in the nodes. The functional units attached to each node. The edges of the cube represent bidirection communication links between neighbouring nodes. In an n-dimensional hypercube, each node is directly connecte to n neighbours. A useful way to label the nodes is to assign binary addresses to them in such way that the address of any two neighbours differ in exactly one-bit position.

Routing messages through the hypercube is particularly easy. The processor at node Ni wishes to send a message to node Nj. The binary addresses of the source i and the destination j are compared from least to most significant bits. The message gets closure to destination node Nj with each of these hops from one node to another. Eg: a message from node N_2 to node N_5 requires 3 hops, passing through nodes N_3 and N_1. The maximum distance that any message needs to travel in a n-dimensional hypercubes is n hops.

Scanning address patterns from right to left is only one of the methods that can be used to determine message routing. The existence of multiple paths b/w two nodes means that when faulty links are encountered, they can usually be avoided by Simple, local routing decision. If one of the Shortest routes is not available, a message may be sent over a long path.

Hypercube interconnection networks have been used in a no.of machines. Eg's include Intel iPSC, which used a 7-dimensional cube to Connect upto 128 nodes and NCUBE's NCUBE/ten which had upto 1024 nodes in a 10-dimensional Cube.

Mesh Networks:-

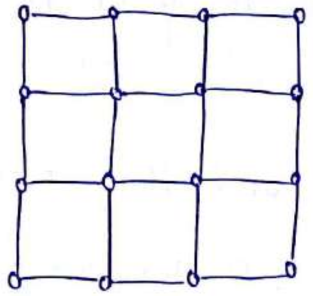


Fig: 2-dimensional mesh network

Interconnecting a large no.of nodes is by means of a mesh. Again the links b/w the nodes are birectional. Mesh gained popularity in the early 1990's and essentially displaced hypercubes for interconnection networks in large multiprocessors.

One of the Simplest and most effective possibilities is to Choose a path b/w a source node Ni and destination node Nj such that the transfer first takes place in the horizontal direction from Ni toward Nj.

Eg's: Intel's paragon and the experimental machine Dash and Flash at stanford University and Alewife at MIT.

If a waparound Connection is made b/w the nodes at the Opposite edges in the result in a network that Consists of a set of bidirectional rings in the x torus, the average latency of information is reduced, but the cost of greater Complexity. Such an interconnection network is used in Fujitsu's AP3000 machines.

Both the regular mesh and the torus Schemes Can also be implemented as 3-D networks, in which the links are b/w neighbors in the x, y & z directions.

Eg's of 3-D torus is found in Cray's T3E multiprocessor.

Tree Networks:-

A hierarchically structured network implemented in the form of tree is another interconnection topology. The (fig a) Shows a 4-way tree that interconnects 16-modules. In this tree, each pattern node allows Communication b/w two of its Children at a time. An intermediate-level node Can provide a Connection from one of its Child nodes to its parent. This enables two leaf nodes that are any distance apart to Communicate.

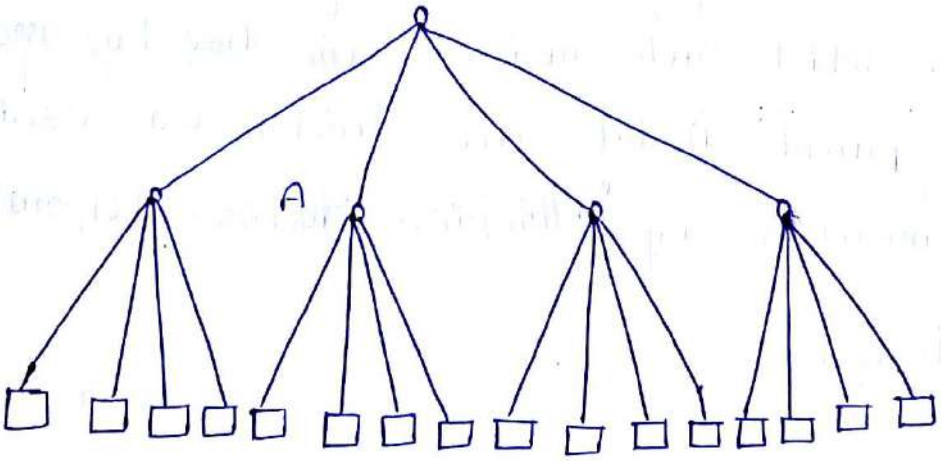
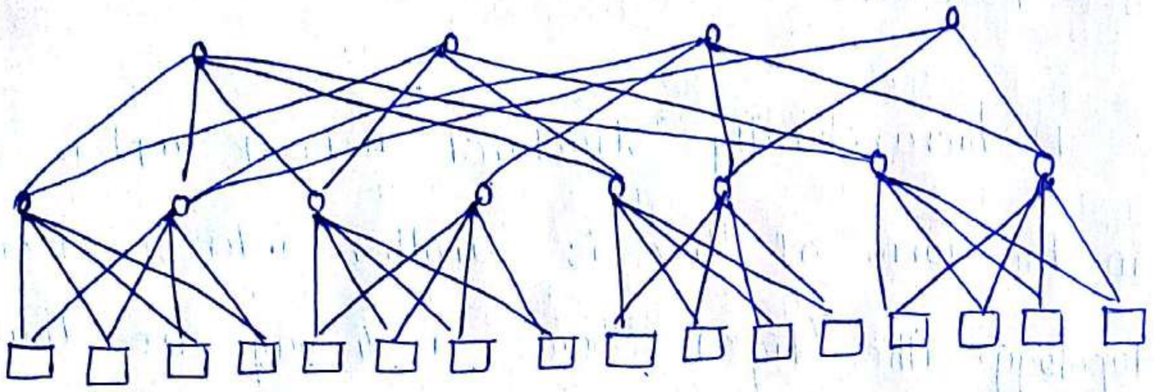


fig: A four-way tree.

A tree network performs well if there is a large amount of locality in Communication that is only a Small portion of network traffic goes through the Single rote noode.



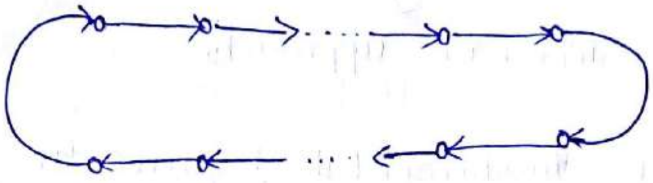
(b) fat Tree

If this is not these case, performance detoriorates rapidly because the root node becomes a bottleneck.

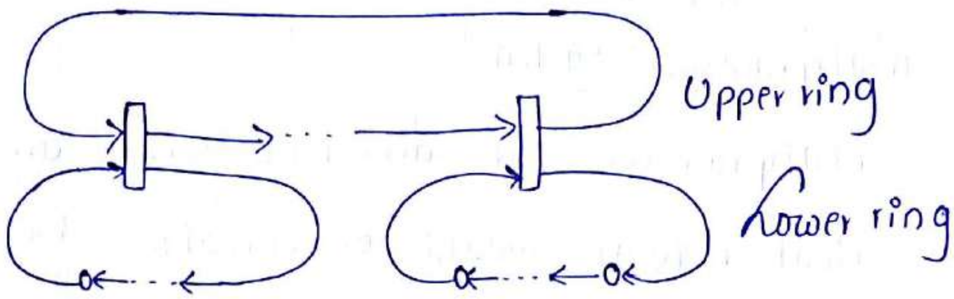
To reduce the possibility of a bottleneck the no.of links in the upperlevels of a tree hierarchy can be increased. This is done in a "fat tree" network, in which each node in the tree has more than one parent. A fat tree structure was used in CM-5 meachine by ".Thinking Machines Corporation."

Ring Networks:-

One of the Simplest network topologies is uses a ring to interconnect the node in the system. The main advantage of this arrangement is that the ring is easy to implement. Links in the ring can be wide, usually accommodating a Complete packet in parallel, because each node is Connected to only two neighbours. However, it is not useful to construct a very long ring to connect many nodes because the latency of information transfer would be unacceptably large.



(a) Single Ring



(b) Hierarchy of rings

Having short rings reduces substantially the latency of transfers that involves nodes on the same ring. The drawback of this scheme is that the highest-level ring may become a bottleneck for traffic.

Commercial machines that feature ring network include Exemplar V2600 by Hewlett-Packard and KSR-2 by Kendal Square Research.

Practical Considerations :-

Several different topologies can be used to implement the interconnection network in a multiprocessor system. Each has certain advantages and disadvantages when comparing different approaches.

The most fundamental requirement is that to Communication network be fast enough and have Sufficient throughput to Satisfy the traffic demand in multiprocessor system.

Multiprocessor of different sizes are needed. The ideal network would be Suitable for all sizes ranging from just a few processor to possibly thousands of processors. The term "Scalability" is often used to describe the ability of multiprocessor architecture to provide increased performance as the size of system increases, while the increase in cost is proportional to the increase in size.

In addition to providing the basic Communication b/w Sources and destinations it is useful to have broadcasting Capability where a message traverses the entire network and is received by all nodes.

The ability to Send a message to only a Subset of the network nodes is to also beneficial. Such transfers are Called multicasting.

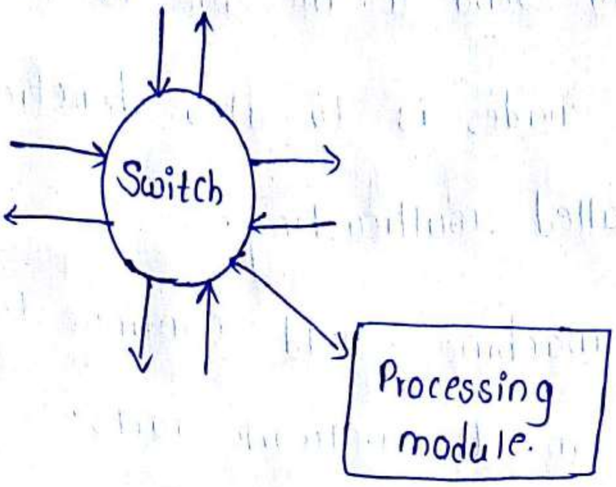
Ideally, the machine could Continue to function even if Some link in the network fails.

Meshes and Rings:-

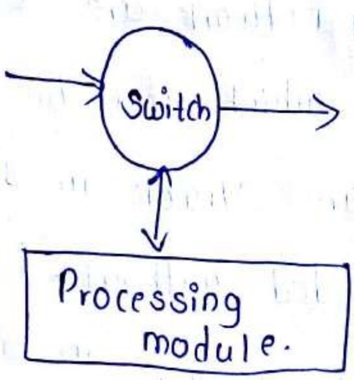
Both mesh and ring network are Characterized by point-to-point links which can be driven at high clock rates. Both are Viable in small Configurations and can be expanded without difficulty. Incremental expansion is Smaller in a ring network than in mesh network.

The Switch block includes both the Circuitry that selects the path for transfer and the buffers needed to hold the data being transferred. Data are transferred from the buffer in one node to the buffer in the next node in one Clock cycle.

The bidirectional Communication is needed in both X & Y directions, eight distinct network link must be Connected to the node.



(a) node in a mesh



(b) Node in a ring

The width of the links are limited by total no.of wires that can be used, taking into the account the Cost and packaging. The term flit [Flow Clontrol digIT] is often used to refer to a portion of the packet that can be accepted by switching circuitry in the node for forwarding, or buffering in case the forward path is blocked by another transfer.

A Straightforward scheme known as the

"store-and-forward" method is to provide a large enough buffer in each node to hold all flits of packets. Thus, an entire packet is transferred from one module node to another where it is stored until it can be forwarded to the next node.

An attractive alternative is the "wormhole" routing scheme in which sequence of flits that constitute a packet can be viewed as a worm that moves through the network. The first flit in a worm contains a header that includes the address of the destination node.

Wormhole routing is an application of a strategy known as circuit switching which is a familiar concept from telephone networks, where a path through the network is established when a number is dialed. The conservation takes place along this path called circuit. Once a circuit is established, however, the remaining flits of the packet move toward the destination without experiencing any contention. The stratagies in which an entire packet is buffered at each node as in the store-and-forward method called "packet switching".

The main disadvandage of a hierarchical ring network that the ring at the top of the hierarchy may become a bottleneck if too many packets need to be transfered over it. This will occur if the locality in comminication is low. The ring base system has hundreds of processors. In contrast mesh-based system scale well to thousands of processors.

The Ring-based System are easier to implement but do not scale as well as mesh-based system. The mesh systems are suitable for use in both small and very large systems.

Mixed Topology Networks:-

Designers of multiprocessors system strive to achieve superior performance at a reasonable cost. In an effort to exploit most advantageous characteristics of different topologies, many sucessful machines feature mixed topologies. Bus and crossbar are excellent choices for connecting a few processors together.

Data General's AV25000 system uses nodes where processors are connected by the bus.

These nodes are interconnected using a ring network. Hewlett- packard's Exemplar V2600 also uses a ring network to interconnect nodes. where each nodes has a Crossbar Switching Connect the processors. Compaq's Alpha Server sc uses a fat tree to interconnect the node that Comprice processors Connected by Crossbar Switch.

Symmetric Multiprocessors :-

A multiprocessor System in which all processors have identical access to all memory modules and all I/O devices. If any processor can excecute either the Operating System kernel or use programs, the machine is called a "Symmetric Multiprocessor "[SMP]". This also implies that any processor can initiate an I/O Operation to On any I/O device and it can handle any external interrupt.

SMP's are usually implemented using either a bus or Crossbar network. Often. an smp is used as an node in much large multiprocessor system.