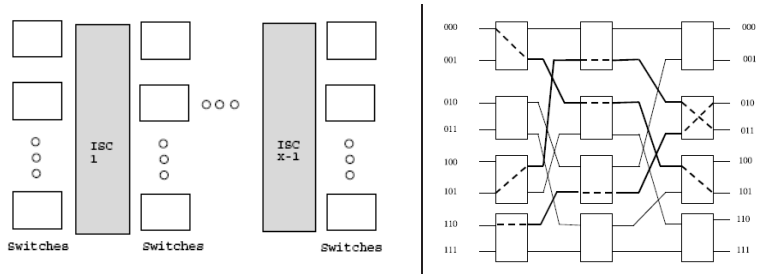

Figure 1: Multistage interconnection network and an example 8×8 Shuffle-Exchange network (SEN).

0.0.1 Multistage Networks

- Multistage interconnection networks (MINs) were introduced as a means to improve some of the limitations of the single bus system (to improve is the availability of only one single path) while keeping the cost within an affordable limit.

- Such MINs provide a number of simultaneous paths between the processors and the memory modules (see Fig. 1a).

- A general MIN consists of a number of stages each consisting of a set of 2×2 switching elements. Stages are connected to each other using Inter-stage Connection (ISC) Pattern. These patterns may follow any of the routing functions such as Shuffle-Exchange, Butterfly, Cube, and so on.

- Figure 1b shows an example of an 8×8 MIN that uses the 2×2 SEs described before. This network is known in the literature as the Shuffle-Exchange network (SEN).

- The figure shows how three simultaneous paths connecting the three pairs of input/output 000 $\longrightarrow$ 101, 101 $\longrightarrow$ 011, and 110 $\longrightarrow$ 010 can be established. It should be noted that the interconnection pattern among stages follows the shuffle operation.

- In MINs, the routing of a message from a given source to a given destination is based on the destination address (self-routing). There exist $log_2 N$ stages in an $N \times N$ MIN.

- The number of bits in any destination address in the network is $log_2 N$. Each bit in the destination address can be used to route the message

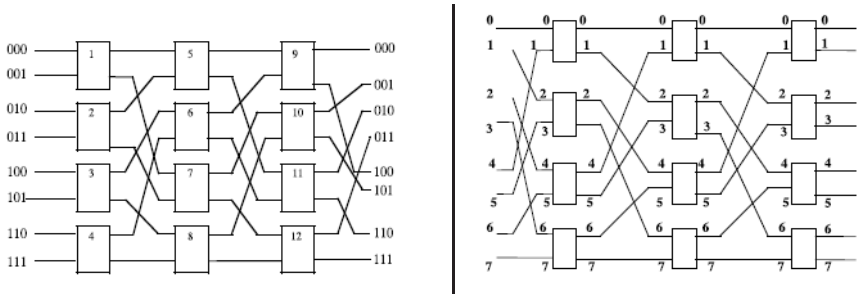

Figure 2: Multistage interconnection network and an example 8×8 Shuffle-Exchange network (SEN).

through one stage. The destination address bits are scanned from left to right and the stages are traversed from left to right.

- The first (most significant bit) is used to control the routing in the first stage; the next bit is used to control the routing in the next stage, and so on. The convention used in routing messages is that if the bit in the destination address controlling the routing in a given stage is 0, then the message is routed to the upper output of the switch. On the other hand if the bit is 1, the message is routed to the lower output of the switch.

- Consider, for example, the routing of a message from source input 101 to destination output 011 in the 8×8 SEN shown in Figure 1b. Since the first bit of the destination address is 0, therefore the message is first routed to the upper output of the switch in the first (leftmost) stage. Now, the next bit in the destination address is 1, thus the message is routed to the lower output of the switch in the middle stage. Finally, the last bit is 1, causing the message to be routed to the lower output in the switch in the last stage.

- **The Banyan Network**; A number of other MINs exist, among these the Banyan network is well known (see Fig. 2a, an example of an 8×8 Banyan network).

- If the number of inputs, for example, processors, in an MIN is N and the number of outputs, for example, memory modules, is N, the number of MIN stages is $log_2 N$ and the number of SEs per stage is $N/2$, and hence the network complexity, measured in terms of the total number of SEs is $O(Nlog_2 N)$.

- The time complexity, measured by the number of SEs along the path from input to output, is $O(log_2 N)$. For example, in a 16×16 MIN, the length of the path from input to output is 4.

- The total number of SEs in the network is usually taken as a measure for the total area of the network. The total area of a 16×16 MIN is 32 SEs.

- **The Omega Network;** The Omega Network represents another well-known type of MINs. A size N omega network consists of n ($n = log_2 N$ single-stage) Shuffle-Exchange networks. Each stage consists of a column of N=2, two-input switching elements whose input is a shuffle connection. (Figure 2b illustrates the case of an $N = 8$ Omega network.

- As can be seen from the figure, the inputs to each stage follow the shuffle interconnection pattern. Notice that the connections are identical to those used in the 8×8 Shuffle-Exchange network (SEN) shown in Fig. 2a.

- Owing to its versatility, a number of university projects as well as commercial MINs have been built. These include the Texas Reconfigurable Array Computer (TRAC) at the University of Texas at Austin, the Cedar at the University of Illinois at Urbana-Champaign, the RP3 at IBM, the Butterfly by BBN Laboratories, and the NYU Ultracomputer at New York University.

- The NYU Ultracomputer is an experimental shared memory MIMD architecture that could have as many as 4096 processors connected through an Omega MIN to 4096 memory modules.

- The MIN is an enhanced network that can combine two or more requests bound for the same memory address. The network interleaves consecutive memory addresses across the memory modules in order to reduce conflicts in accessing different data elements.

- The switch nodes in the NYU Ultracomputer are provided with queues (queue lengths of 8 to 10 messages) to handle messages collision at the switch. The system achieves one-cycle processor to memory access.

0.0.2 Blockage in Multistage Interconnection Networks

- A number of classification criteria exist for MINs. Among these criteria is the criterion of blockage.

- **Blocking Networks;** Blocking networks possess the property that in the presence of a currently established interconnection between a pair of input/output, the arrival of a request for a new interconnection between two arbitrary unused input and output may or may not be possible.

- Examples of blocking networks include Omega, Banyan, Shuffle-Exchange, and Baseline. Consider, for example the SEN shown in Figure 1b. In the presence of a connection between input 101 and output 011, a connection between input 100 and output 001 is not possible. This is because the connection 101 to 011 uses the upper output of the third switch from the top in the first stage. This same output will be needed by the requested connection 100 to 001.

- This contention will lead to the inability to satisfy the connection 100 to 001, that is, blocking. Notice however that while connection 101 to 011 is established, the arrival of a request for a connection such as 100 to 110 can be satisfied.

- **Nonblocking Networks;** Nonblocking networks are characterized by the property that in the presence of a currently established connection between any pair of input/output, it will always be possible to establish a connection between any arbitrary unused pair of input/output. The *Clos* is a well-known example of nonblocking networks.

0.1 Static Interconnection Networks

Static (fixed) interconnection networks are characterized by having fixed paths, unidirectional or bidirectional, between processors. Two types of static networks can be identified. These are completely connected networks (CCNs) and limited connection networks (LCNs).

0.1.1 Completely Connected Networks

- In a completely connected network (CCN) each node is connected to all other nodes in the network. Completely connected networks guarantee fast delivery of messages from any source node to any destination node (only one link has to be traversed).

- Completely connected networks are, however, expensive in terms of the number of links needed for their construction.

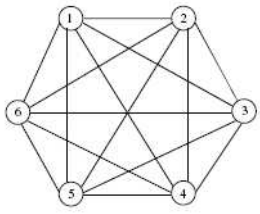

Figure 3: Completely Connected Network.

- It should be noted that the number of links in a completely connected network is given by $N(N-1)/2$, that is, $O(N^2)$.
- The delay complexity of CCNs, measured in terms of the number of links traversed as messages are routed from any source to any destination is constant, that is, $O(1)$.
- An example having $N = 6$ nodes is shown in Fig. 3. A total of 15 links are required in order to satisfy the complete interconnectivity of the network.

0.1.2 Limited Connection Networks

- Limited connection networks (LCNs) do not provide a direct link from every node to every other node in the network. Instead, communications between some nodes have to be routed through other nodes in the network.
- Two other conditions seem to have been imposed by the existence of limited interconnectivity in LCNs.

  1. the need for a pattern of interconnection among nodes and
     - linear arrays;
       * in the worst possible case, when node 1 has to send a message to node N, the message has to traverse a total of $N-1$ nodes before it can reach its destination.
       * Therefore, although linear arrays are simple in their architecture and have simple routing mechanisms, they tend to be slow. This is particularly true when the number of nodes N is large.
       * The network complexity of the linear array is O(N) and its time complexity is O(N).

- ring (loop) networks;
- tree networks;
  - In a tree network, of which the binary tree (shown in Fig. **??**d) is a special case, if a node at level i (assuming that the root node is at level 0) needs to communicate with a node at level j, where $i > j$ and the destination node belongs to the same root's child subtree, then it will have to send its message up the tree traversing nodes at levels $i-1, i-2, \ldots, j+1$ until it reaches the destination node.
  - If a node at level i needs to communicate with another node at the same level i (or with node at level $j \neq i$ where the destination node belongs to a different root's child subtree), it will have to send its message up the tree until the message reaches the root node at level 0. The message will have to be then sent down from the root nodes until it reaches its destination.
  - It should be noted that the number of nodes (processors) in a binary tree system having k levels can be calculated as:
$$N(k) = 2^0 + 2^1 + 2^2 + \ldots + 2^k$$
  - Notice also that the maximum depth of a binary tree system is $log_2 N$, where N is the number of nodes (processors) in the network.
  - Therefore, the network complexity is $O(2^k)$ and the time complexity is $O(log_2 N)$.
- cube networks;
  - Cube-connected networks are patterned after the n-cube structure. An n-cube (hypercube of order n) is defined as an undirected graph having 2n vertices labeled 0 to $2n-1$ such that there is an edge between a given pair of vertices if and only if the binary representation of their addresses differs by one and only one bit.
  - A 4-cube is shown in Fig. 4. In a cube-based multiprocessor system, processing elements are positioned at the vertices of the graph. Edges of the graph represent the point-to-point communication links between processors.
  - As can be seen from the figure, each processor in a 4-cube is connected to four other processors. In an n-cube, each processor has communication links to n other processors.

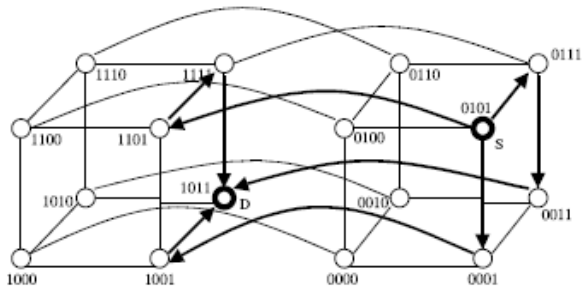

Figure 4: A 4-cube.

* Recall that in a hypercube, there is an edge between a given pair of nodes if and only if the binary representation of their addresses differs by one and only one bit. This property allows for a simple message routing mechanism.
* The route of a message originating at node i and destined for node j can be found by XOR-ing the binary address representation of i and j.
* If the XOR-ing operation results in a 1 in a given bit position, then the message has to be sent along the link that spans the corresponding dimension.
  - For example, if a message is sent from source (S) node 0101 to destination (D) node 1011, then the XOR operation results in 1110.
  - That will mean that the message will be sent only along dimensions 2, 3, and 4 (counting from right to left) in order to arrive at the destination.
  - The order in which the message traverses the three dimensions is not important. Once the message traverses the three dimensions in any order it will reach its destination.
* The hypercube is referred to as a logarithmic architecture. This is because the maximum number of links a message has to traverse in order to reach its destination in an n-cube containing $N = 2^n$ nodes is $log_2 N = n$ links.
* It is worth mentioning that the Intel iPSC is an example of hypercube-based commercial multiprocessor systems. A number of variations to the basic hypercube interconnection have been proposed.

– two-dimensional arrays (nearest-neighbor mesh);

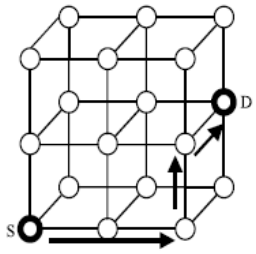

Figure 5: A $3 \times 3 \times 2$ mesh network.

* An n-dimensional mesh can be defined as an interconnection structure that has $K_0 \times K_1 \times \ldots \times K_{n-1}$ nodes where n is the number of dimensions of the network and K_i is the radix of dimension i. Figure 5 shows an example of a $3 \times 3 \times 2$ mesh network.
* A node whose position is (i, j, k) is connected to its neighbors at dimensions $i \pm 1$, $j \pm 1$, and $k \pm 1$. Mesh architecture with wrap around connections forms a torus.
* A number of routing mechanisms have been used to route messages around meshes. One such routing mechanism is known as the dimension-ordering routing.Using this technique, a message is routed in one given dimension at a time, arriving at the proper coordinate in each dimension before proceeding to the next dimension.
  - Consider, for example, a 3D mesh. Since each node is represented by its position (i, j, k), then messages are first sent along the i dimension, then along the j dimension, and finally along the k dimension.
  - At most two turns will be allowed and these turns will be from i to j and then from j to k.
  - In Fig. 5, it is shown the route of a message sent from node S at position $(0, 0, 0)$ to node D at position $(2, 1, 1)$.
* Other routing mechanisms in meshes have been proposed. These include dimension reversal routing, the turn model routing, and node labeling routing.
* Multiprocessors with mesh interconnection networks are able to support many scientific computations very effi-

ciently. It is also known that n-dimensional meshes can be laid out in n dimensions using only short wires and built using identical boards, each requiring only a small number of pins for connections to other boards.

* Another advantage of mesh interconnection networks is that they are scalable. Larger meshes can be obtained from smaller ones without changing the node degree (a node degree is defined as the number of links incident on the node).
* Because of these features, a large number of distributed memory parallel computers utilize mesh interconnection networks. Examples include MPP from Goodyear Aerospace, Paragon from Intel, and J-Machine from MIT.

2. the need for a mechanism for routing messages around the network until they reach their destinations.

0.2 Analysis and Performance Metrics

- For dynamic networks, we discuss the performance issues related to cost, measured in terms of the number of cross points (switching elements), the delay (latency), the blocking characteristics, and the fault tolerance.
- For static networks, we discuss the performance issues related to degree, diameter, and fault tolerance.

0.2.1 Dynamic Networks

- **The Crossbar**;
  - the cost of the crossbar system can be measured in terms of the number of switching elements (cross points) required inside the crossbar. The crossbar possesses a quadratic rate of cost (complexity) given by $O(N^2)$.
  - The delay (latency) within a crossbar switch, measured in terms of the amount of the input to output delay, is constant.The crossbar possesses a constant rate of delay (latency) given by $O(1)$. It should be noted that the high cost (complexity) of the crossbar network pays off in the form of reduction in the time (latency).
  - The crossbar is however a nonblocking network; that is, it allows multiple output connection pattern (permutation) to be achieved.

- A fault-tolerant system can be simply defined as a system that can still function even in the presence of faulty components inside the system. The crossbar can be affected by a single-point failure. Nevertheless, segmenting the crossbar and realizing each segment independently can reduce the effect of a single-point failure in a crossbar.

- **Multiple Bus**;
  - It consists of M memory modules, N processors, and B buses. A given bus is dedicated to a particular processor for the duration of a bus transaction.
  - A processor?memory transfer can use any of the available buses. Given B buses in the system, then up to B requests for memory use can be served simultaneously.
  - A multiple bus possesses an $O(B)$ rate of cost (complexity) growth.
  - The multiple bus possesses an $O(B \times N)$ rate of delay (latency) growth.
  - Multiple bus-multiprocessor organization offers the desirable feature of being highly reliable and fault-tolerant. This is because a single bus failure in a B bus system will leave $(B-1)$ distinct fault-free paths between the processors and thememory modules.
  - On the other hand, when the number of buses is less than the number of memory modules (or the number of processors), bus contention is expected to increase.

- **Multistage Interconnection Networks**;
  - Each stage consists of $N/2$, 2×2 SEs.
  - The network cost (complexity), measured in terms of the total number of SEs, is $O(N \times log_2 N)$.
  - The latency (time) complexity, measured by the number of SEs along the path from input to output, is $O(log_2 N)$.
  - Simplicity of message routing inside a MIN is a desirable feature of such networks. There exists a unique path between a given input?output pair.
  - MINs are characterized as being 0-fault tolerant; that is, a MIN cannot tolerate the failure of a single component.

Network	Delay (Latency)	Cost (Complexity)	Blocking	Degree of Fault Tolerance
Bus	$O(N)$	$O(1)$	Yes	0
Multiple bus	$O(mN)$	$O(m)$	Yes	$(m-1)$
MINs	$O(\log N)$	$O(N \log N)$	Yes	0
Crossbar	$O(1)$	$O(N^2)$	No	0

Figure 6: Performance Comparison of Dynamic Networks.

Based on the above discussion, Fig. 6 provides an overall performance comparison among different dynamic interconnection networks. (N represent the number of inputs (outputs) while m represents the number of buses)

0.2.2 Static Networks

1. Degree of a node, d, is defined as the number of channels incident on the node.

2. Diameter, D, of a network having N nodes is defined as the longest path, p, of the shortest paths between any two nodes. For example, the diameter of a 4×4 Mesh $D = 6$.

3. A network is said to be symmetric if it is isomorphic to itself with any node labeled as the origin; that is, the network looks the same from any node. Rings and Tori networks are symmetric while linear arrays and mesh networks are not.

- **Completely Connected Networks (CCNs)**;
  - the cost of a completely connected network having N nodes, measured in terms of the number of links in the network, is given by $N(N-1)/2$, that is, $O(N^2)$.
  - The delay (latency) complexity of CCNs, measured in terms of the number of links traversed as messages are routed from any source to any destination, is constant, that is, O(1).
  - The degree of a node in CCN is $N-1$, that is, $O(N)$, while the diameter is $O(1)$.
- **Linear Array Networks (LCNs)**

– In this network architecture, each node is connected to its two immediate neighboring nodes. Each of the two nodes at the extreme ends of the network is connected only to its single immediate neighbor.

– The network cost (complexity) measured in terms of the number of nodes of the linear array is $O(N)$.

– The delay (latency) complexity measured in terms of the average number of nodes that must be traversed to reach from a source node to a destination node is $N/2$, that is, $O(N)$.

– The node degree in the linear array is 2, that is, $O(1)$ and the diameter is $(N-1)$, that is, $O(N)$.

Network	Degree (d)	Diameter (D)	Cost (No. of Links)	Symmetry	Worst Delay
CCNs	$N-1$	1	$N(N-1)/2$	Yes	1
Linear array	2	$N-1$	$N-1$	No	N
Binary tree	3	$2(\lceil \log_2 N \rceil - 1)$	$N-1$	No	$\log_2 N$
n-cube	$\log_2 N$	$\log_2 N$	$nN/2$	Yes	$\log_2 N$
2D-mesh	4	$2(n-1)$	$2(N-n)$	No	$\sqrt{N}$
k-ary n-cube	$2n$	$N\lfloor k/2 \rfloor$	$n \times N$	Yes	$k \times \log_2 N$

Figure 7: Performance Characteristics of Static Networks.

In Fig. 7, the basic performance characteristics of a number of static interconnection networks are summarized. (N is the number of nodes and n is the number of dimensions)