Interconnection Networks

By

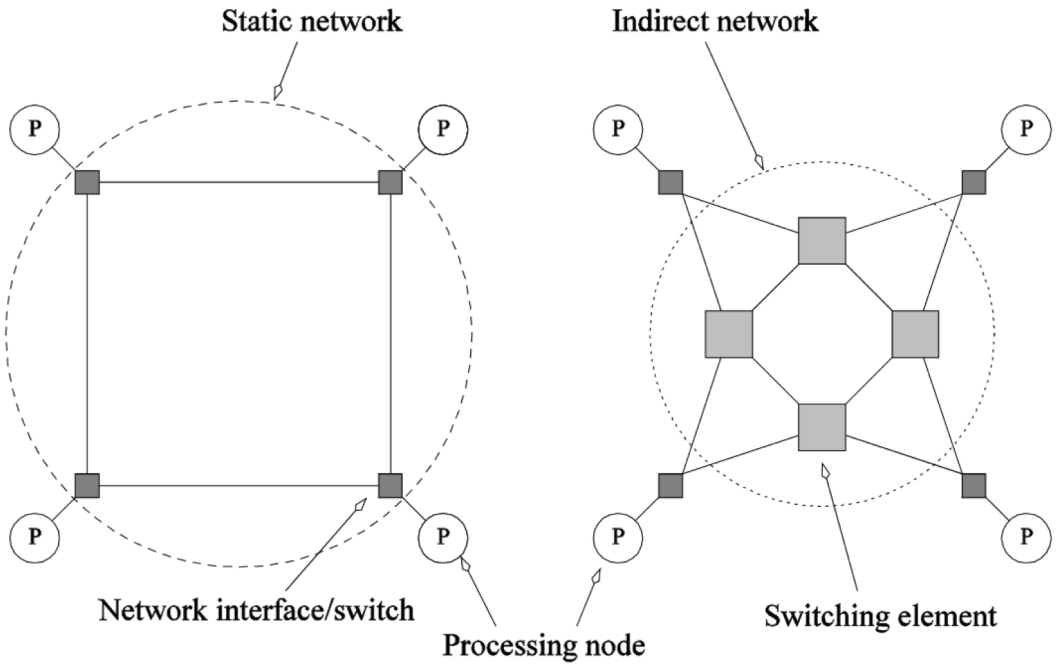
Dr. B. Neelima

Evolution

- Employed in telephone industry in 1950s
- Computer industry to provide faster communication
- Performance requirements of many applications –exceeds the capabilities of single processor
- Any parallel system must employ communication system
- Otherwise, benefits are negated
- Mainly used for transportation of data
- Network topology refers to the layouts of links and switch boxes that establish interconnections

Types of topologies

- Two topologies: static and dynamic
- Static networks consist of point-to-point communication links among processing nodes and are also referred to as direct networks.
- Dynamic networks are built using switches and communication links. Dynamic networks are also referred to as indirect networks.



Classification of interconnection networks: (a) a static network; and (b) a dynamic network.

Terminology

- Network interface-Connects endpoints (e.g. cores) to network. Decouples computation/communication
- Links-Bundle of wires that carries a signal
- Switch/router-Connects fixed number of input channels to fixed number of output channels
- Channel-A single logical connection between routers/switches
- The diameter of a network is defined as the largest minimum distance between any pair of nodes
- The diameter can be used to compare the relative performance characteristics of different networks

Terminology

- Node-A network endpoint connected to a router/switch
- Message-Unit of transfer for network clients (e.g. cores, memory)
- Packet-Unit of transfer for network
- Flit-Flow control digit-Unit of flow control within network
- Direct or Indirect Networks-Endpoints sit “inside” (direct) or “outside” (indirect) the network-E.g. mesh is direct; every node is both endpoint and switch

Terminology

- Switches map a fixed number of inputs to outputs
- The total number of ports on a switch is the degree of the switch
- The cost of a switch grows as the square of the degree of the switch, the peripheral hardware linearly as the degree, and the packaging costs linearly as the number of pins
- The relative speeds of the I/O and memory buses impact the performance of the network
- A variety of network topologies have been proposed and implemented
- These topologies tradeoff performance for cost
- Commercial machines often implement hybrids of multiple topologies for reasons of packaging, cost, and available components

Properties of topology/network

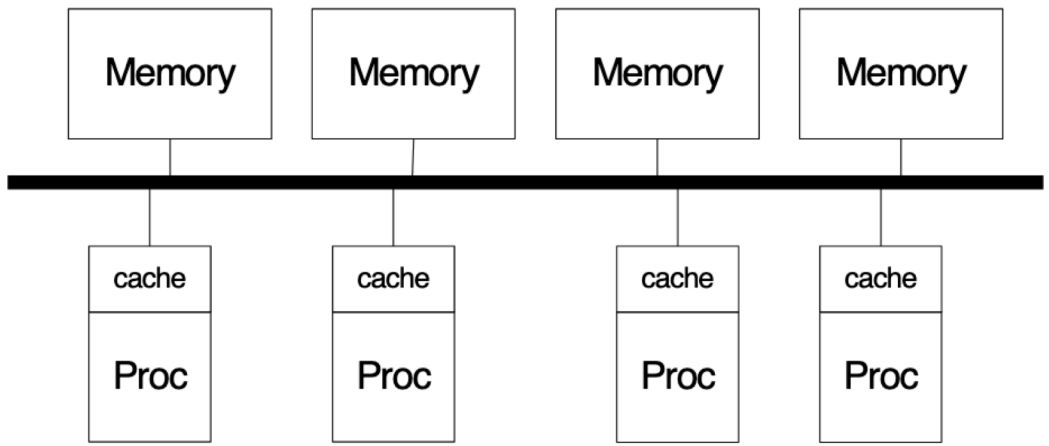
- Regular or Irregular-regular if topology is regular graph (e.g. ring, mesh)
- Routing Distance -number of links/hops along route
- Diameter -maximum routing distance
- Average Distance-average number of hops across all valid routes
- Bisection Bandwidth-Often used to describe network performance
  - Cut network in half and sum bandwidth of links severed
  - (Min # channels spanning two halves) * (BW of each channel)
  - Meaningful only for recursive topologies
  - Can be misleading, because does not account for switch and routing efficiency
- Blocking vs. Non-Blocking-If connecting any permutation of sources & destinations is possible, network is non-blocking; otherwise network is blocking.

Static networks

- Categorized by node degree
  - Degree 1: shared bus
  - Degree 2: linear array, ring
  - Degree 3: binary tree, fat tree, shuffle-exchange
  - Degree 4: two-dimensional mesh (Illiac, torus)
  - Varying degree: n-cube, n-dimensional mesh, k-ary n-cube

Bus

+ **Simple**
+ **Cost effective for a small number of nodes**
+ **Easy to implement coherence** (snooping)
- **Not scalable to large number of nodes**
  (limited bandwidth, electrical loading → reduced frequency)
- **High contention**



Crossbar

- **Every node connected to all others** (non-blocking)
- **Good for small number of nodes**

+ Low latency and high throughput

- Expensive

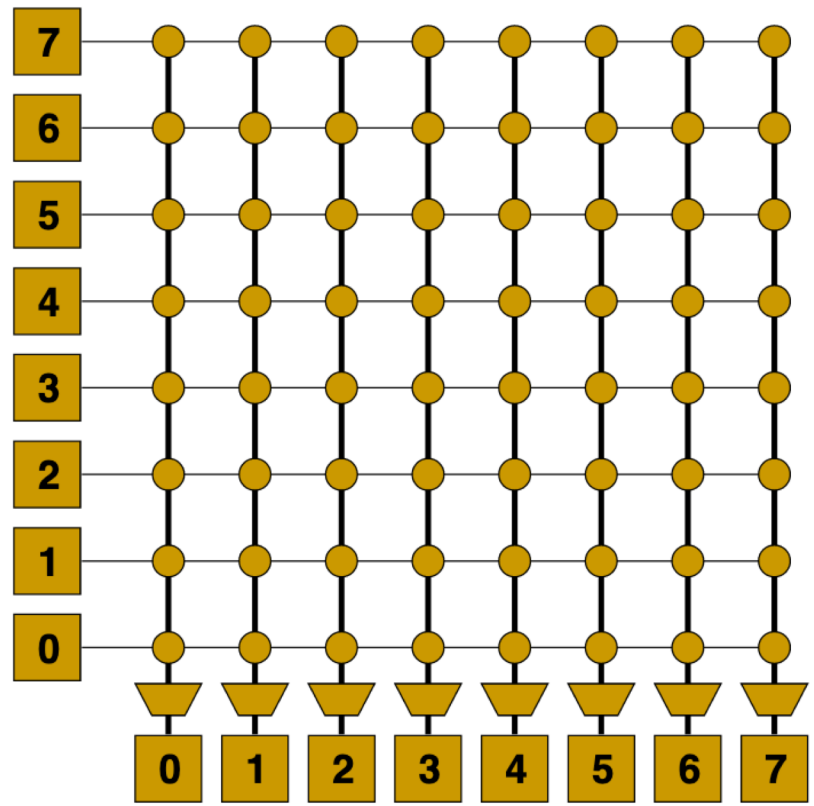
- Not scalable → O(N^2) cost

- Difficult to arbitrate

Core-to-cache-bank networks:

- IBM POWER5

- Sun Niagara I/II

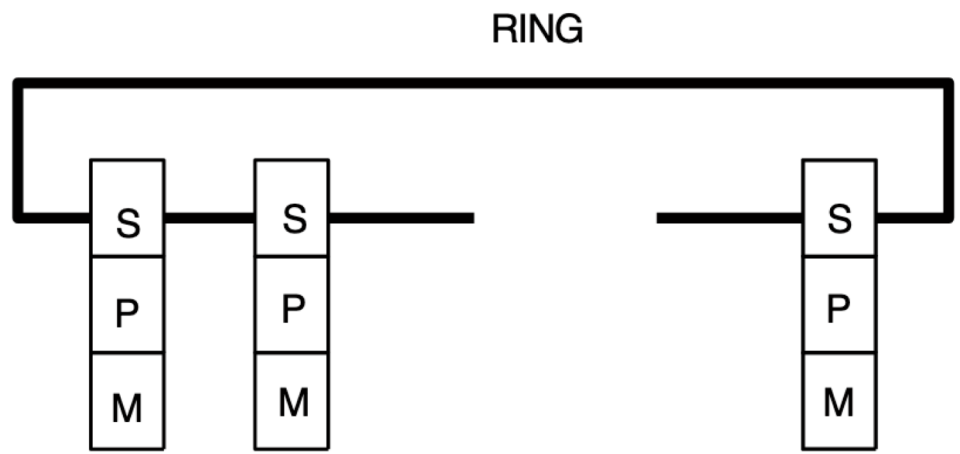


Ring

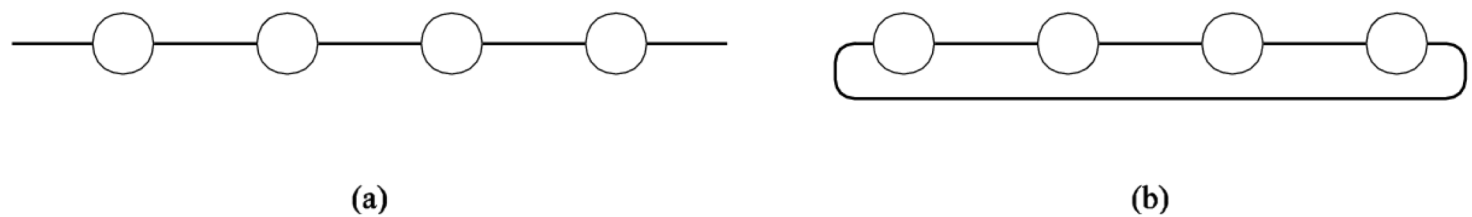
+ Cheap: O(N) cost
- High latency: O(N)
- Not easy to scale
  - Bisection bandwidth remains constant

Used in:

- Intel Larrabee/Core i7
- IBM Cell



- In a linear array, each node has two neighbors, one to its left and one to its right. If the nodes at either end are connected, we refer to it as a 1-D torus or a ring.
- A generalization to 2 dimensions has nodes with 4 neighbors, to the north, south, east, and west.
- A further generalization to *d* dimensions has nodes with *2d* neighbors.
- A special case of a *d*-dimensional mesh is a hypercube. Here, $d = \log p$, where p is the total number of nodes.

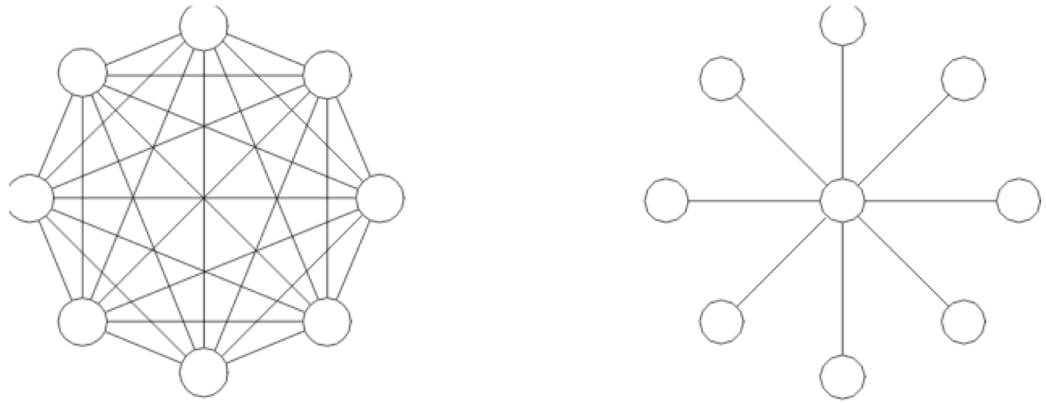


Linear arrays: (a) with no wraparound links; (b) with wraparound link.

Completely Connected and Star Connected

- Completely connected:
  - Each processor is connected to every other processor
  - The number of links in the network scales as $O(p^2)$
  - While the performance scales very well, the hardware complexity is not realizable for large values of p
  - In this sense, these networks are static counterparts of crossbars.
- Star Connected:
  - Every node is connected only to a common node at the center
  - Distance between any pair of nodes is O(1).However, the central node becomes a bottleneck
  - In this sense, star connected networks are static counterparts of buses.

Example of an 8-node completely connected network.

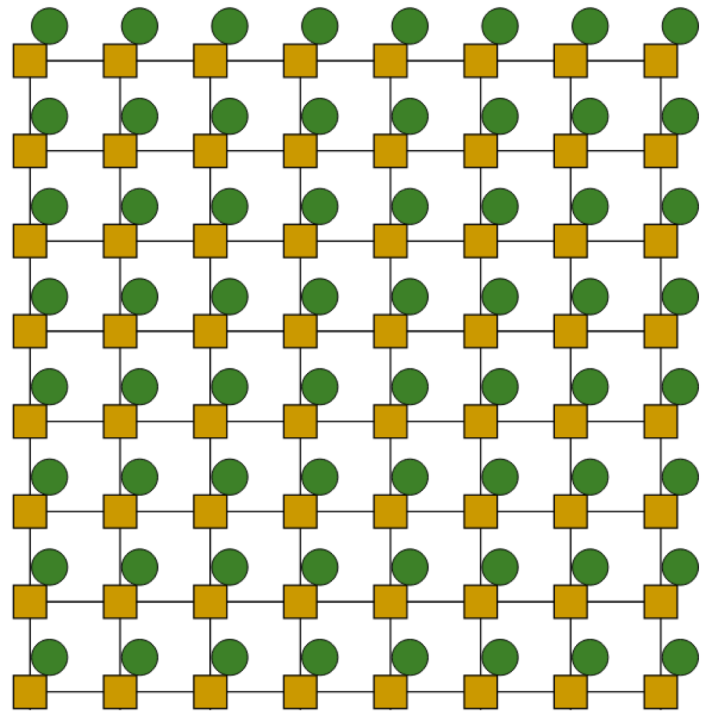


(a) A completely-connected network of eight nodes;
(b) a star connected network of nine nodes.

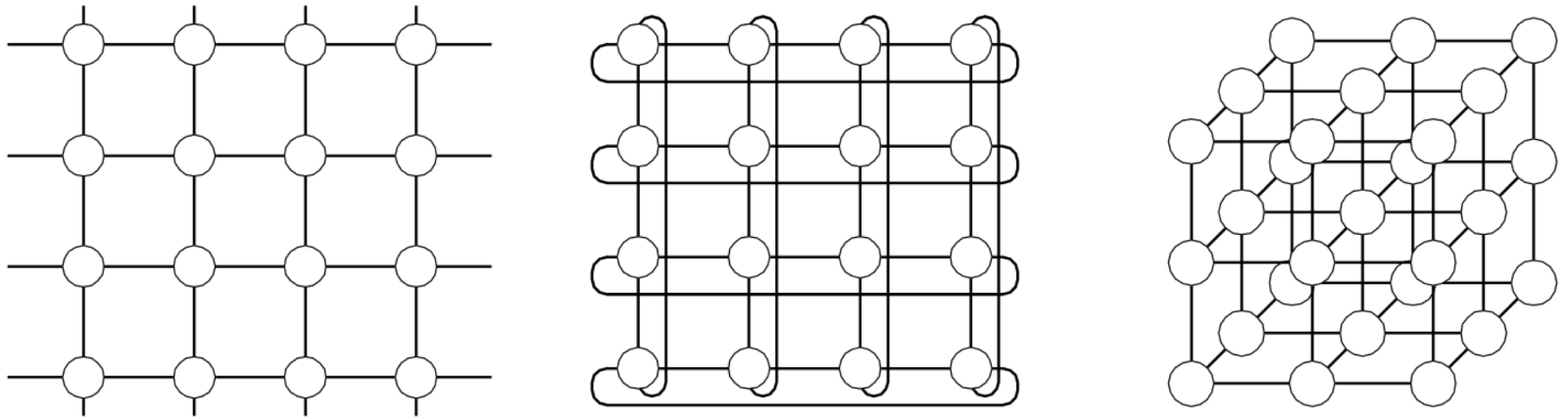
Mesh

- O(N) cost
- Average latency: O(sqrt(N))
- Easy to layout on-chip: regular & equal-length links
- Path diversity: many ways to get from one node to another

- Used in:
  - Tilera 100-core CMP
  - On-chip network prototypes

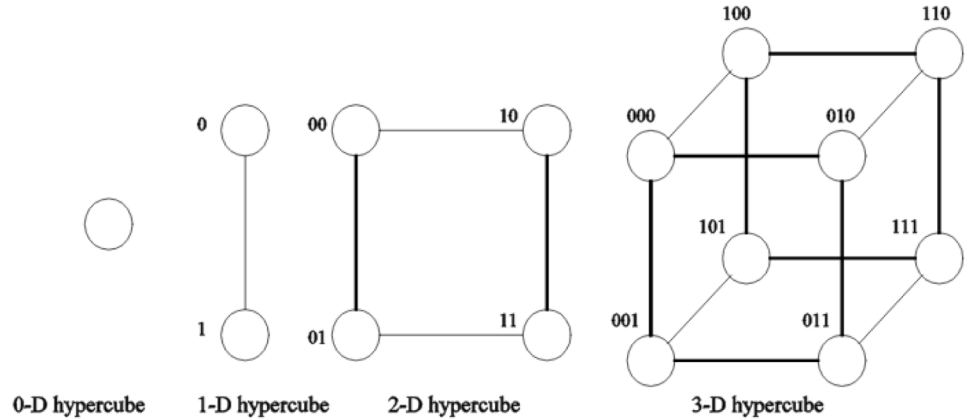


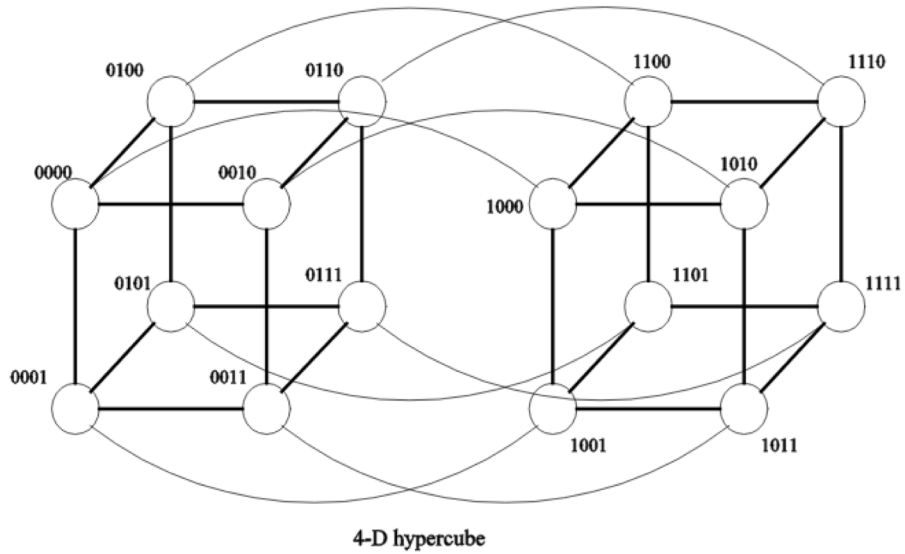
Two and Three-Dimensional Meshes



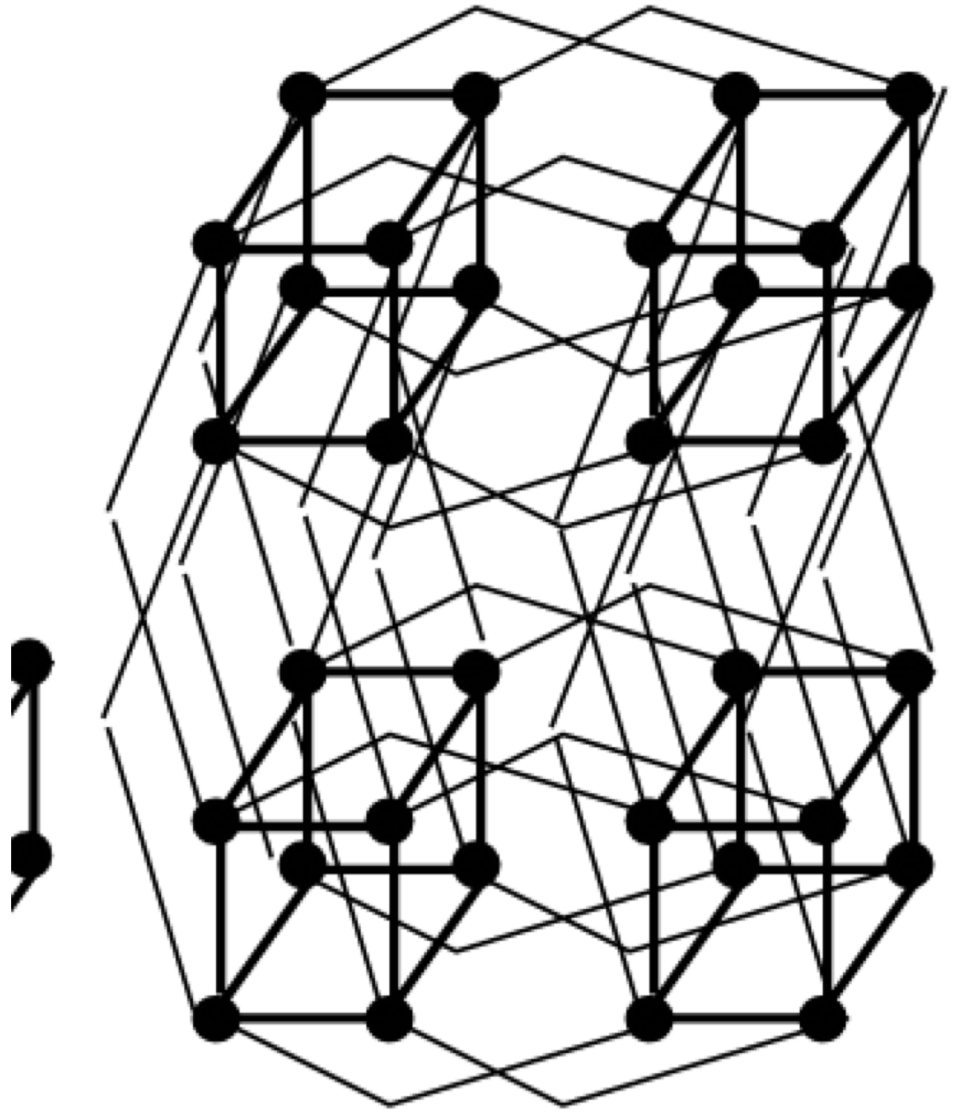
Two and three dimensional meshes: (a) 2-D mesh with no wraparound; (b) 2-D mesh with wraparound link (2-D torus); and (c) a 3-D mesh with no wraparound.

Hypercube and their construction



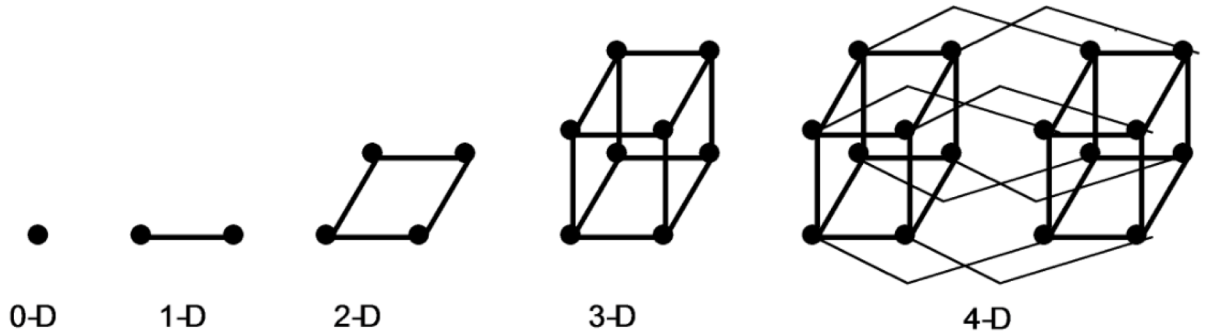


Construction of hypercubes from hypercubes of lower dimension.



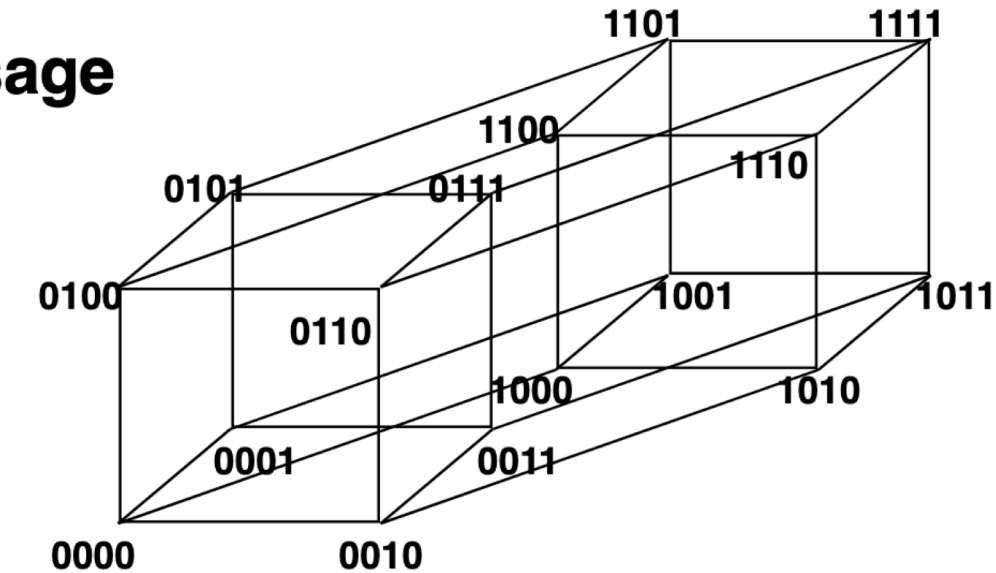
Hypercube

- Latency: O(logN)
- Radix: O(logN)
- #links: O(NlogN)



+ Low latency
- Hard to lay out in 2D/3D

- Used in some early message passing machines, e.g.:
  - Intel iPSC
  - nCube



Torus

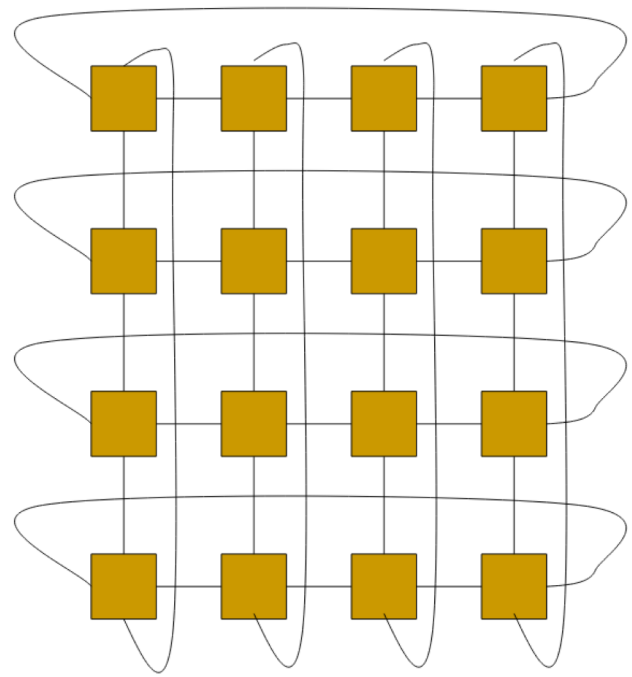
- Mesh is not symmetric on edges: performance very sensitive to placement of task on edge vs. middle
- Torus avoids this problem

+ Higher path diversity (& bisection bandwidth) than mesh

- Higher cost

- Harder to lay out on-chip

  - Unequal link lengths



Trees

Planar, hierarchical topology

Latency: O(logN)

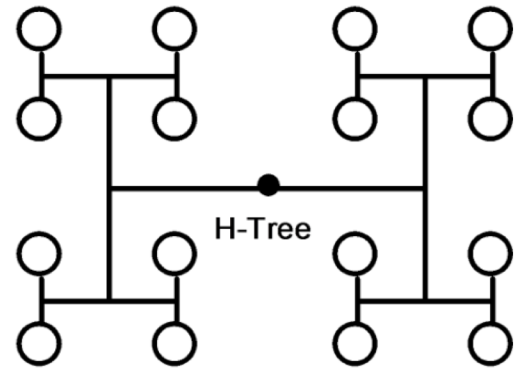
Good for local traffic

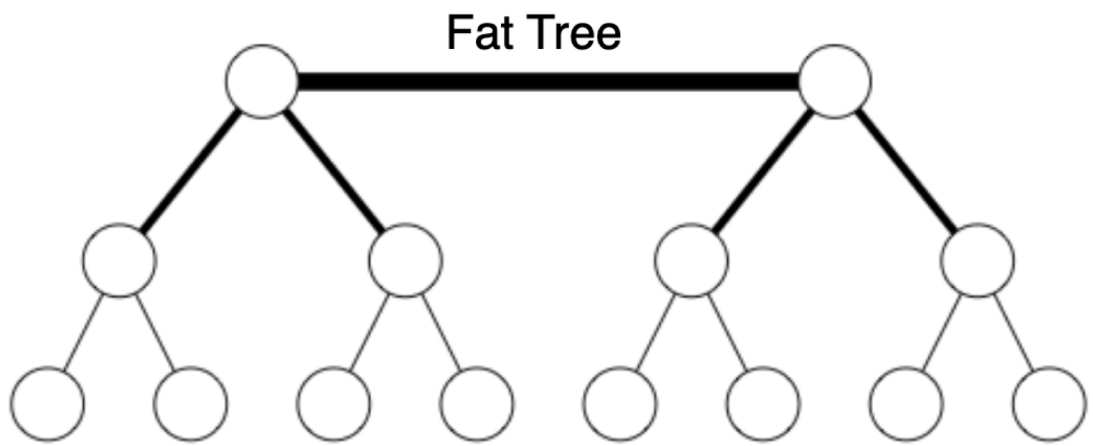
+ Cheap: O(N) cost

+ Easy to Layout

- Root can become a bottleneck

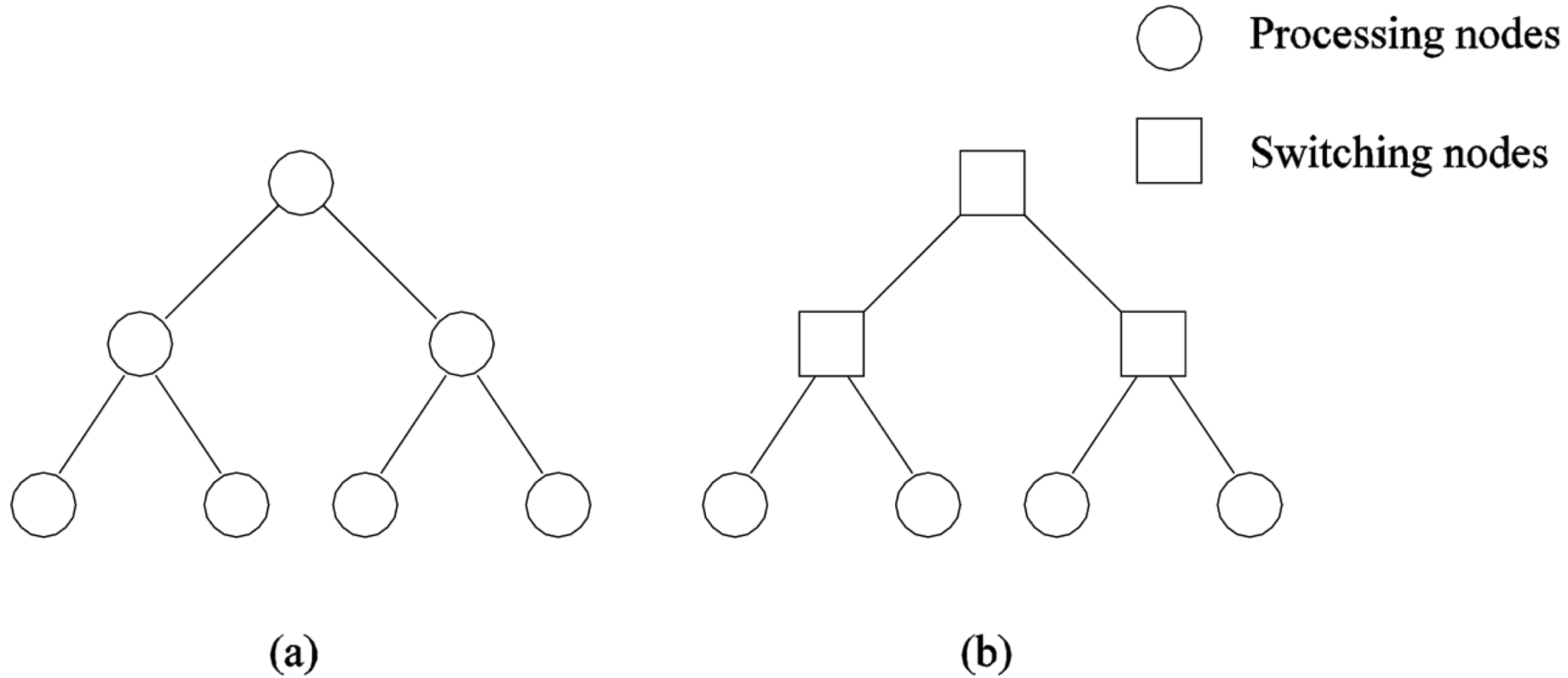
Fat trees avoid this problem (CM-5)



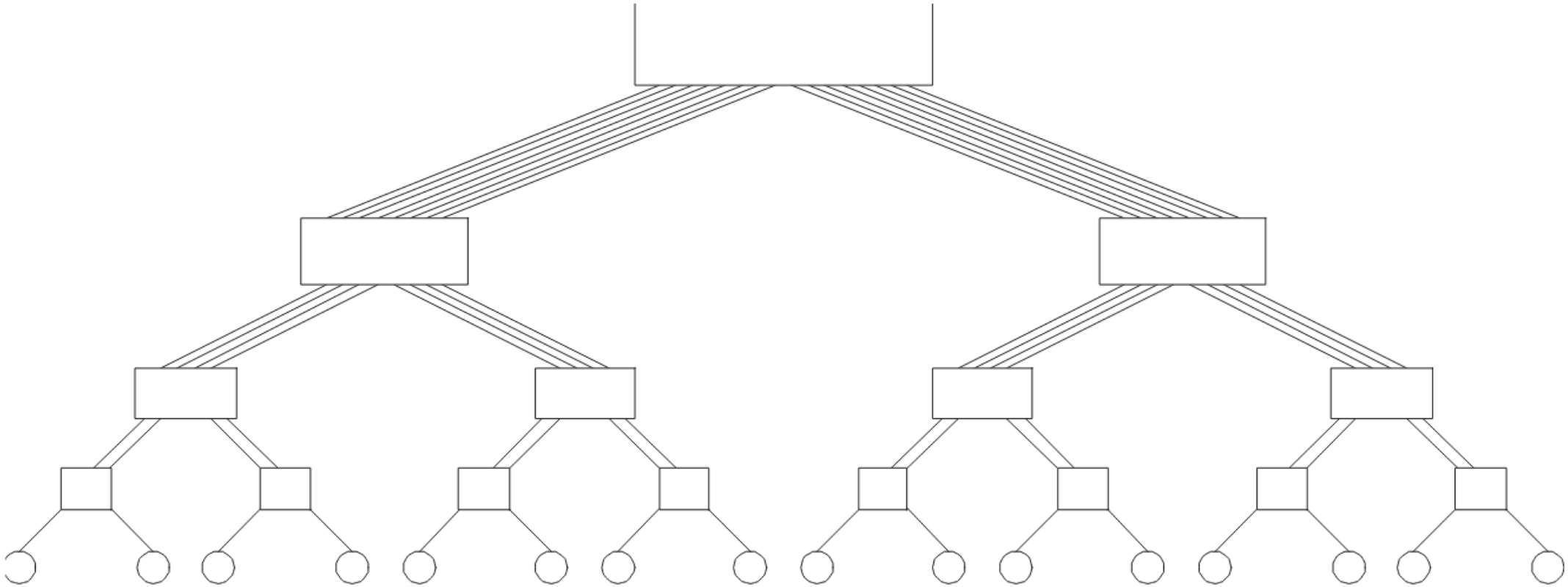


Tree-properties

- The distance between any two nodes is no more than 2logp
- Links higher up the tree potentially carry more traffic than those at the lower levels
- For this reason, a variant called a fat-tree, fattens the links as we go up the tree
- Trees can be laid out in 2D with no wire crossings. This is an attractive property of trees.



omplete binary tree networks: (a) a static tree network; and (b) a dynamic tree network.



A fat tree network of 16 processing nodes.

Evaluating Static Interconnection Network

Network	Diameter	Bisection Width
Completely-connected	1	$p^2/4$
Star	2	1
Complete binary tree	$2\log((p+1)/2)$	1
Linear array	$p-1$	1
2-D mesh, no wraparound	$2(\sqrt{p}-1)$	$\sqrt{p}$
2-D wraparound mesh	$2\lfloor\sqrt{p}/2\rfloor$	$2\sqrt{p}$
Hypercube	$\log p$	$p/2$
Wraparound *k*-ary *d*-cube	$d\lfloor k/2\rfloor$	$2k^{d-1}$

Topology	Degree	Diameter	Ave Dist	Bisection
1D Array	2	N-1	N / 3	1
1D Ring	2	N/2	N/4	2
2D Mesh	4	$2 (N^{1/2} - 1)$	$2/3\ N^{1/2}$	$N^{1/2}$
2D Torus	4	$N^{1/2}$	$1/2\ N^{1/2}$	$2N^{1/2}$
k-ary n-cube	2n	nk/2	nk/4	nk/4
Hypercube	n =log N		n	n/2

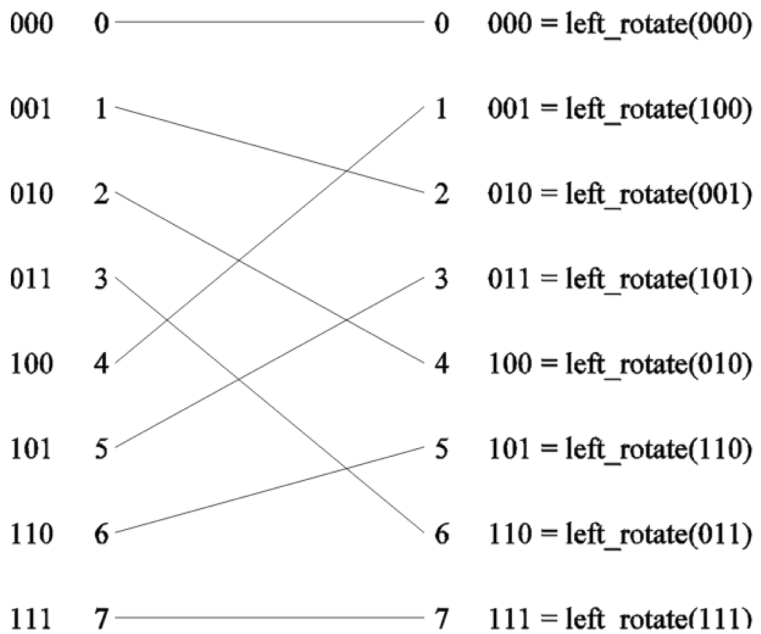
Multistage networks

- Crossbars have excellent performance scalability but poor cost scalability
- Buses have excellent cost scalability, but poor performance scalability
- Multistage interconnects strike a compromise between these extremes
- One of the most commonly used multistage interconnects is the Omega network
- This network consists of log p stages, where p is the number of inputs/outputs
- At each stage, input i is connected to output j if:

$$j = \begin{cases} 2i, & 0 \leq i \leq p/2 - 1 \\ 2i + 1 - p, & p/2 \leq i \leq p - 1 \end{cases}$$

Multistage Omega Network

Each stage of the Omega network implements a perfect shuffle as follows:



A perfect shuffle interconnection for eight inputs and outputs.

Multistage Omega Network

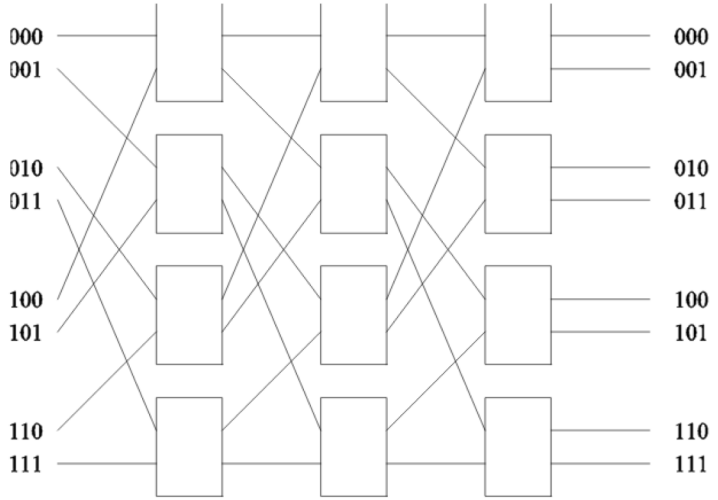
- The perfect shuffle patterns are connected using 2×2 switches.
- The switches operate in two modes – crossover or passthrough.

(a) (b)

Two switching configurations of the 2 × 2 switch:
(a) Pass-through; (b) Cross-over.

Multistage Omega Network

A complete Omega network with the perfect shuffle interconnects and switches can now be illustrated:



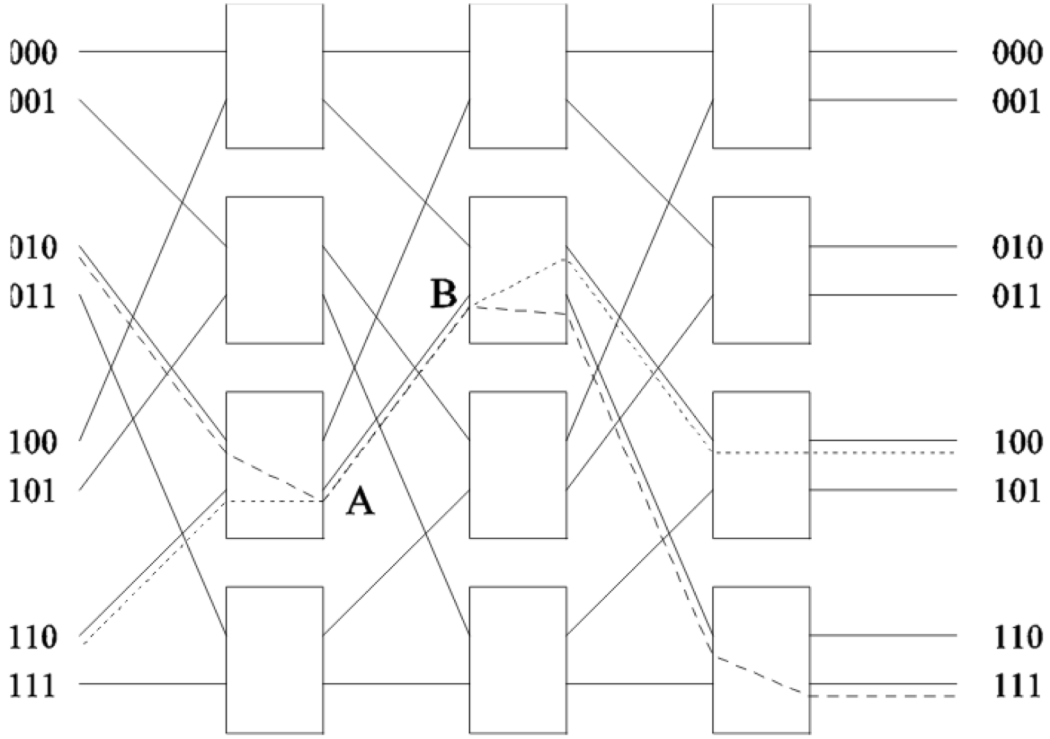
A complete omega network connecting eight inputs and eight outputs.

An omega network has $p/2 \times \log p$ switching nodes, and the cost of such a network grows as $(p \log p)$.

Multistage Omega Network -- Routing

- Let *s* be the binary representation of the source and *d* be that of the destination processor.
- The data traverses the link to the first switching node. If the most significant bits of *s* and *d* are the same, then the data is routed in pass-through mode by the switch else, it switches to crossover.
- This process is repeated for each of the *log p* switching stages.
- Note that this is not a non-blocking switch.

Multistage Omega Network -- Routing



An example of blocking in omega network: one of the messages (010 to 111 or 110 to 100) is blocked at link AB.

Evaluating Static Interconnection Network

- *Diameter:* The distance between the farthest two nodes in the network. The diameter of a linear array is $p - 1$, that of a mesh is $2(\sqrt{p} - 1)$, that of a tree and hypercube is $\log p$, and that of a completely connected network is $O(1)$.
- *Bisection Width:* The minimum number of wires you must cut to divide the network into two equal parts. The bisection width of a linear array and tree is *1*, that of a mesh is $\sqrt{p}$, that of a hypercube is $p/2$ and that of a completely connected network is $p^2/4$.
- *Cost:* The number of links or switches (whichever is asymptotically higher) is a meaningful measure of the cost. However, a number of other factors, such as the ability to layout the network, the length of wires, etc., also factor in to the cost.

Evaluating Dynamic Interconnection Network

Network	Diameter	Bisection Width	Arc Connectivity	Cost (No. of links)
Crossbar	1	p	1	p^2
Omega Network	$\log p$	$p/2$	2	$p/2$
Dynamic Tree	$2 \log p$	1	2	$p - 1$

Network characterization

- Topology (what)
  - physical interconnection structure of the network graph
  - direct vs indirect
- Routing Algorithm (which)
  - restricts the set of paths that msgs may follow
- Switching Strategy (how)
  - how data in a msg traverses a route
  - circuit switching vs. packet switching
- Flow Control Mechanism (when)
  - when a msg or portions of it traverse a route
  - what happens when traffic is encountered?
- *Interplay of all of these determines performance*