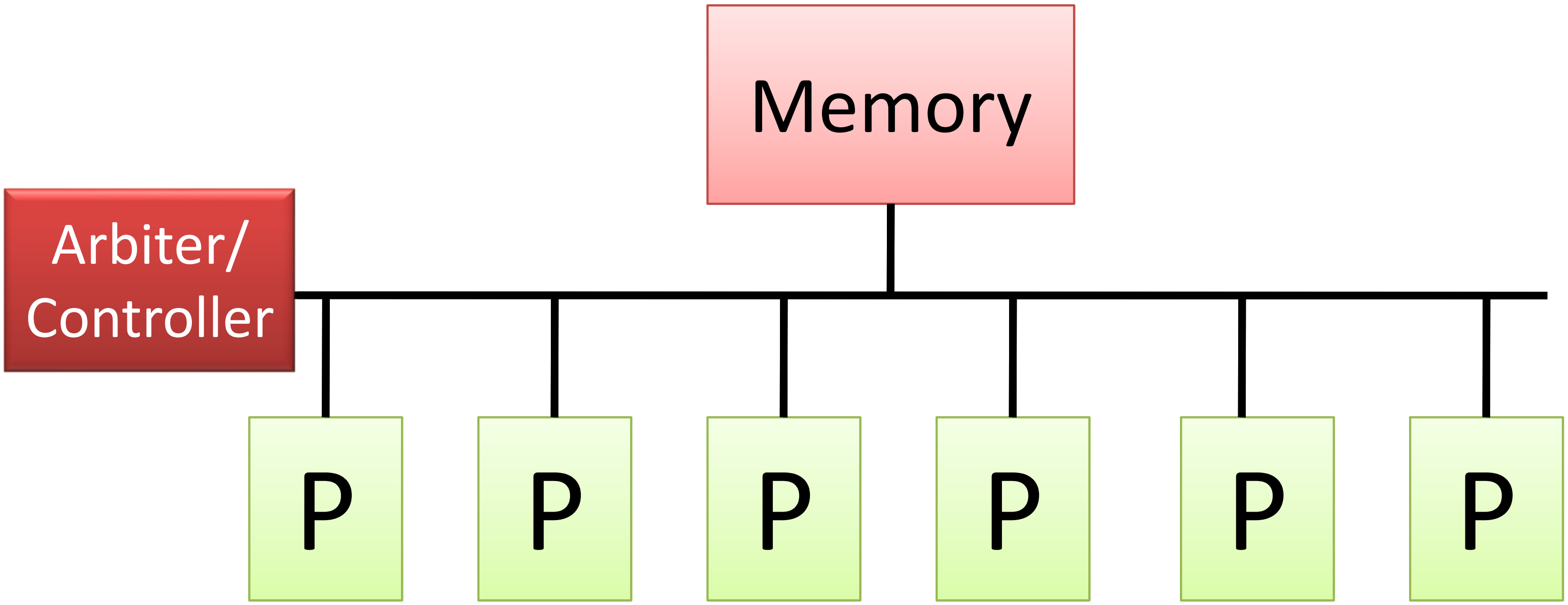
CS528
Interconnection Network of HPC

A Sahu
Dept of CSE, IIT Guwahati

Outline

- Multi-node Architecture
  - Static Network: Parameters and Performance
  - Dynamic Network
  - Interconnection and Topology Embedding
- Amdhal’s Law
- **Cilk**

Bus interconnection/Shared Memory



Switched Networks

BUS	Switched Network
• Shared media	• Switched paths
• Lower Cost	• Higher cost
• Lower throughput	• Higher throughput
• Scalability poor	• Scalability better

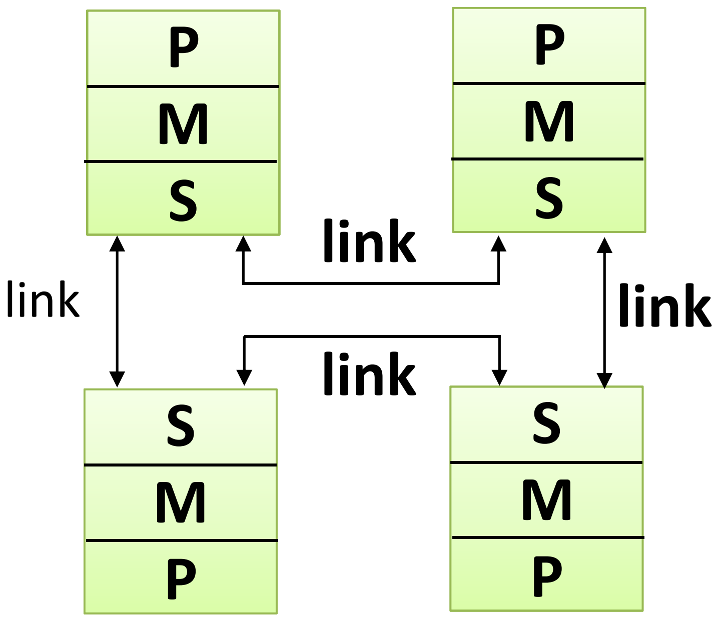
Interconnection Networks

- Topology : who is connected to whom ?
- Direct / Indirect : where is switching done ?
- Static / Dynamic : when is switching done ?
- Circuit switching / packet switching : how are connections established ?

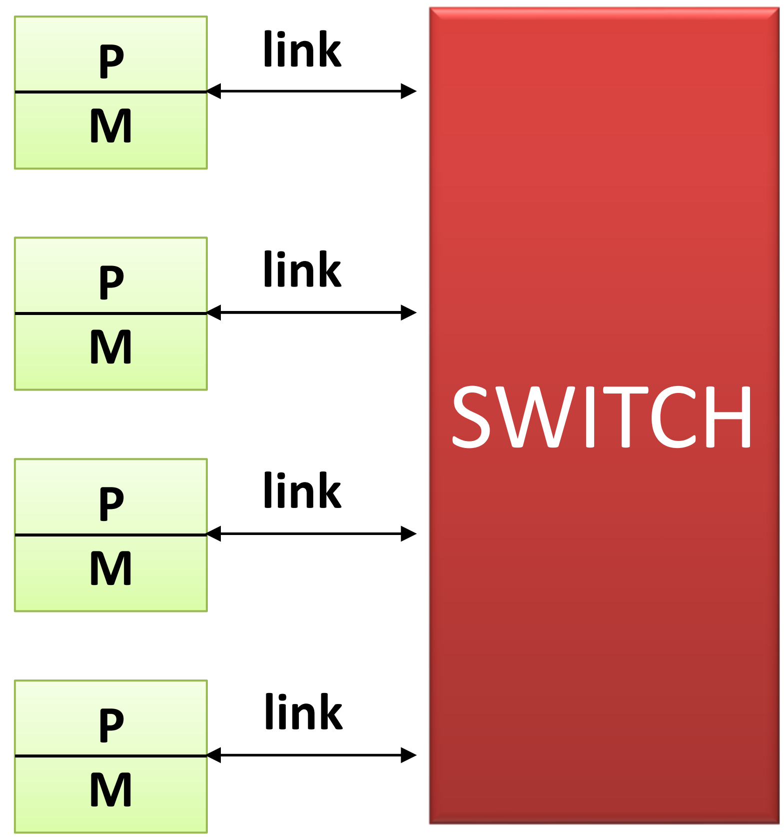
Interconnection Networks

- Store & forward / worm hole routing : how is the path determined ?
- Centralized / distributed : how is switching controlled ?
- Synchronous/asyn : mode of operation?

Direct and Indirect Networks



DIRECT



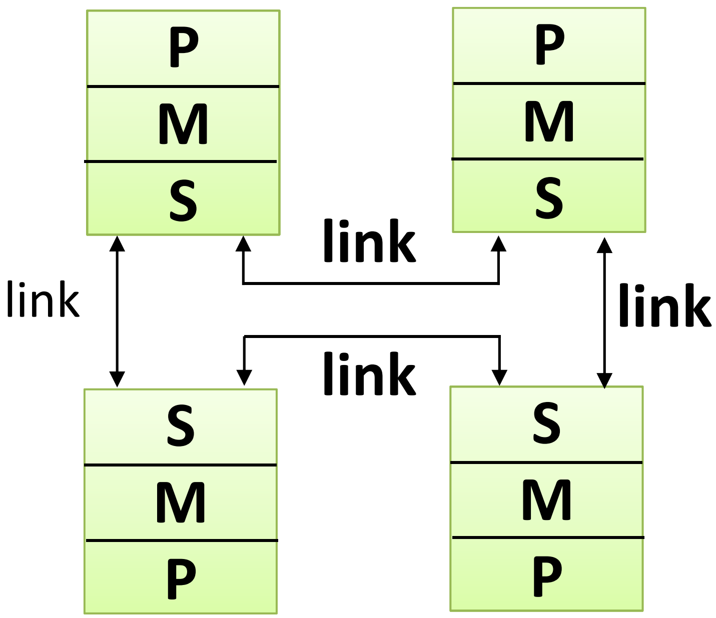
INDIRECT

Static and Dynamic Networks

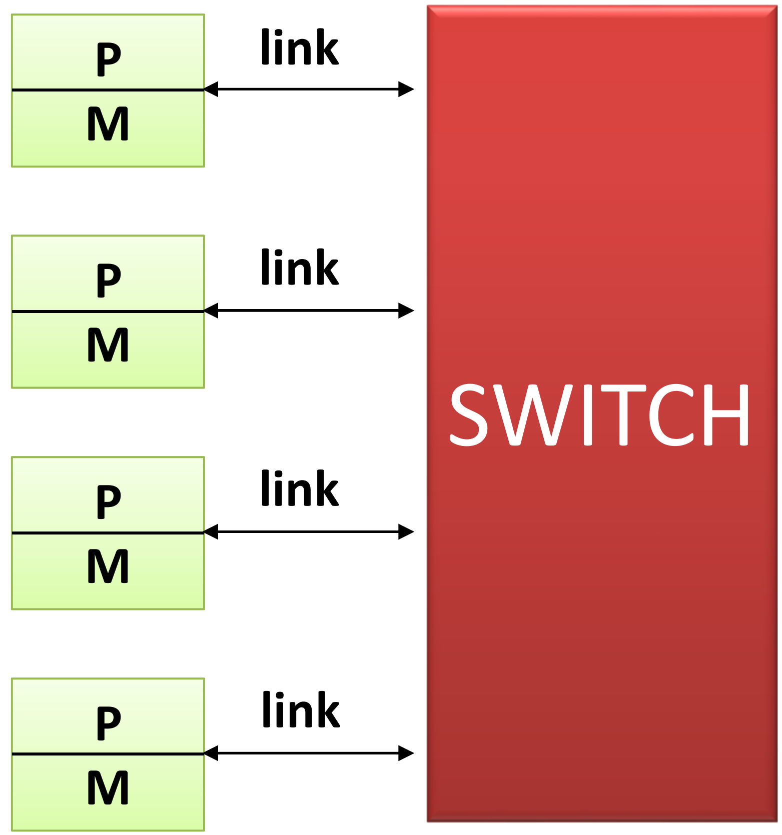
- Static Networks
  - fixed point to point connections
  - usually direct
  - each node pair may not have a direct connection
  - routing through nodes
- Dynamic Networks
  - connections established as per need
  - usually indirect
  - path can be established between any pair of nodes
  - routing through switches

Dynamic Network

Direct and Indirect Networks

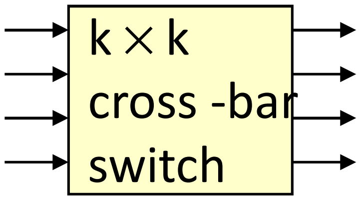


DIRECT



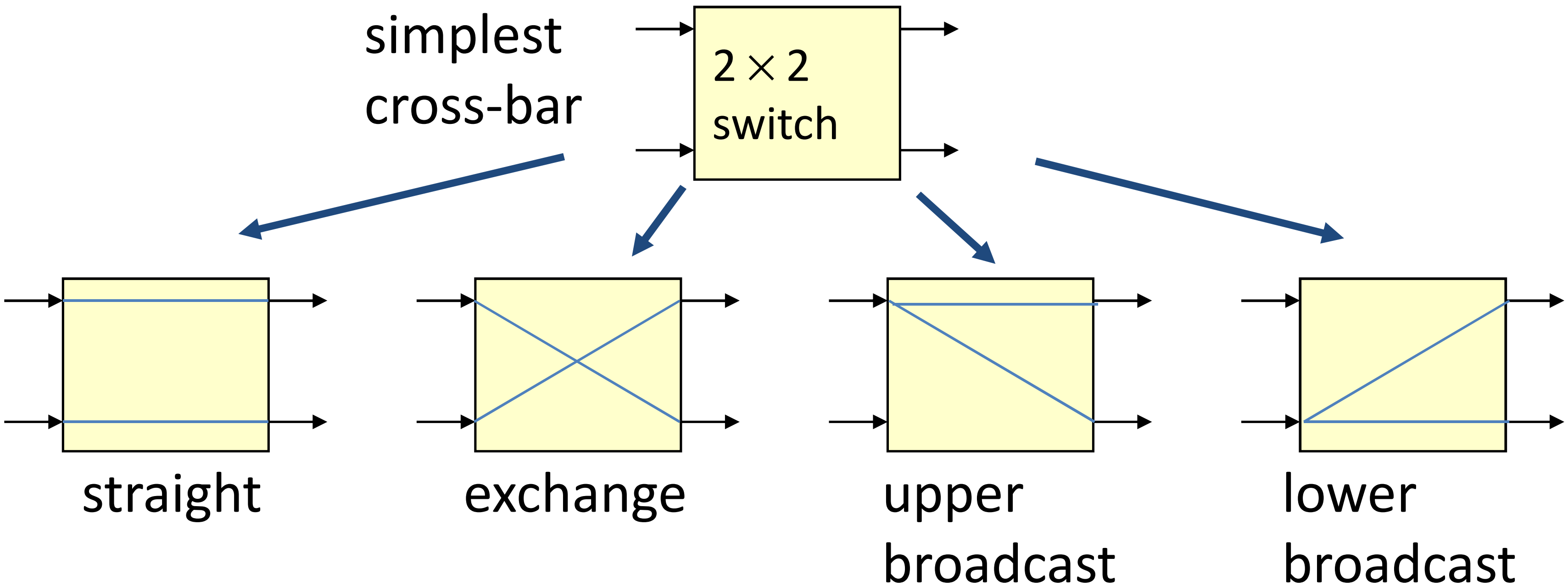
INDIRECT/Dynamic

Dynamic Networks

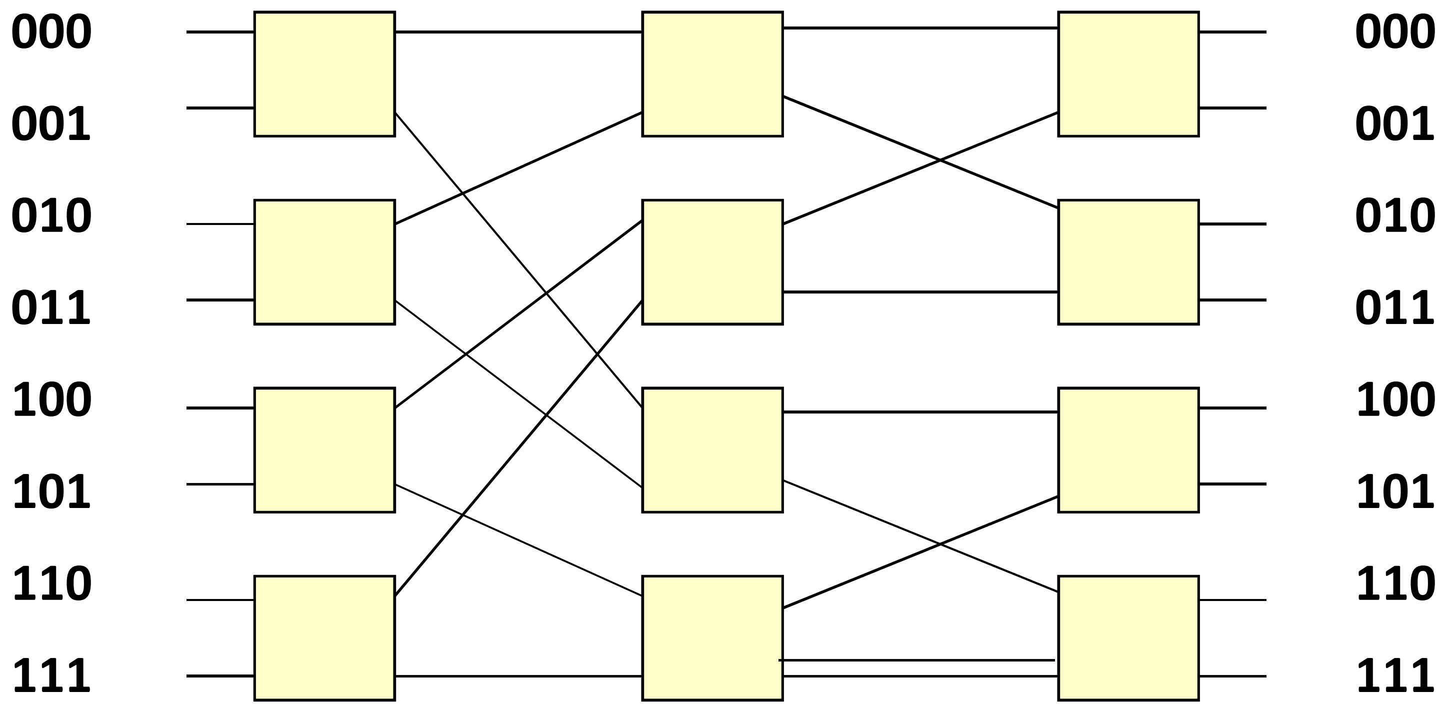


building block for multi-stage dynamic networks

simplest cross-bar



Baseline Network

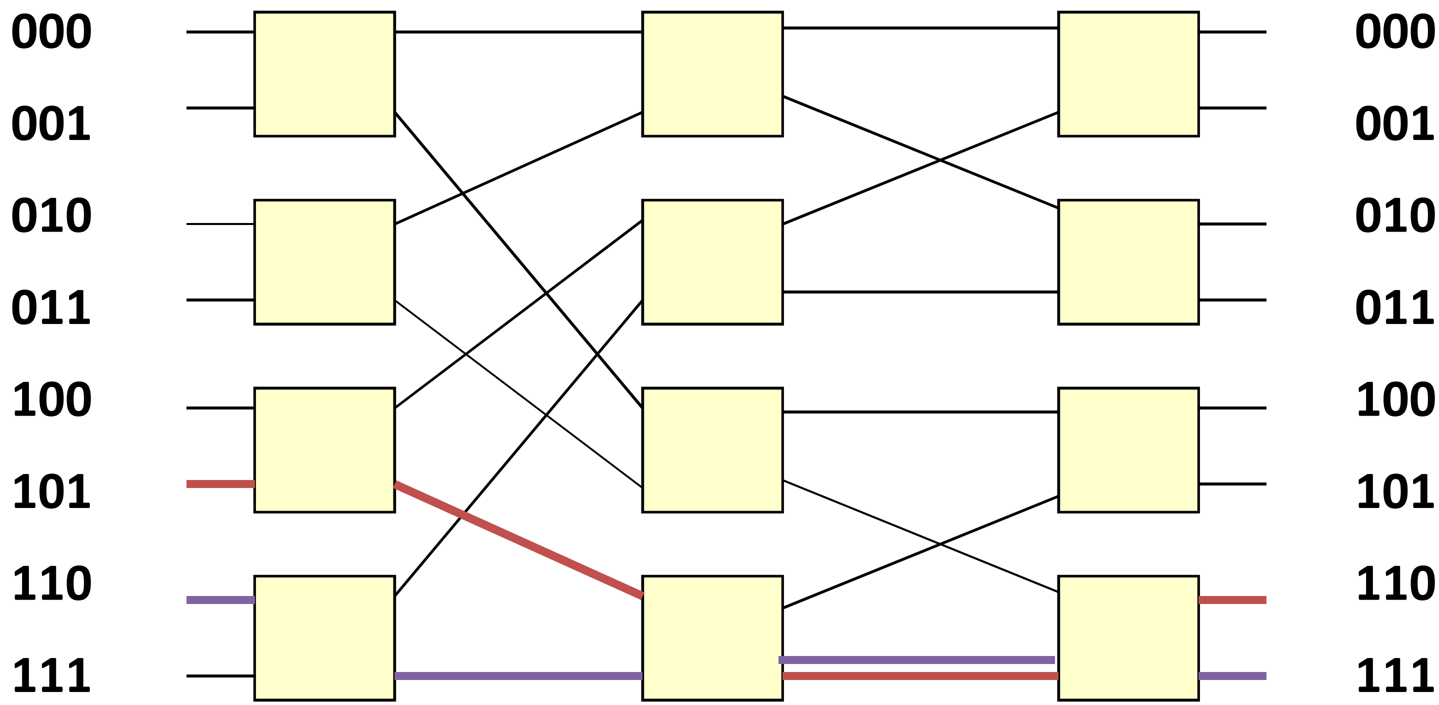


blocking can occur

Diameter=Num Stage=$\text{Log}_K N$

24 links but 8C2=28

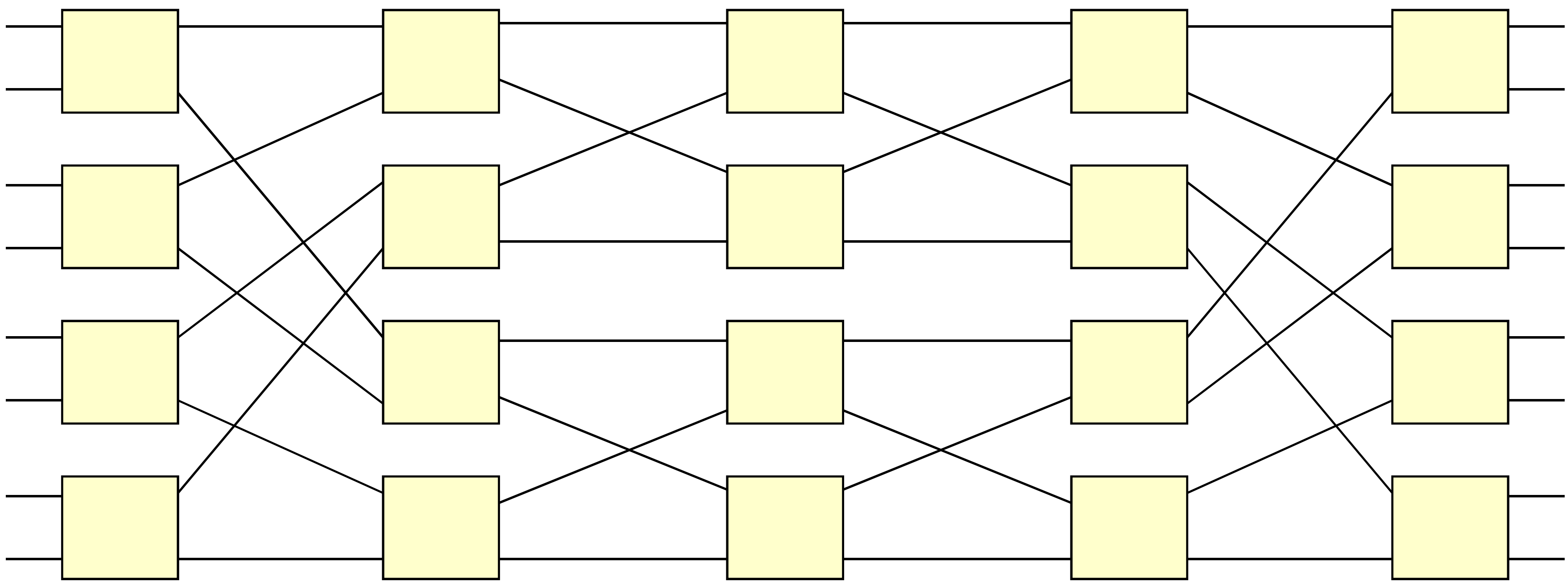
Baseline Network : Blocking



blocking can occur

Benes Network

non-blocking

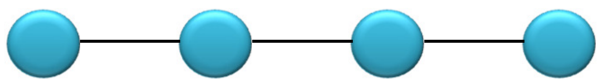


Diameter=Num Stage=$2\text{Log}_K N-1$

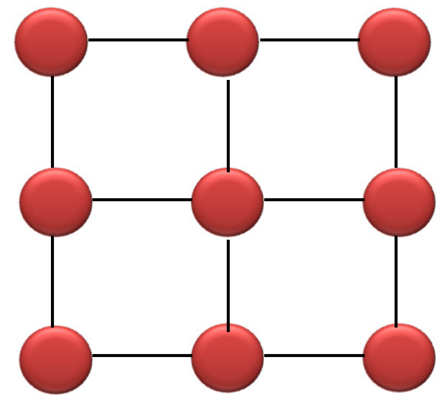
Static Network

Static Network Topologies

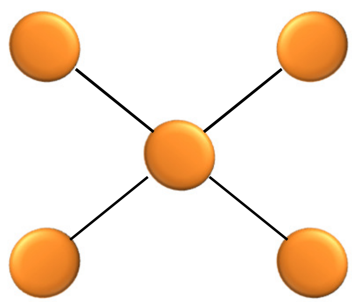
Non-uniform connectivity



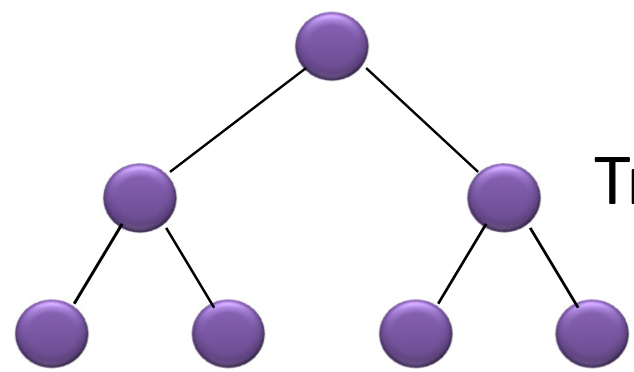
Linear



2D-Mesh



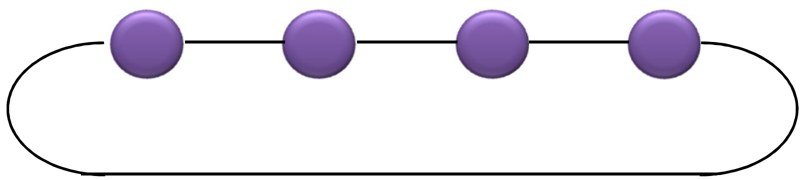
Star



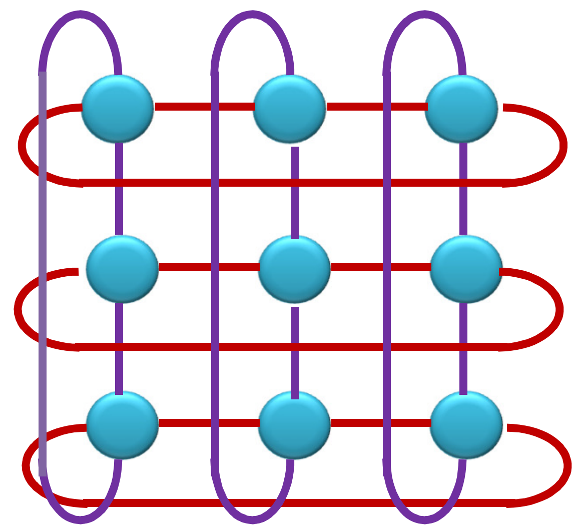
Tree

Static Networks Topologies- contd.

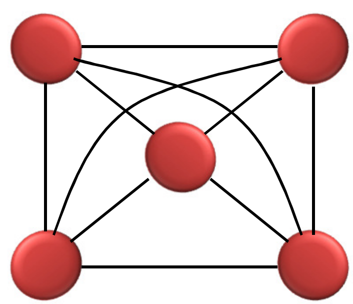
Uniform connectivity



Ring

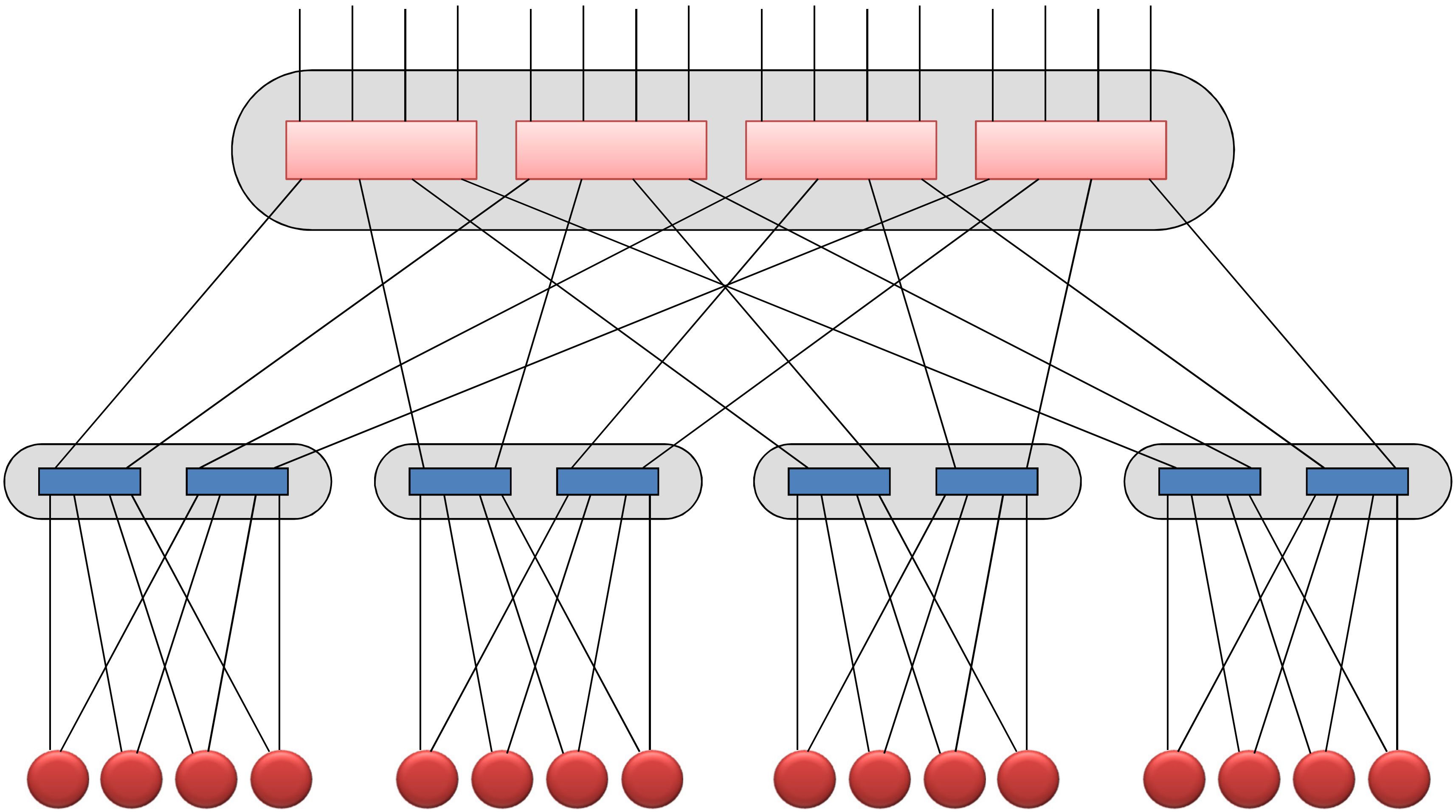


Torus



Fully Connected

Fat Tree Network



Switch / Network Topology

Quality of Topology based on:

- **Degree**: number of links from a node
- **Diameter**: max number of links crossed between two nodes
- **Average distance**: number of links to random destination

Switch / Network Topology

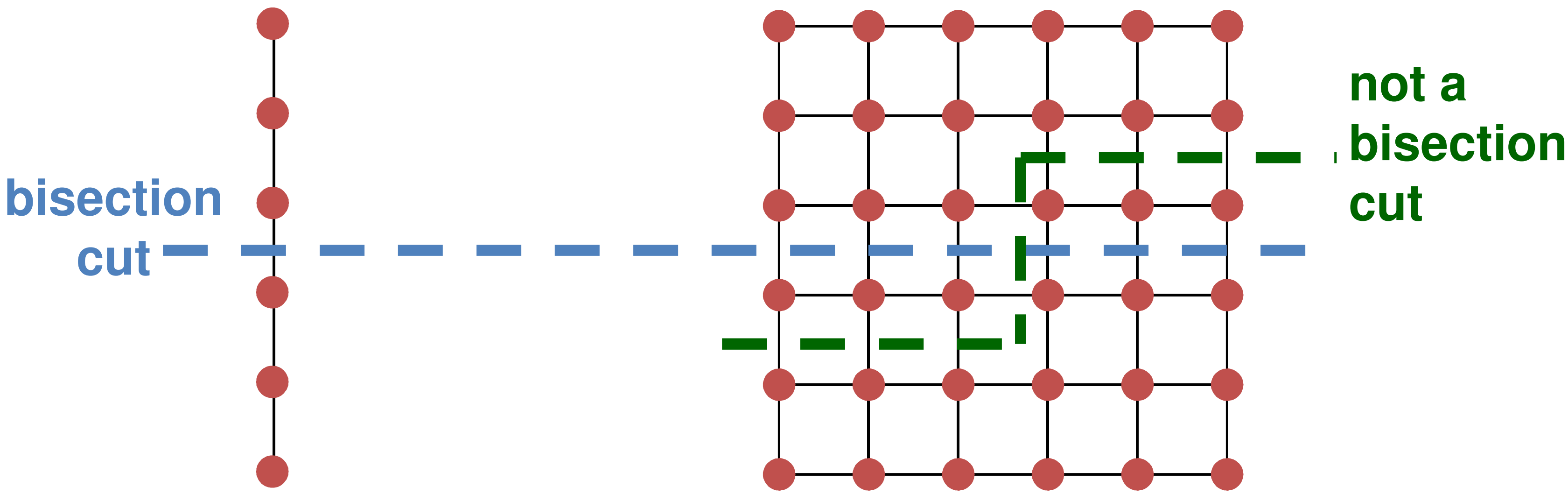
Quality of Topology based on:

- **Bisection**: minimum number of links that separate the network into two halves

- **Bisection bandwidth** = link bandwidth * bisection

Bisection Bandwidth

- Bandwidth across smallest cut that divides network into two equal halves
- Bandwidth across "narrowest" part of the network

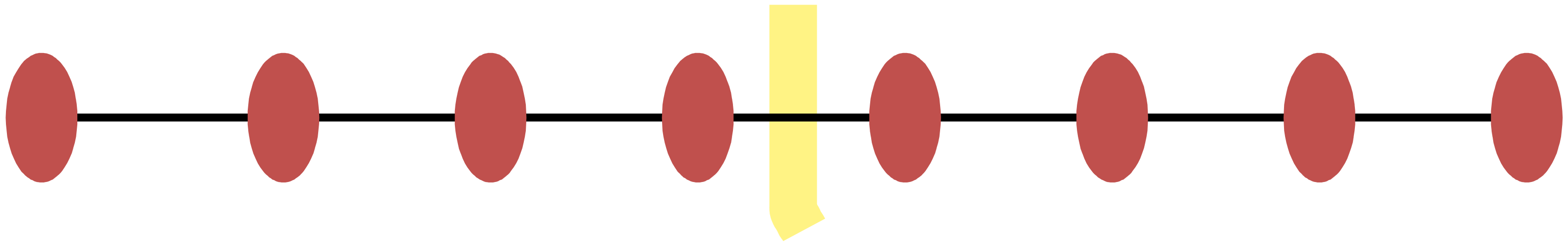


bisection bw= link bw

bisection bw = sqrt(n) * link bw

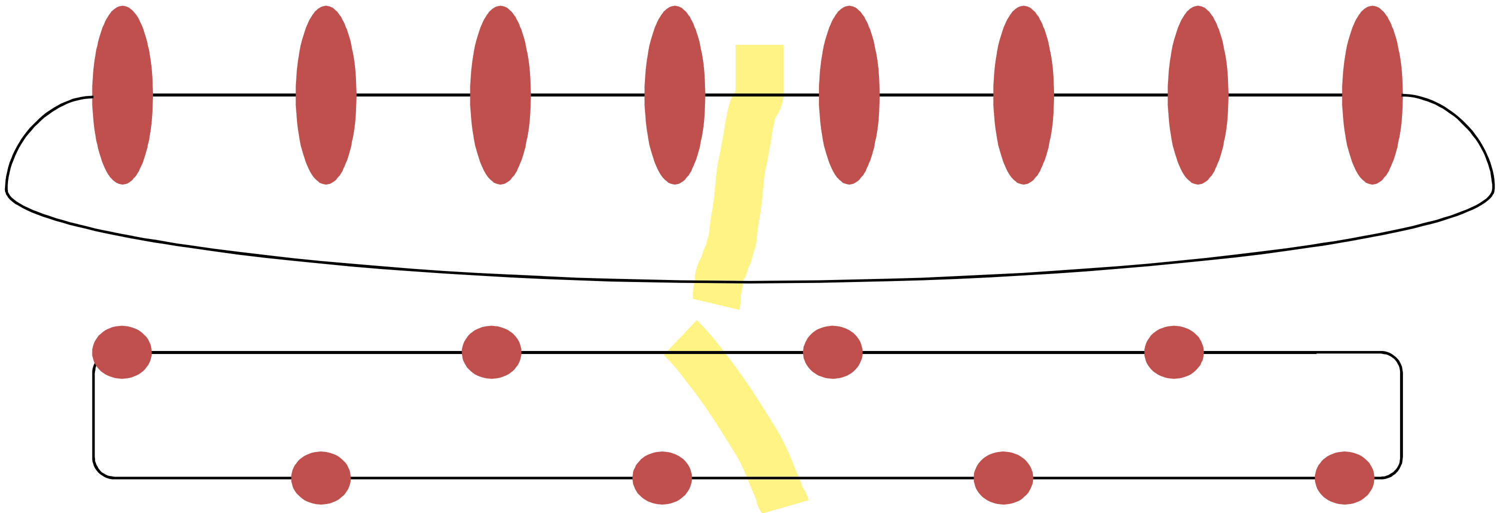
- BB important for algorithms in which all processors need to communicate with all others

Linear Array



- Diameter = n-1
- Average distance ~n/3
- Bisection bandwidth = 1 (in units of link bandwidth)

Ring /Ring Torus



- Diameter = n/2
- Average distance = n/4
- Bisection bandwidth = 2
- Natural for Algo that work with 1D arrays

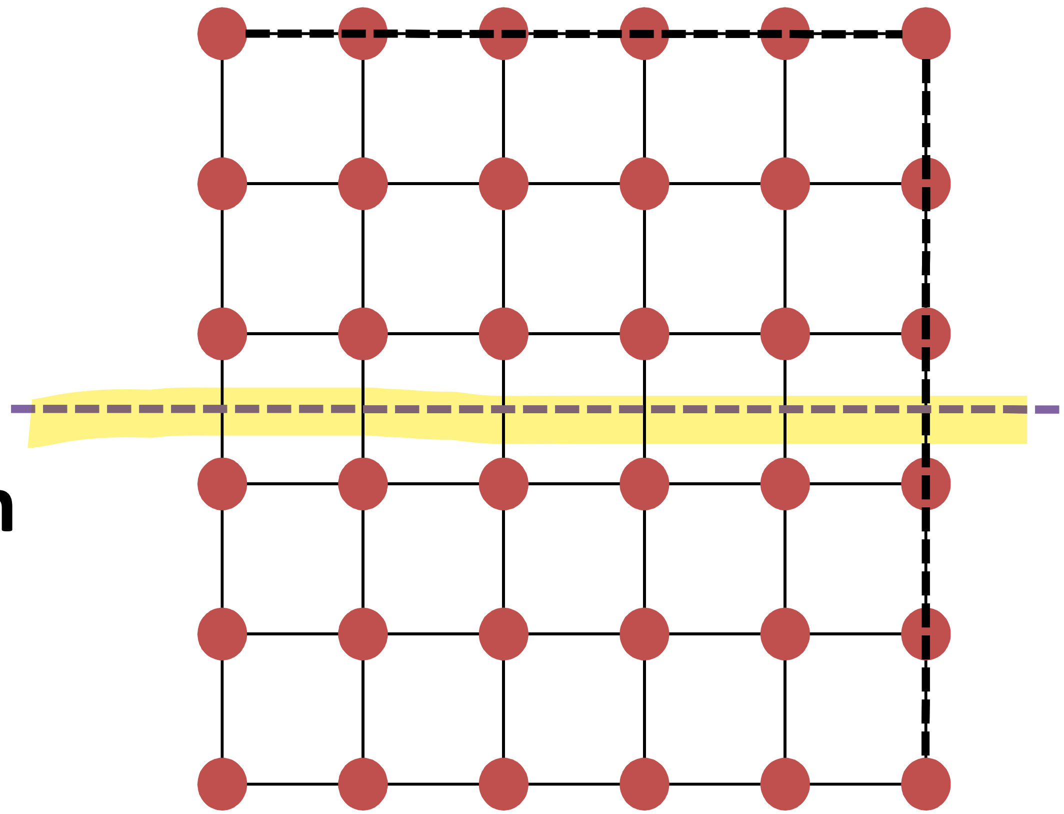
Meshes

- Diameter

  $= 2*(\text{sqrt}(n) - 1)$

- Bisection Bandwidth

  $= \text{sqrt}(n)$

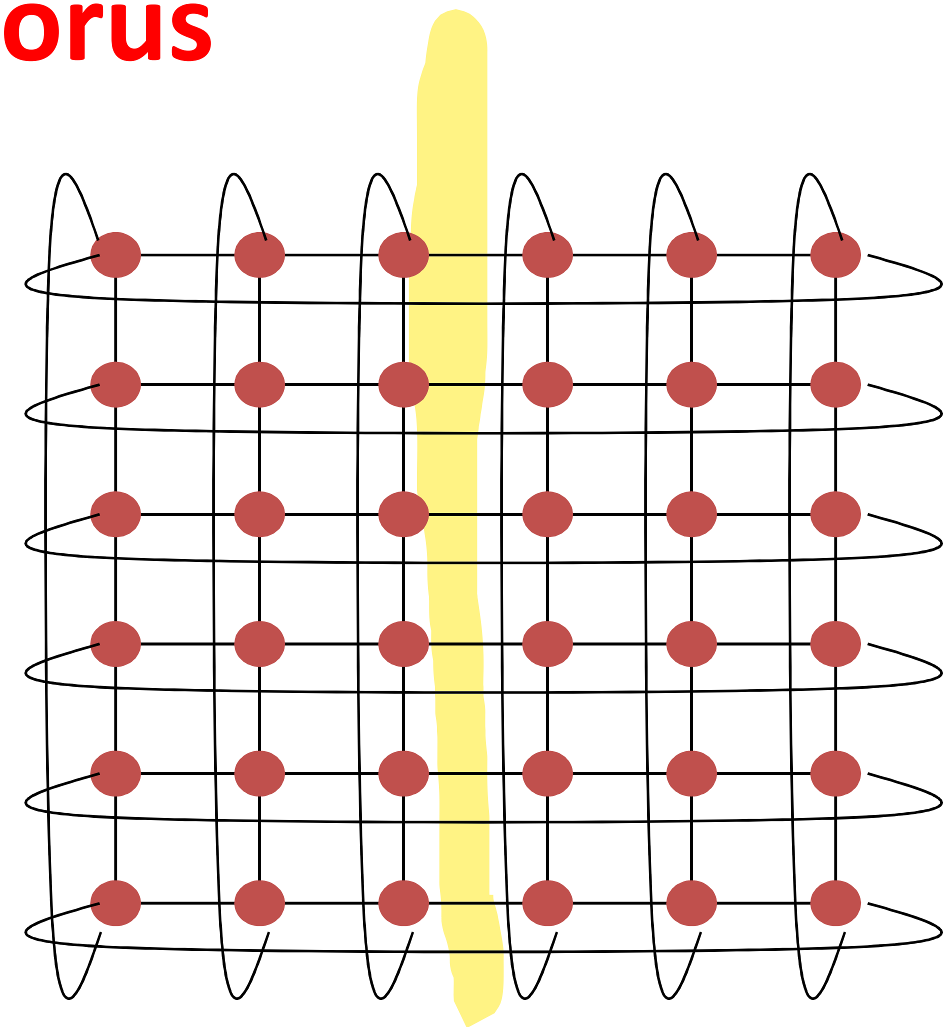


- Generalizes to higher dimensions
- Natural for algorithms that work with 2D and/or 3D arrays

2D Torus

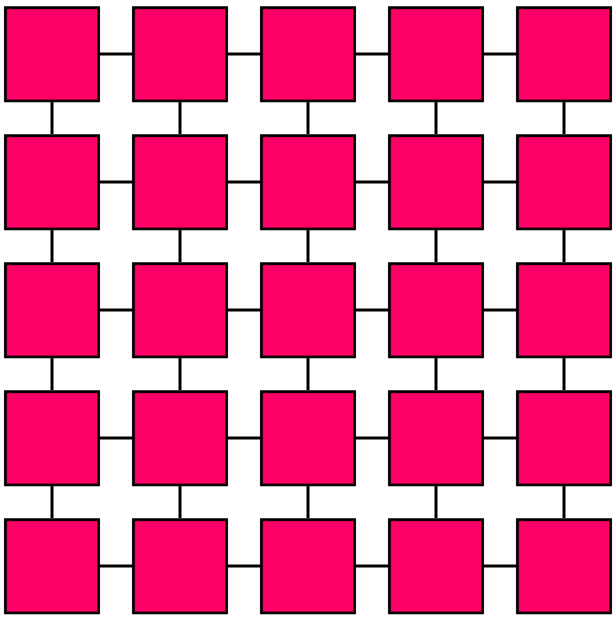
Two dimensional torus

- **Diameter = sqrt(n)**
- **Bisection BW = 2*sqrt(n)**



- **Generalizes to higher dimensions**
- **Natural for algorithms that work with 2D and/or 3D arrays**

Mesh/Torus



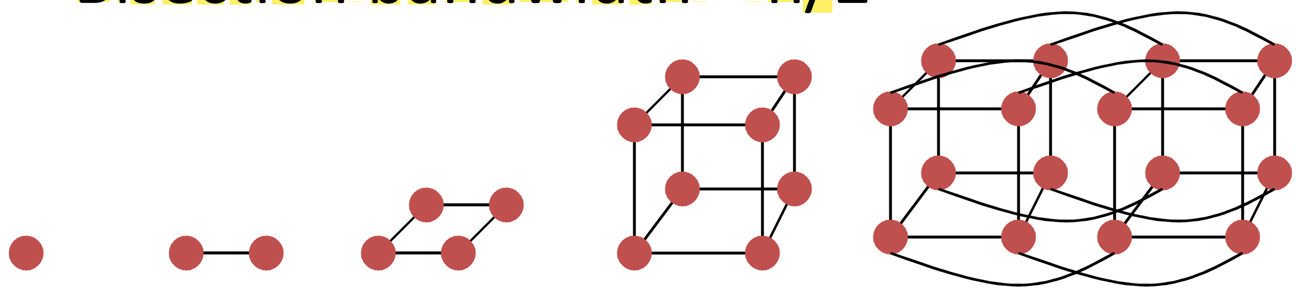
2D mesh

Diameter $\Theta(\sqrt{n})$

Bisection width $\Theta(\sqrt{n})$

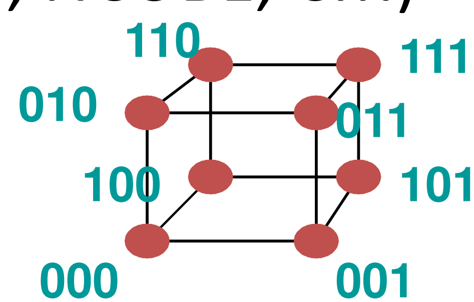
Hyper-cubes

- Number of nodes n = 2d   for dimension d
  - Diameter = d = Log(N)
  - Bisection bandwidth = n/2

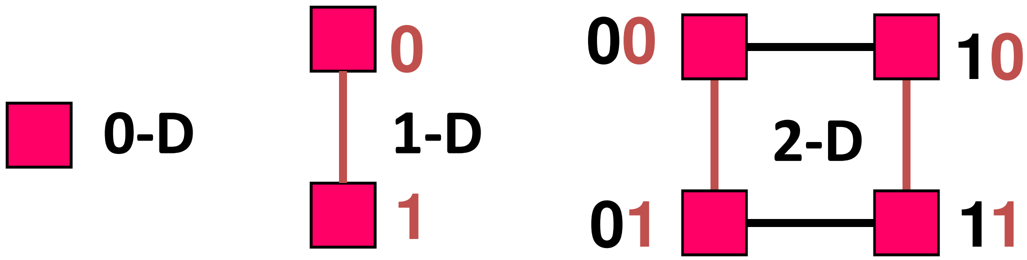


0d 1d 2d 3d 4d

- Popular in early machines (Intel iPSC, NCUBE, CM)
- Grey code addressing:
  - Each node connected to others with 1 bit different

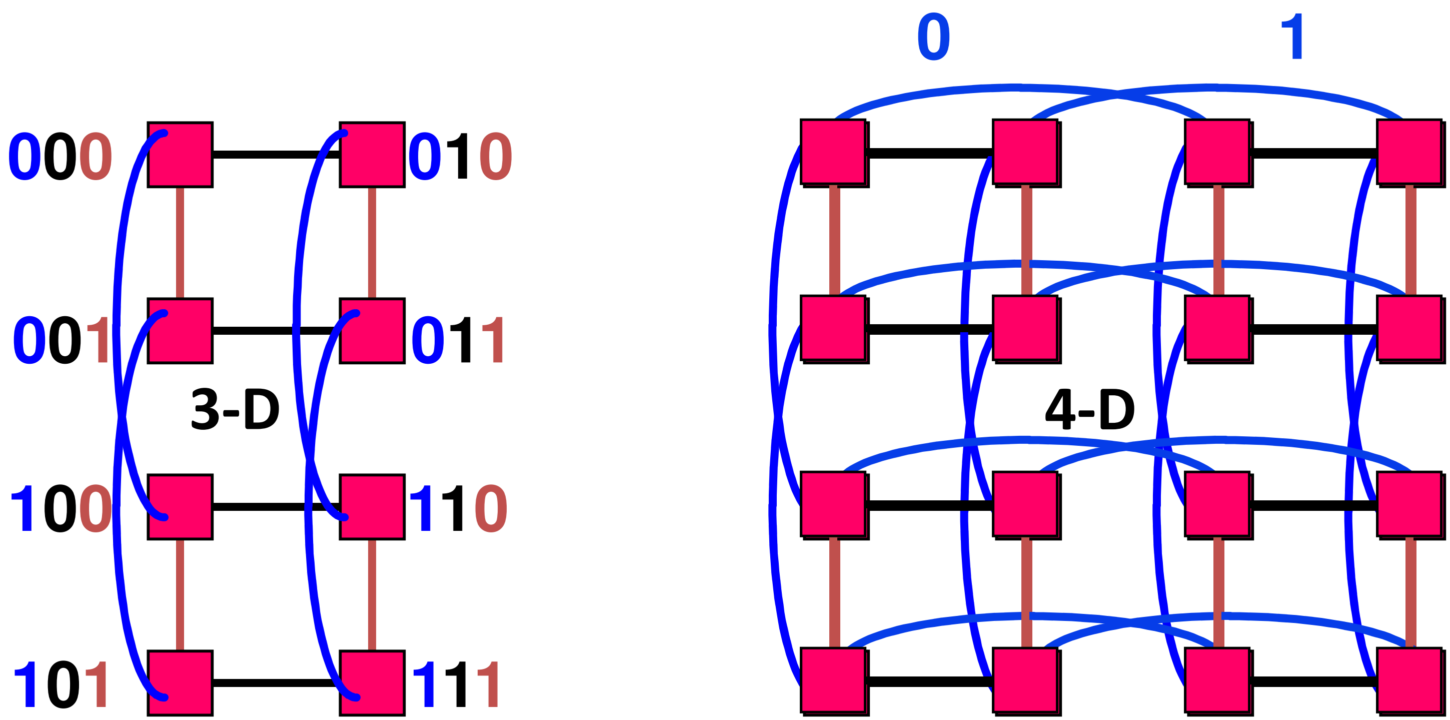


Hypercube

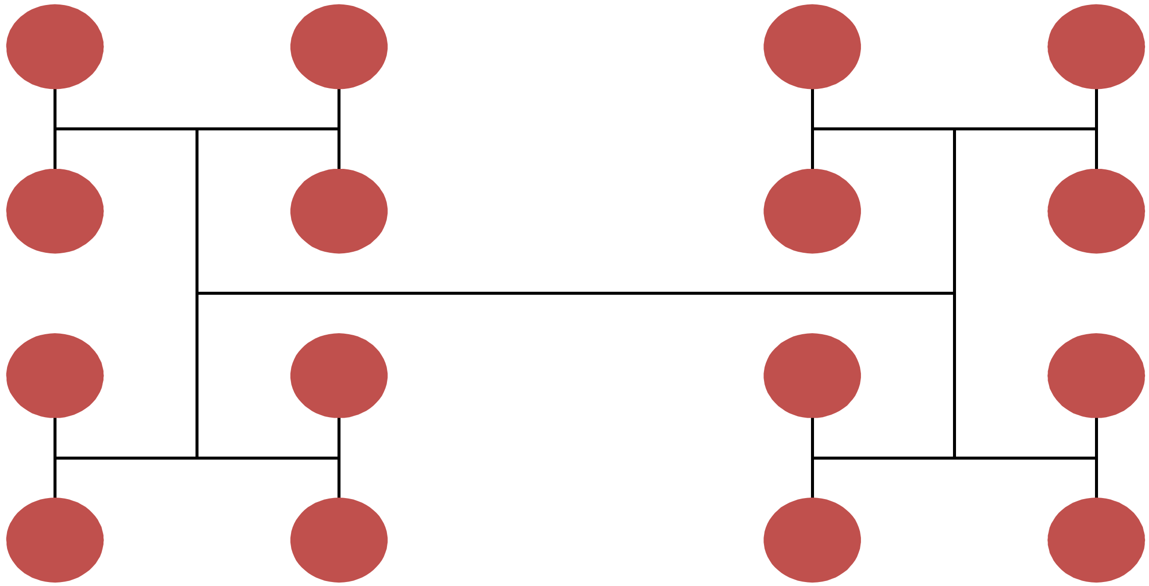


Diameter O(log n)

Bisection width Θ(n)



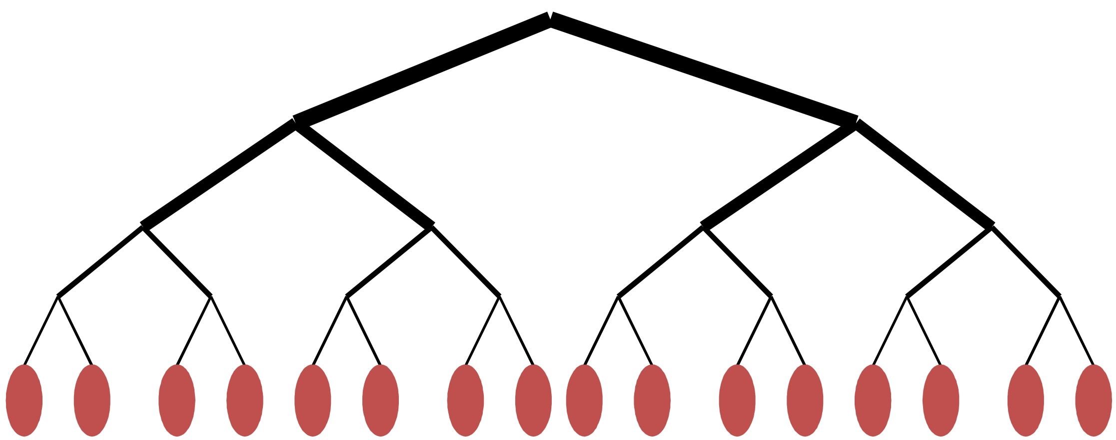
Trees

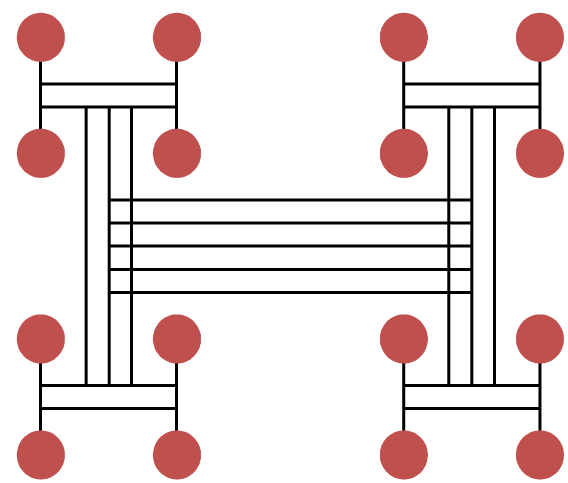


- Diameter **= log n**.
- Bisection bandwidth = **1**
- Easy layout as planar graph
- Many tree algorithms (e.g., summation)

Fat-Trees

- **Fat trees** avoid bisection bandwidth problem of tree:
  - More (or wider) links near top
  - Example: Thinking Machines CM-5





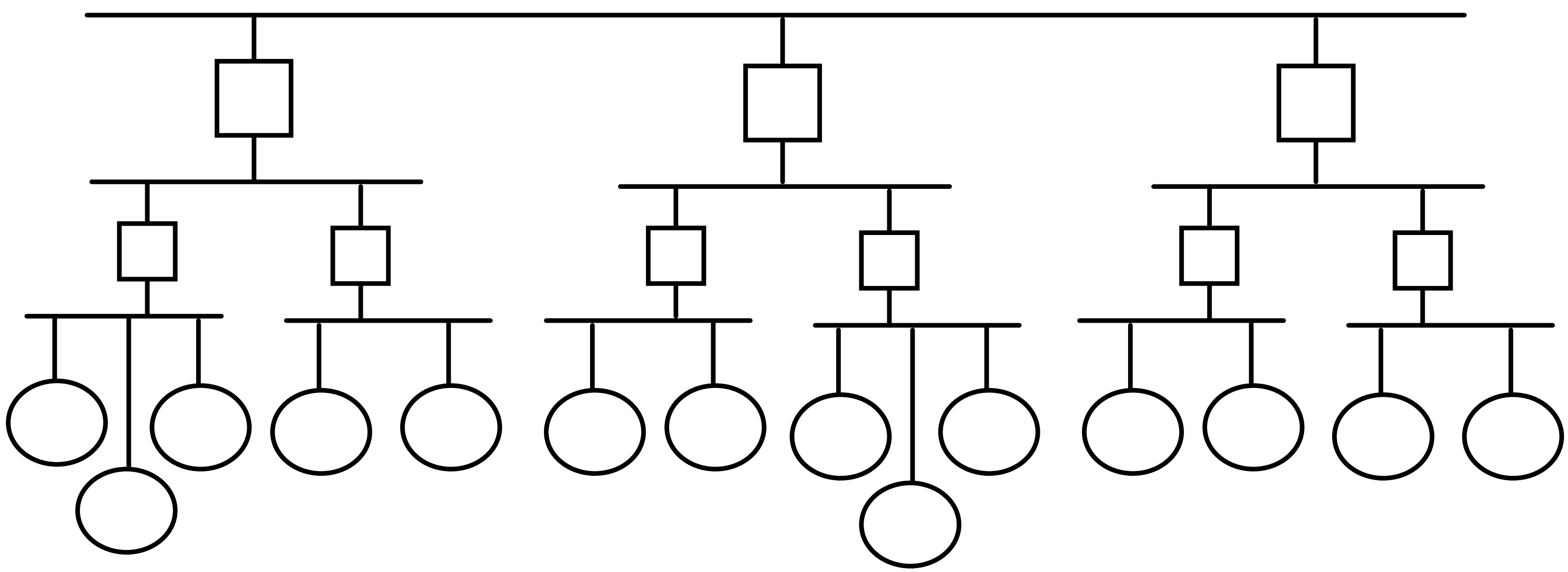
Common Topologies

Type	Degree	Diameter	Ave Dist	Bisection
1D mesh	2	N-1	N/3	1
2D mesh	4	$2(N^{1/2} - 1)$	$2N^{1/2} / 3$	$N^{1/2}$
3D mesh	6	$3(N^{1/3} - 1)$	$3N^{1/3} / 3$	$N^{2/3}$
nD mesh	2n	$n(N^{1/n} - 1)$	$nN^{1/n} / 3$	$N^{(n-1) / n}$
Ring	2	N/2	N/4	2
2D torus	4	$N^{1/2}$	$N^{1/2} / 2$	$2N^{1/2}$
Hypercube	Log_2N	$n=Log_2N$	n/2	N/2
2D Tree	3	$2Log_2N$	$\sim 2Log_2 N$	1
Crossbar	N-1	1	1	$N^2/2$

N = number of nodes, n = dimension

Hierarchical (Multilevel) Networks

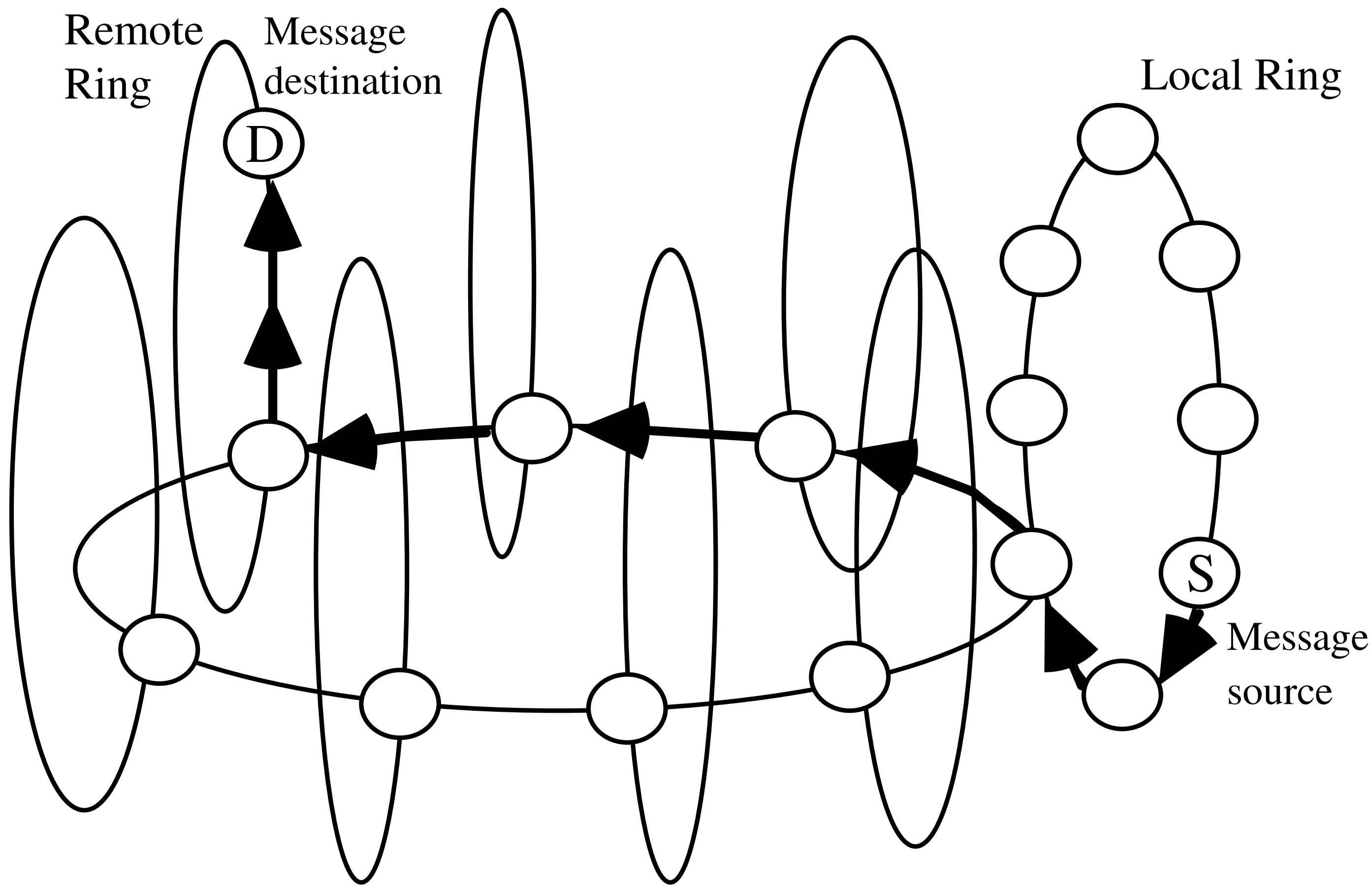
We have already seen several examples of hierarchical networks: multilevel buses



Hierarchical or multilevel bus network.

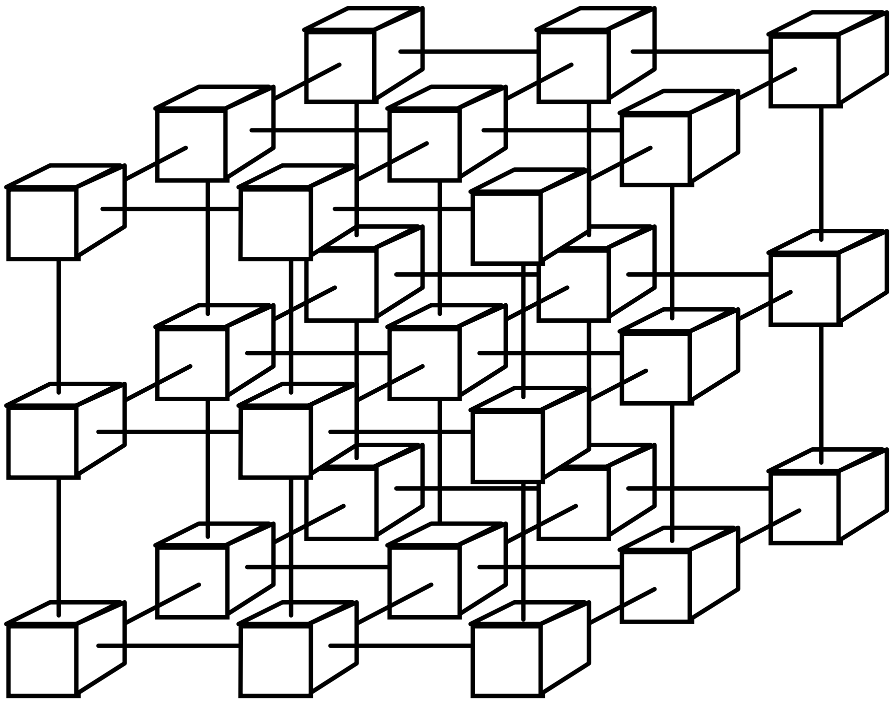
Ring of Ring

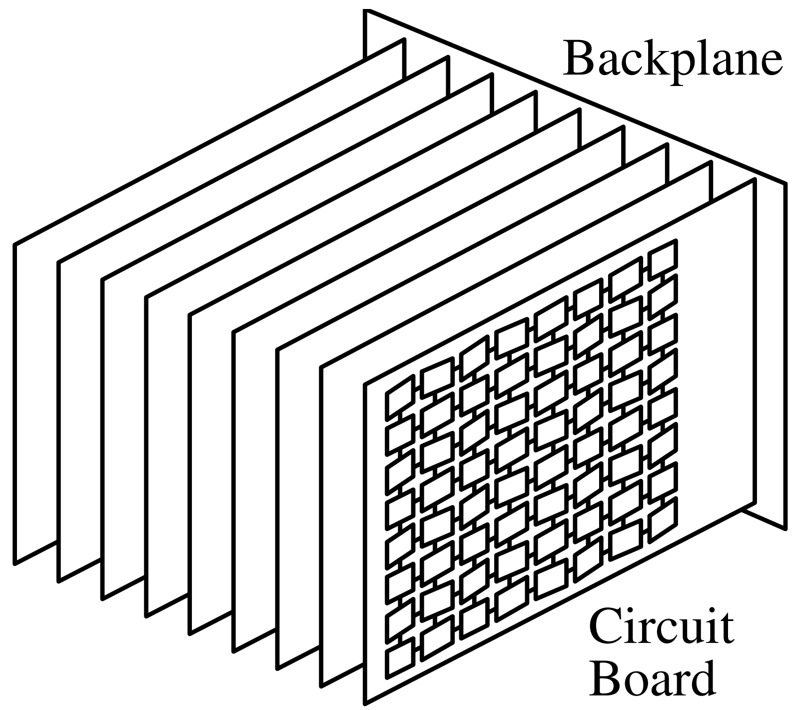
Rings are simple, but have low performance and lack robustness



A 64-node ring-of-rings architecture composed of eight 8-node local rings and one second-level ring.

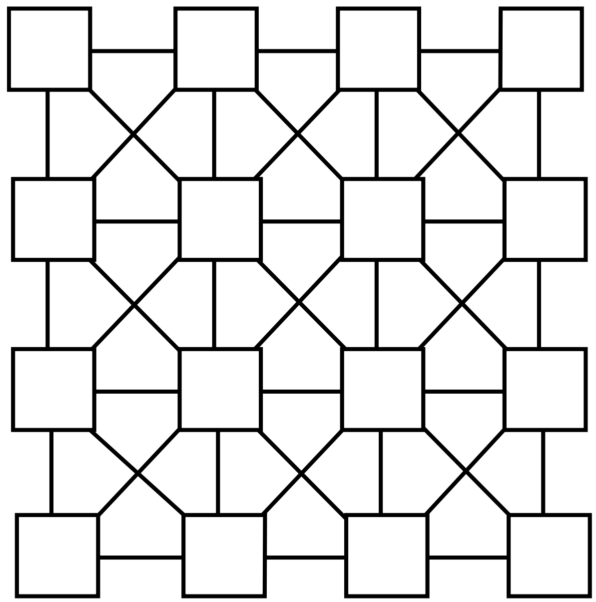
2.5D and 3D MESH

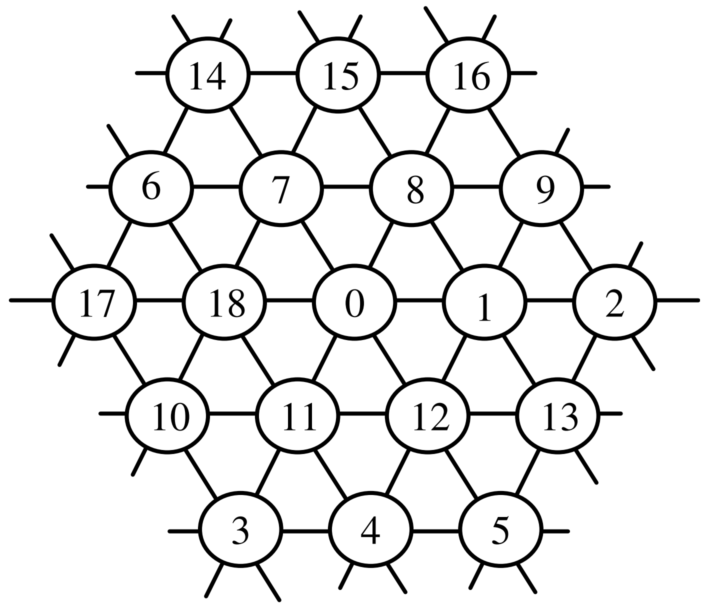




3D and 2.5D physical realizations of a 3D mesh.

Stronger and Weaker Connectivities MESH





Node i connected to i ± 1, i ± 7, and i ± 8 (mod 19).

Fortified meshes and other models with stronger connectivities:

Eight-neighbor
Six-neighbor
Triangular
Hexagonal

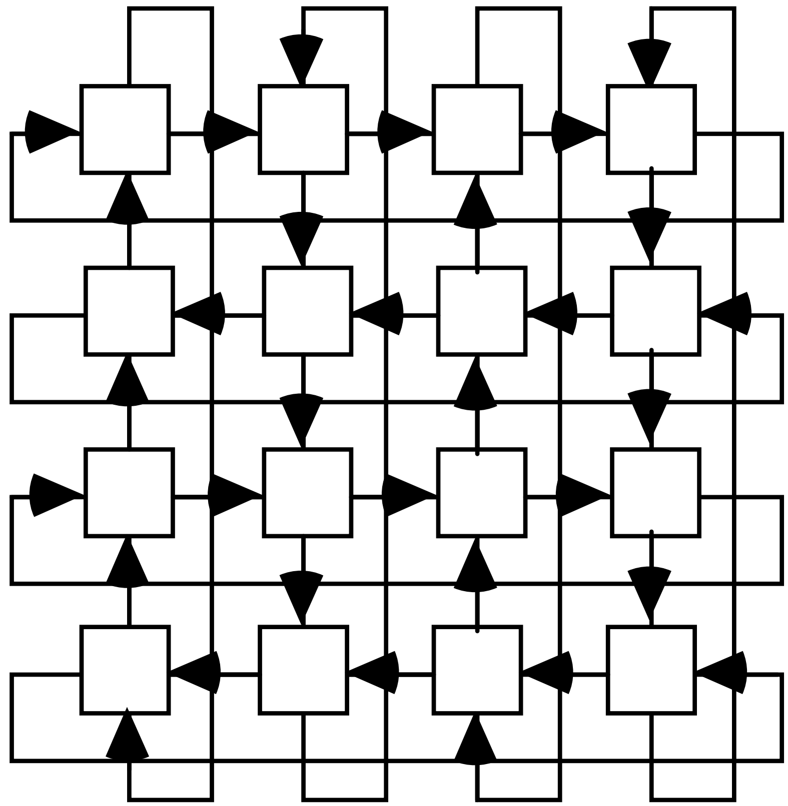
Eight-neighbor and hexagonal (hex) meshes.

As in higher-dimensional meshes, greater connectivity does not automatically translate into greater performance

Area and signal-propagation delay penalties must be factored in

Simplification via Link Orientation

Two in- and out-channels per node, instead of four



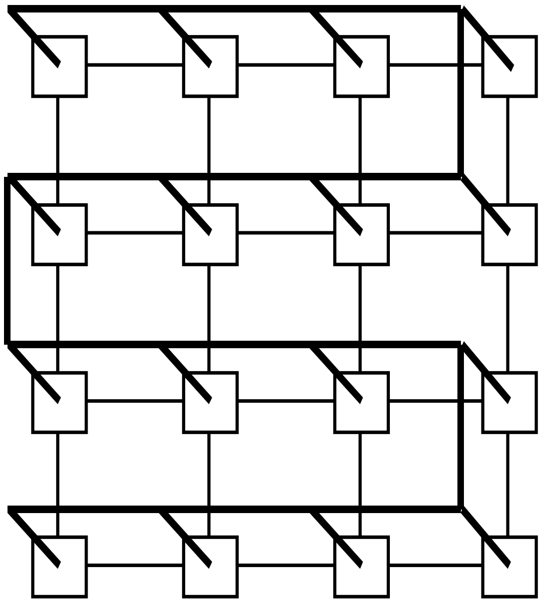
With even side lengths, the diameter does not change

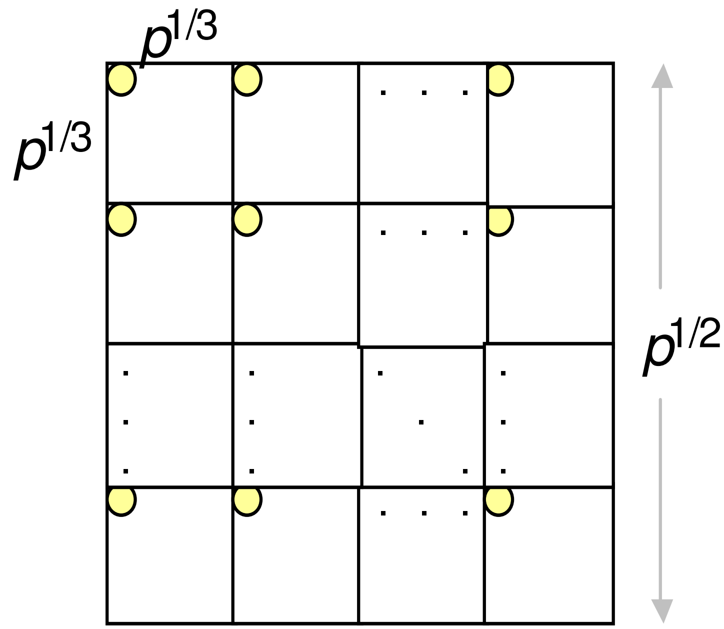
Some shortest paths become longer, however

Can be more cost-effective than 2D mesh

4×4 Manhattan street network.

Using a Single Global Bus





Mesh with a global bus

The single bus increases the bisection width by 1

Broadcast the result to all nodes (one step)