HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Interconnection Networks

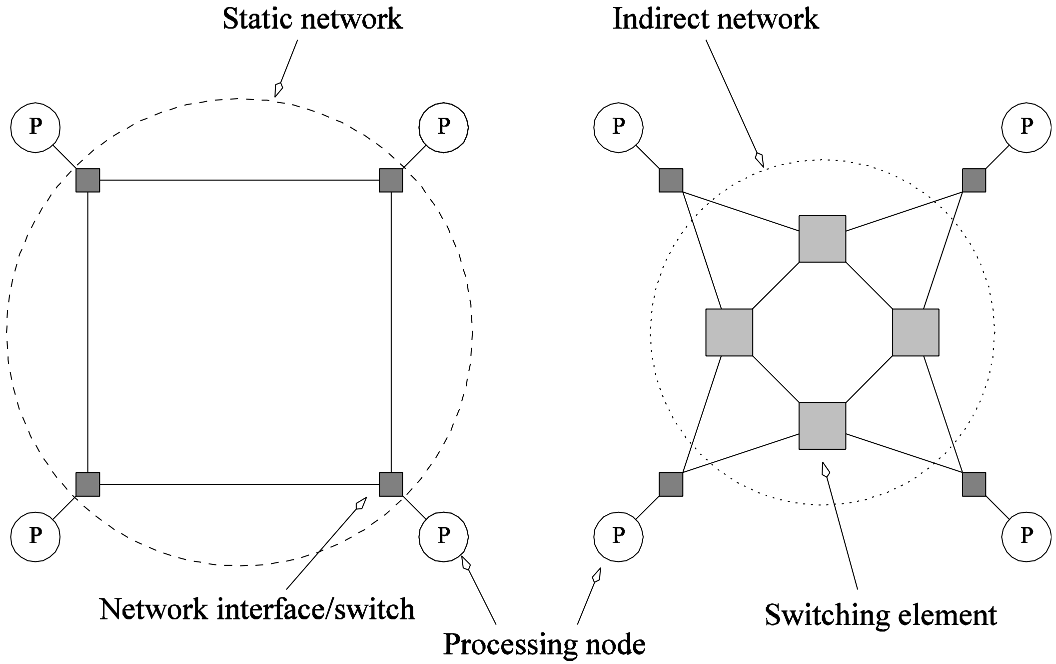
Programmierung Paralleler und Verteilter Systeme (PPV)

Sommer 2015

Frank Feinbube, M.Sc., Felix Eberhardt, M.Sc.,
Prof. Dr. Andreas Polze

Interconnection Networks

- SIMD systems demand structured connectivity
  - Processor-to-processor interaction
  - Processor-to-memory interaction
- Static network
  - Point-to-point links, fixed route
- Dynamic network
  - Consists of links and switching elements
  - Flexible configuration of processor interaction



Interconnection networks

Optimization criteria

Connectivity – ideally direct links between any two stations

High number of parallel connections

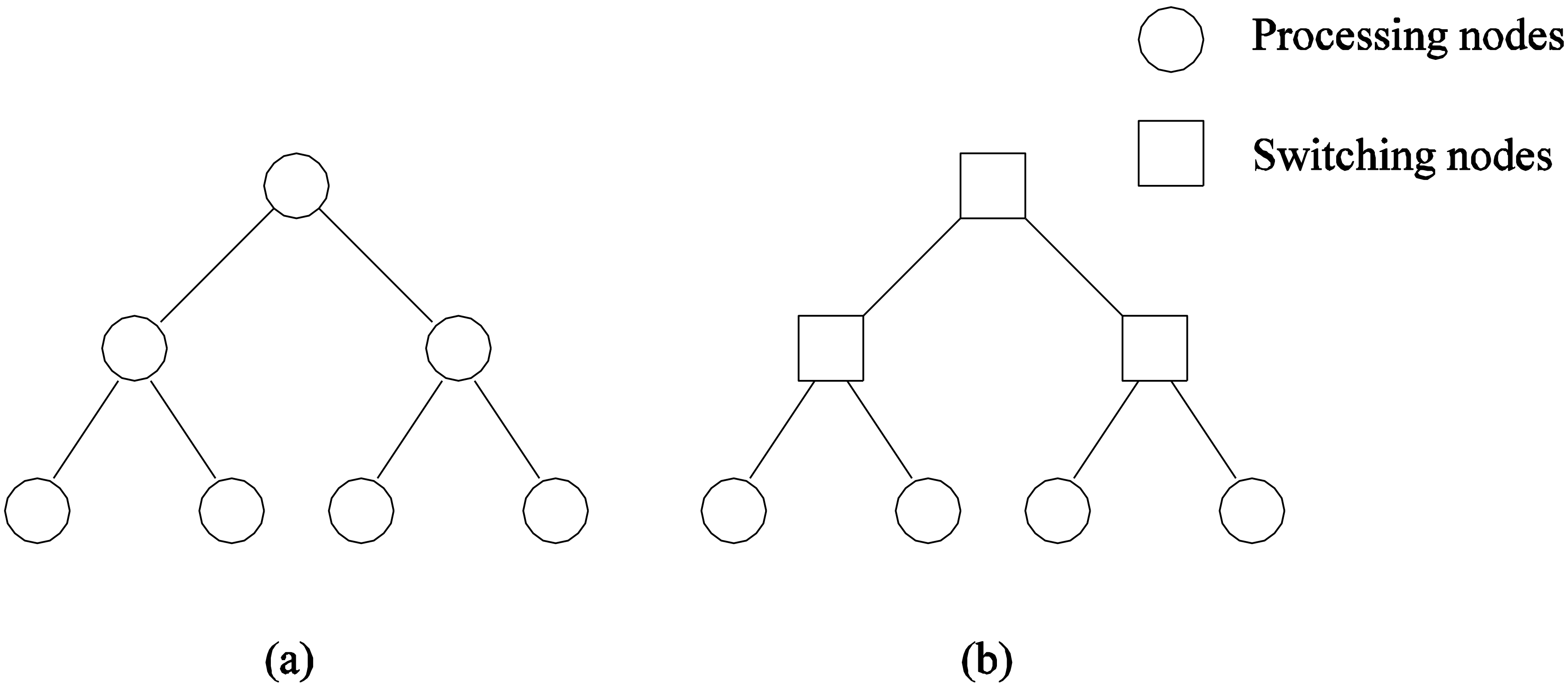
Cost model

Production cost - # connections

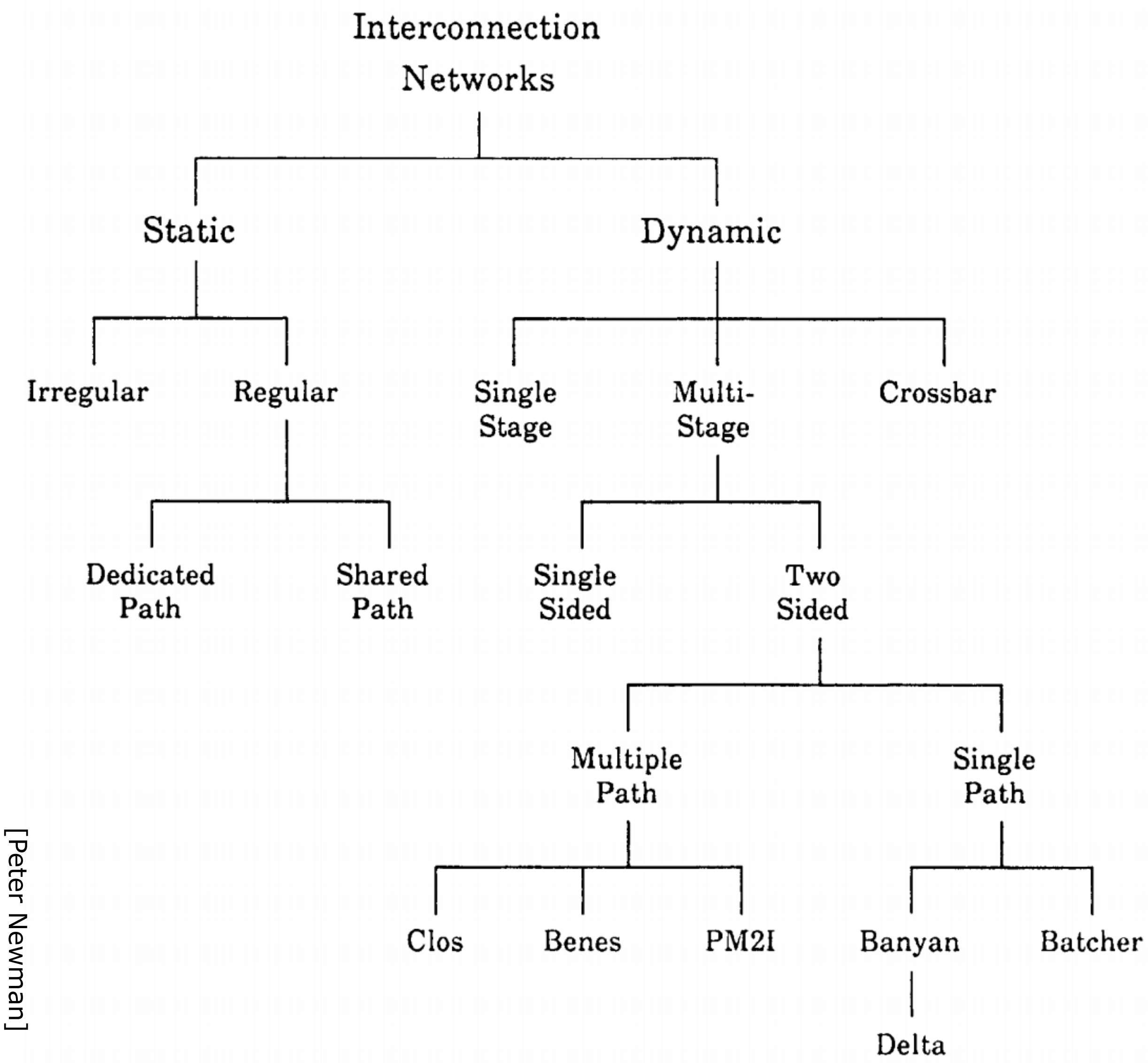
operational cost – distance among PEs

Bus networks, switching networks, point-to-point interconnects

Interconnection Networks



Interconnection Networks



[Peter Newman]

- *Dynamic networks* are built from a graph of configurable switching elements
- General packet switching network counts as *irregular static network*

Interconnection Networks

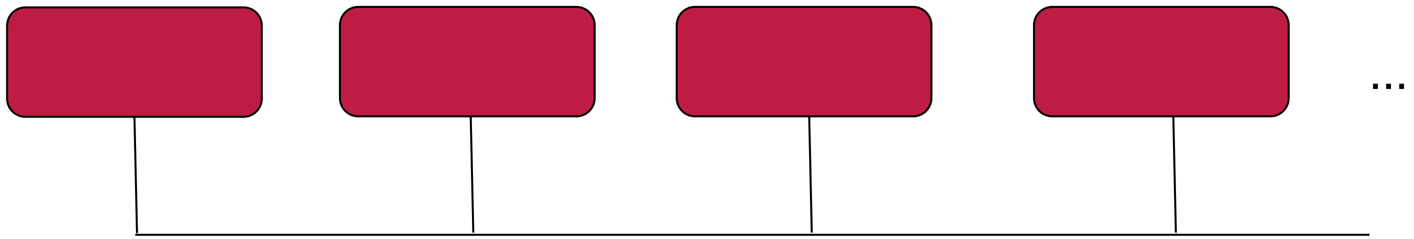
- Network Interfaces
  - Processors talk to the network via a *network interface connector (NIC)* hardware
  - Network interfaces attached to the interconnect
    - Cluster vs. tightly-coupled multi-computer
  - SIMD hardware bundles NIC with the processor
- Switching elements map a fixed number of inputs to outputs
  - Total number of ports is the **degree** of the switch
  - The **cost** of a switch grows as square of the degree
  - The **peripheral hardware** grows linearly as the degree

Interconnection Networks

- A variety of network topologies proposed and implemented
- Each topology has a performance / cost tradeoff
- Commercial machines often implement hybrids
  - Optimize packaging and costs
- Metrics for an interconnection network graph
  - **Diameter**: Maximum distance between any two nodes
  - **Connectivity**: Minimum number of edges that must be removed to get two independent graphs
  - **Link width / weight**: Transfer capacity of an edge
  - **Bisection width**: Minimum transfer capacity given between any two halves of the graph
  - **Costs**: Number of edges in the network
- Often optimization for connectivity metric

Bus Systems

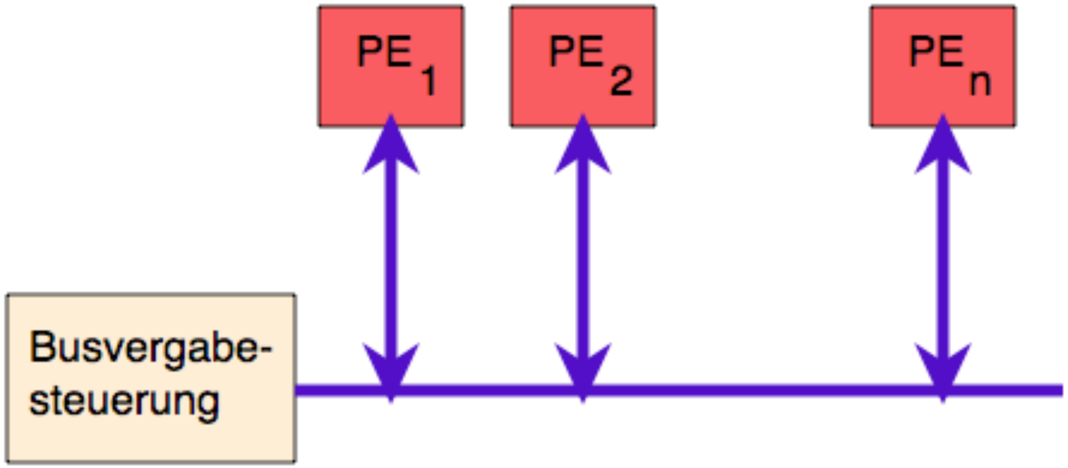
- Static interconnect technology
- Shared communication path, broadcasting of information
  - Diameter: O(1)
  - Connectivity: O(1)
  - Bisection width: O(1)
  - Costs: O(p)



Bus network

Optimal #connection per PE: 1

Constant distance among any two PEs



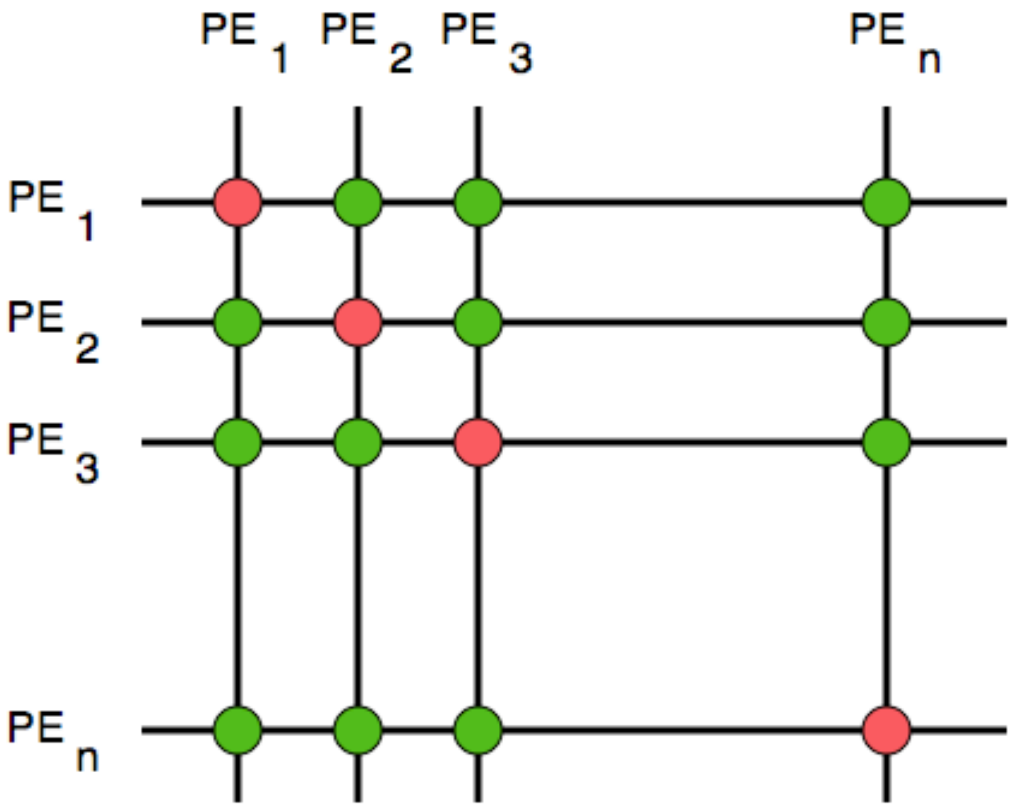
Crossbar switch (Kreuzschienenverteiler)

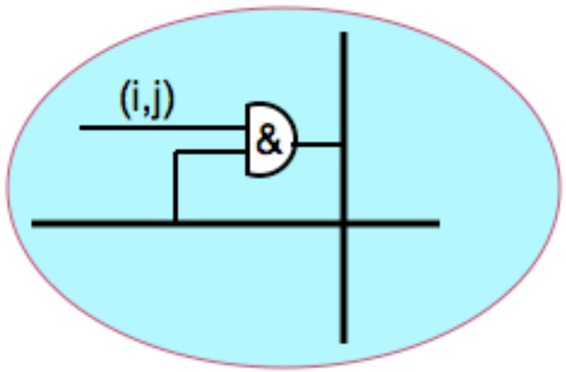
Arbitrary number of permutations

Collision-free data exchange

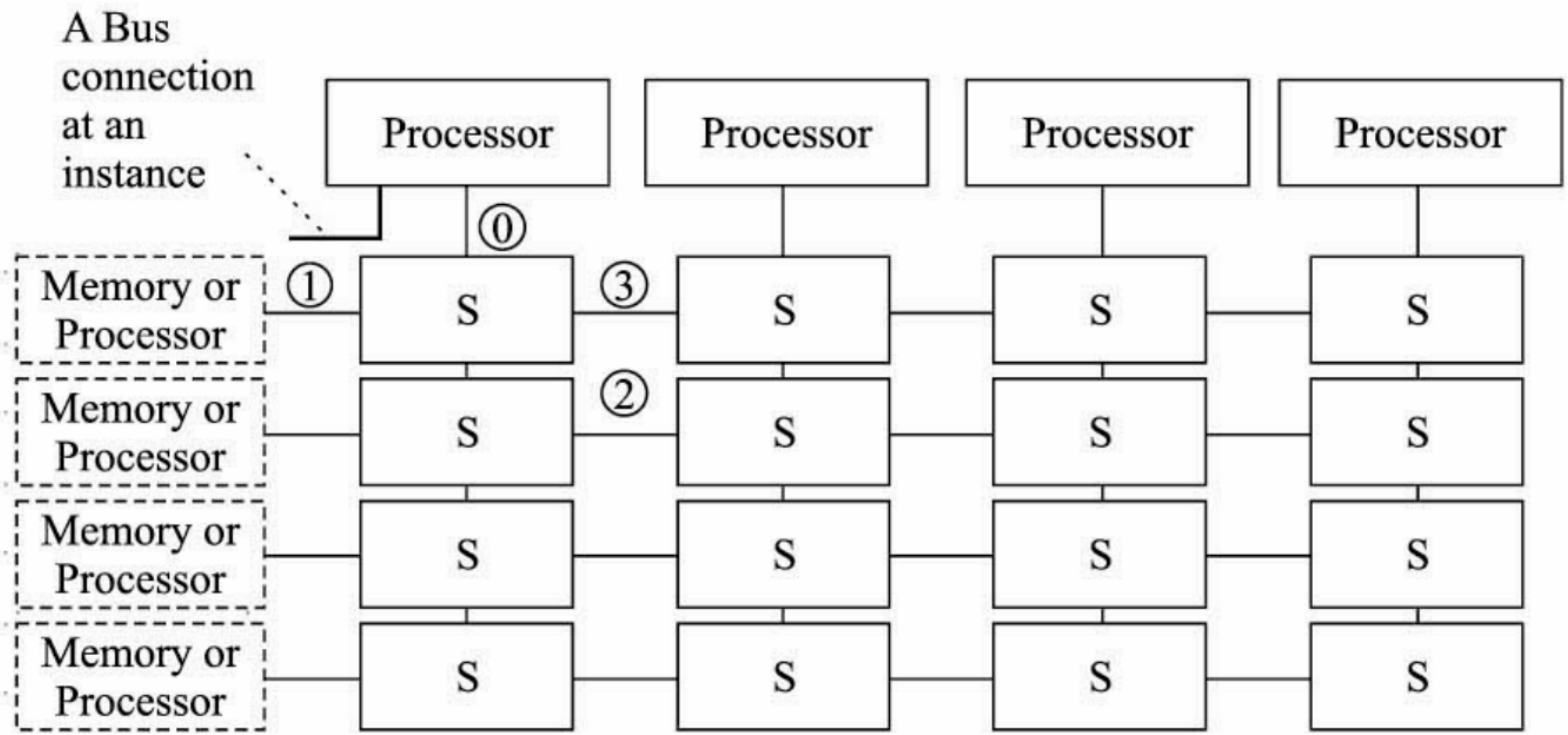
High cost, quadratic growth

n * (n-1) connection points



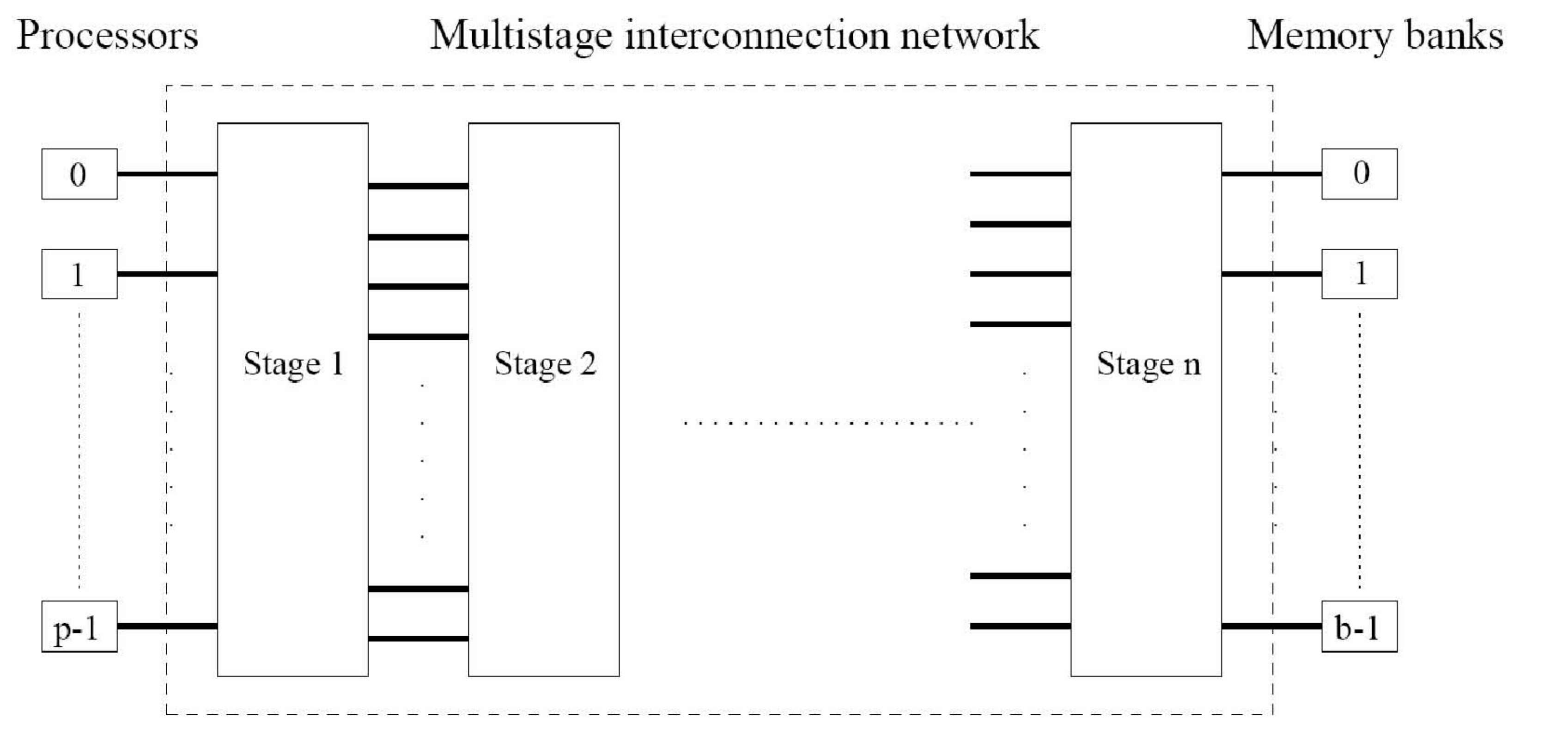


Crossbar Switch



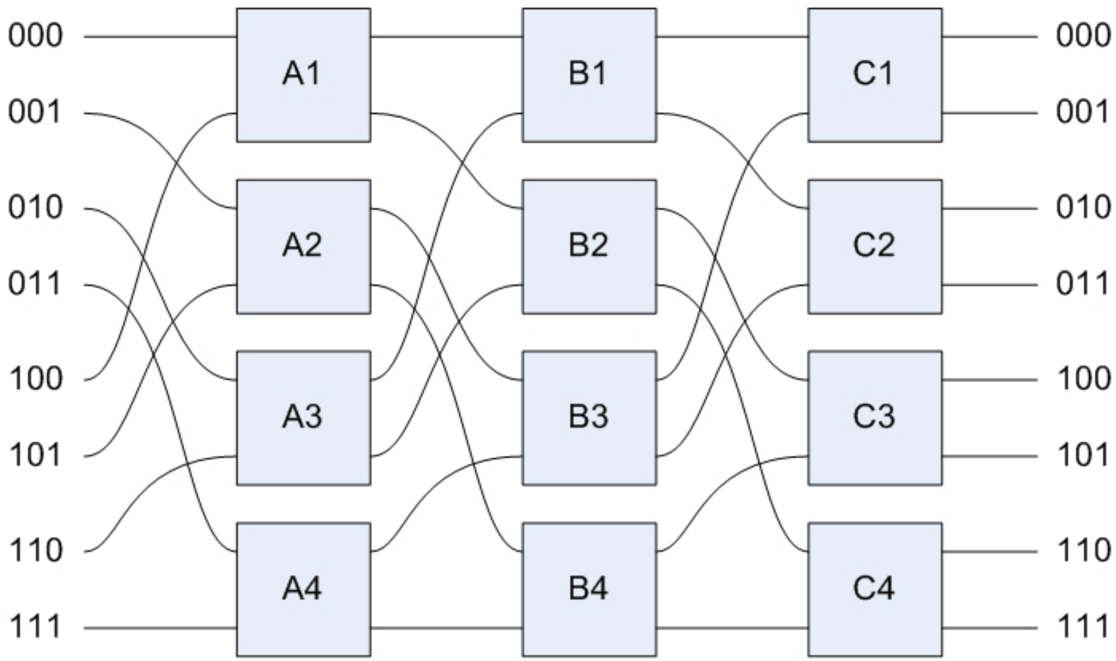
Multistage Interconnection Networks

- Connection by switching elements
- Typical solution to connect processing and memory elements
- Can implement sorting or shuffling in the network routing



Omega Network

- Inputs are crossed or not, depending on routing logic
  - Destination-tag routing: Use positional bit for switch decision
  - XOR-tag routing: Use positional bit of XOR result for decision
- For N PE's, N/2 switches per stage, $\log_2 N$ stages
- Decrease bottleneck probability on parallel communication

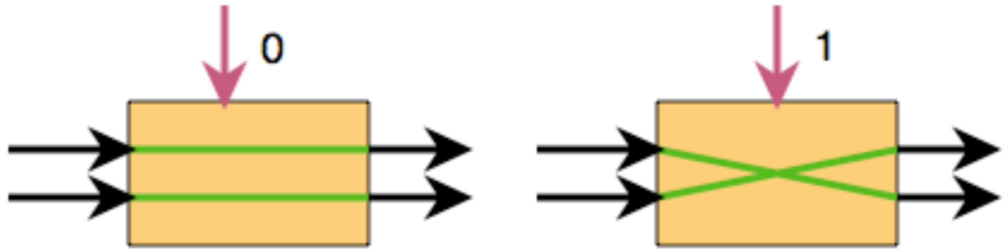


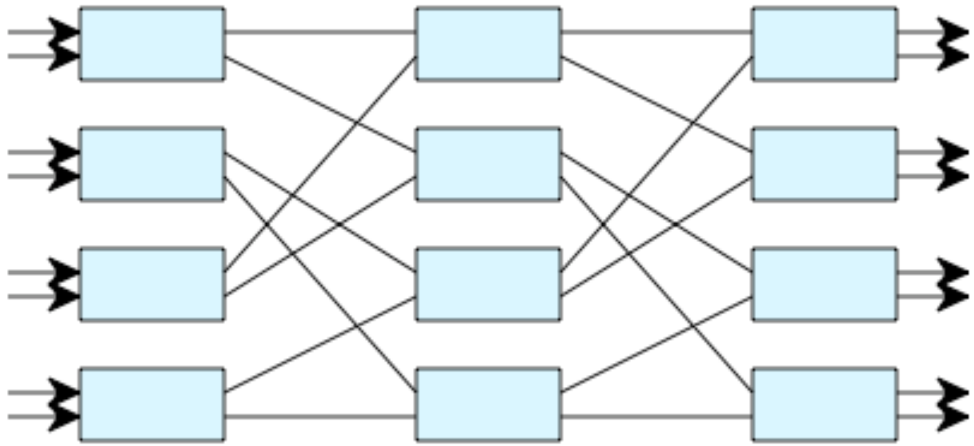
Delta networks

Only n/2 log n delta-switches

Limited cost

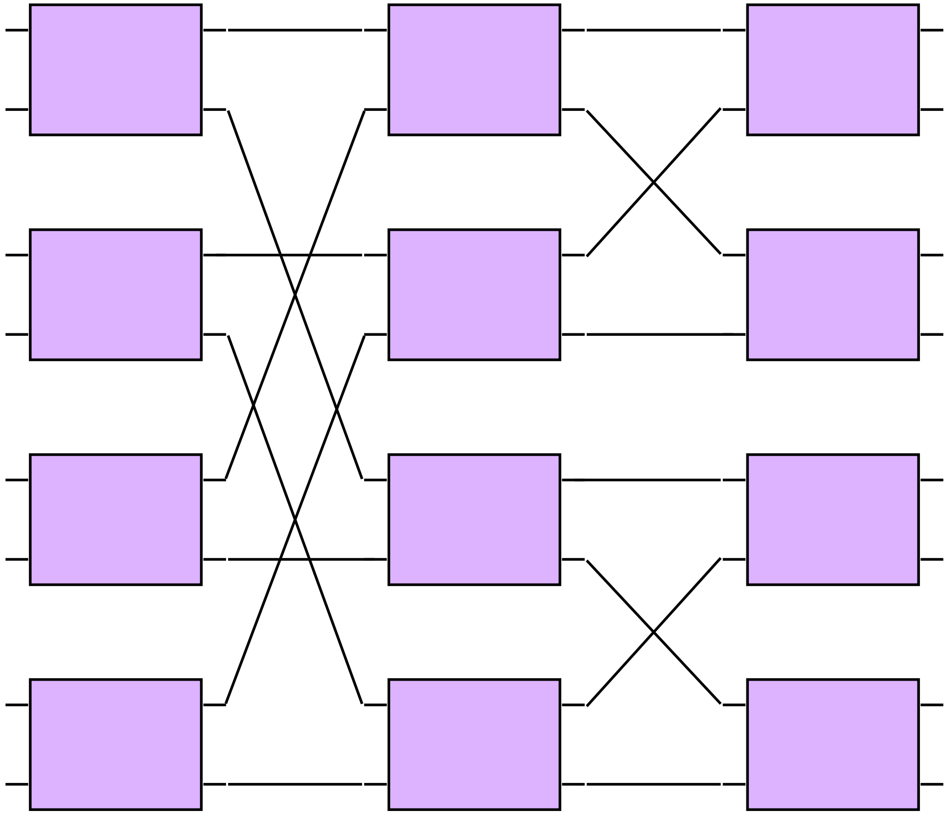
Not all possible permutations operational in parallel





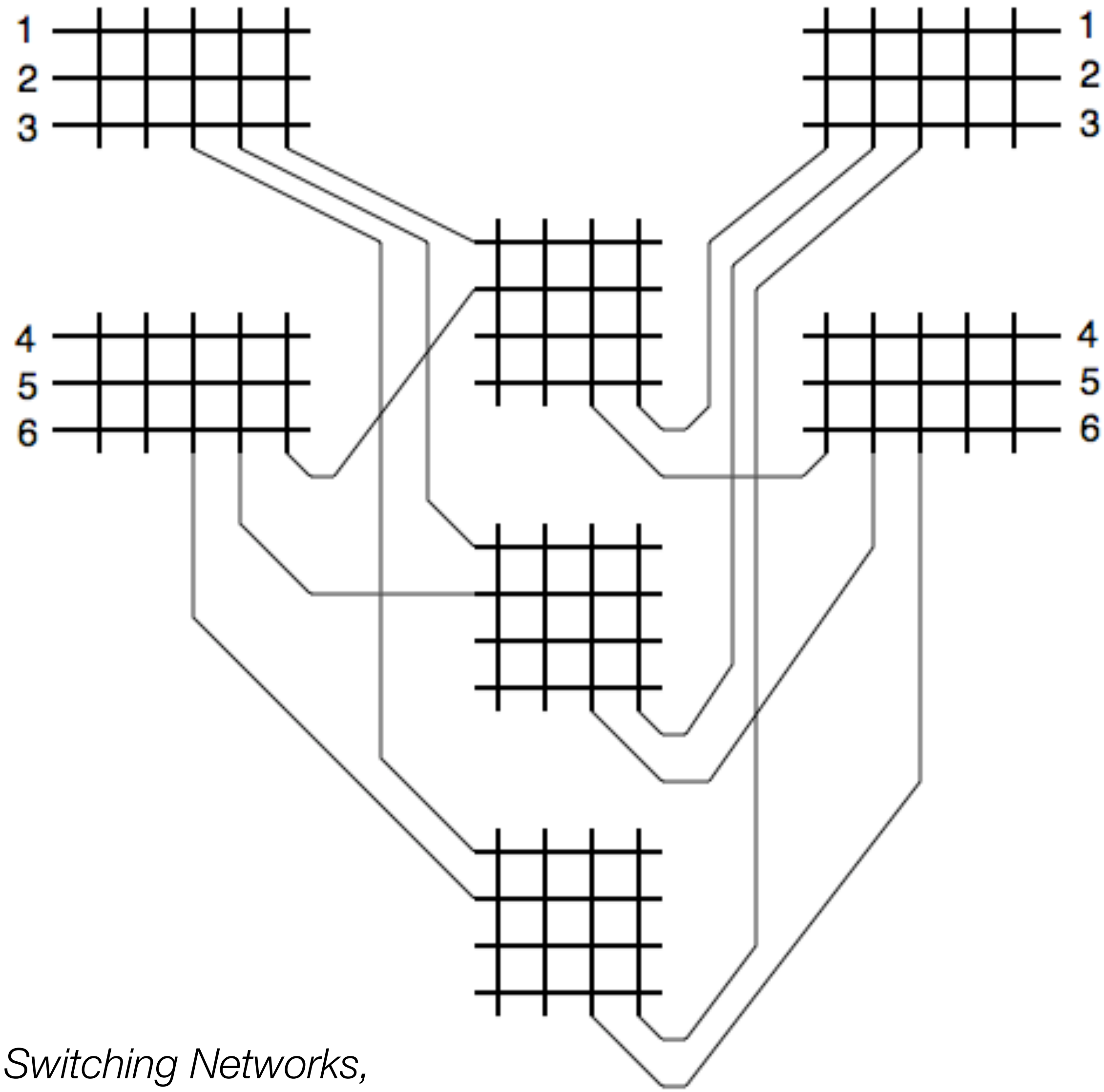
Delta Networks operation

- Stage n checks bit k of the destination tag
- Possible effect of ‚output port contention‘ and ‚path contention‘



Clos coupling networks

Combination of delta network and crossbar

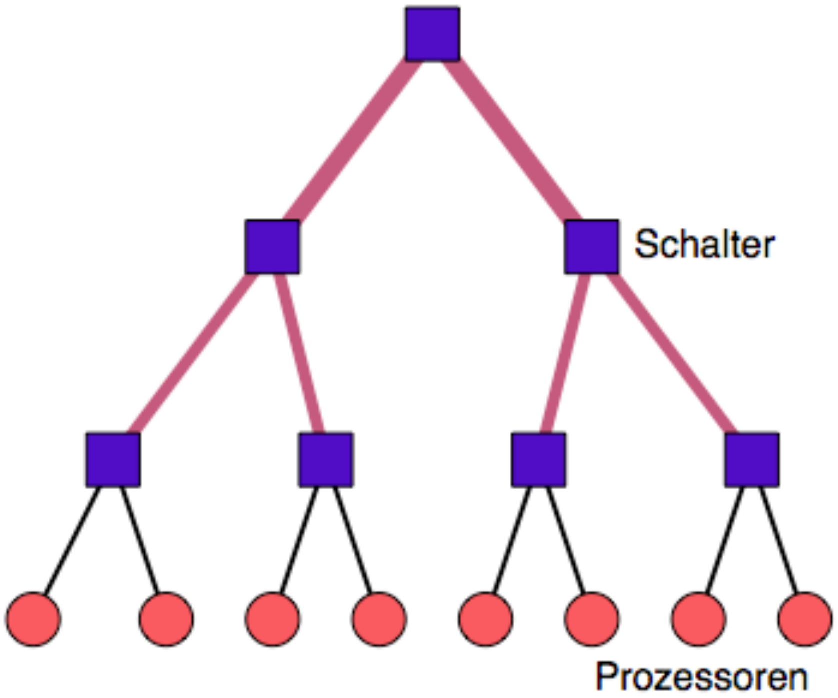


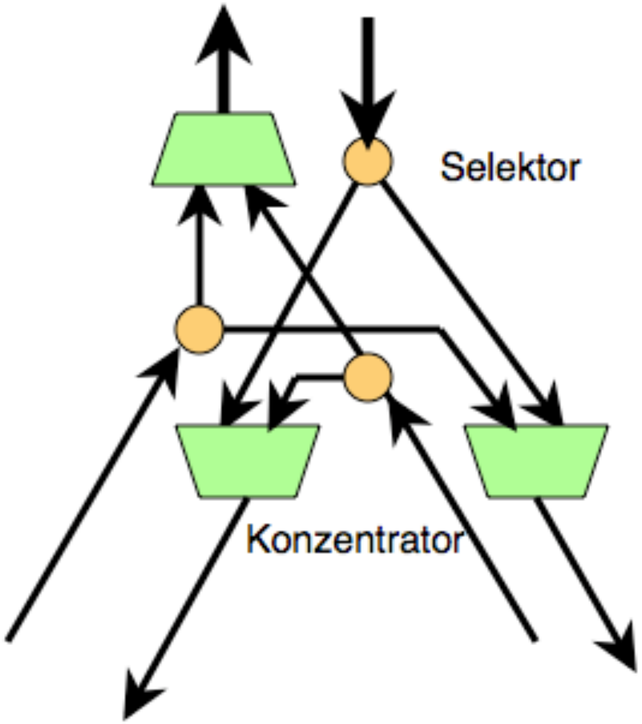
C.Clos, *A Study of Nonblocking Switching Networks, Bell System Technical Journal, vol. 32, no. 2, 1953, pp. 406-424(19)*

Fat-Tree networks

PEs arranged as leafs on a binary tree

Capacity of tree (links) doubles on each layer

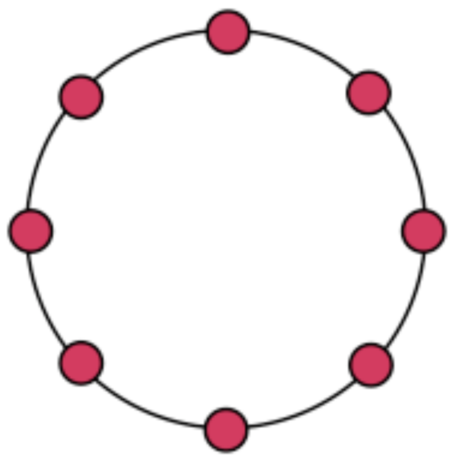




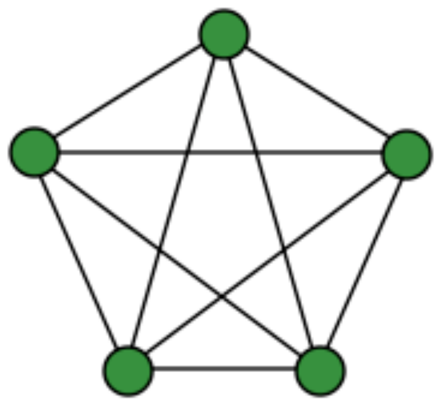
Point-to-point networks: ring and fully connected graph

Ring has only two connections per PE (almost optimal)

Fully connected graph – optimal connectivity (but high cost)



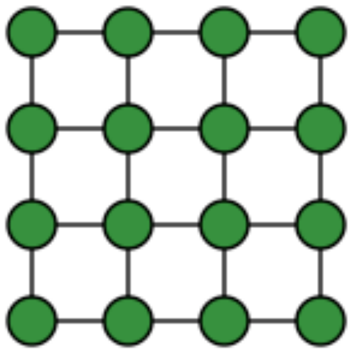
Ring



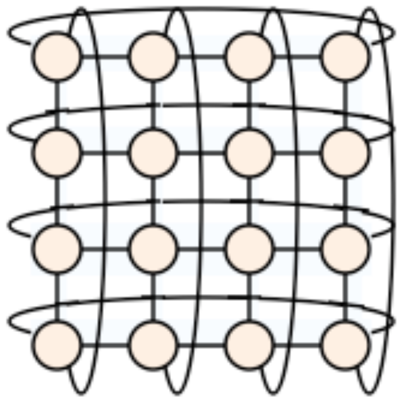
vollständiger Graph

Mesh and Torus

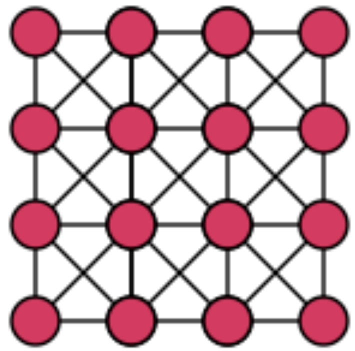
Compromise between cost and connectivity



Quadratisches Gitter (4-way)



Quadratischer Torus (4-way)

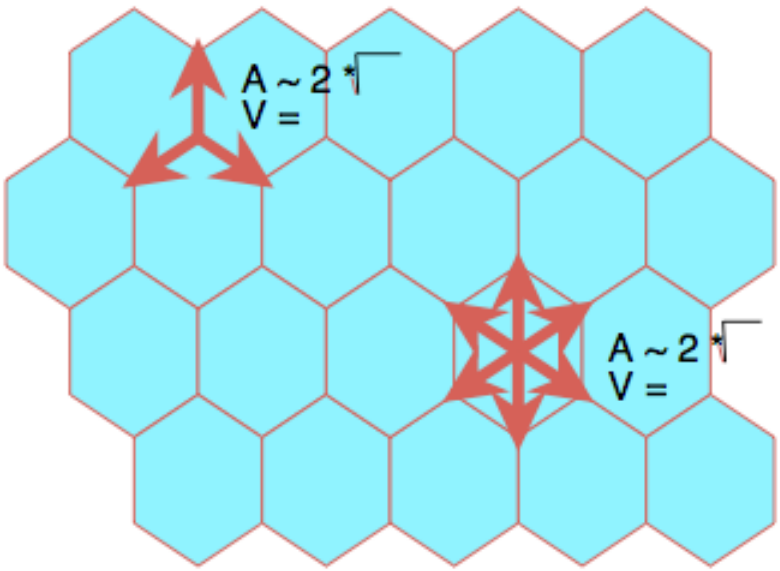


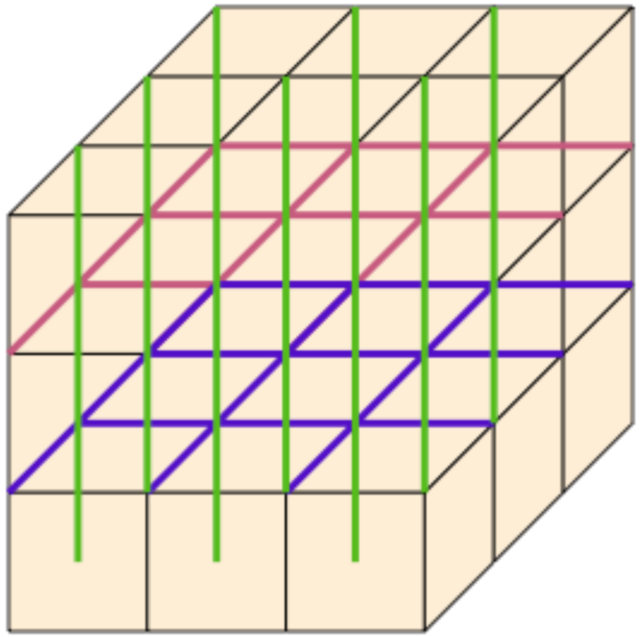
Quadratisches Gitter (8-way)

Cubic Mesh

PEs are arranged in a cubic fashion

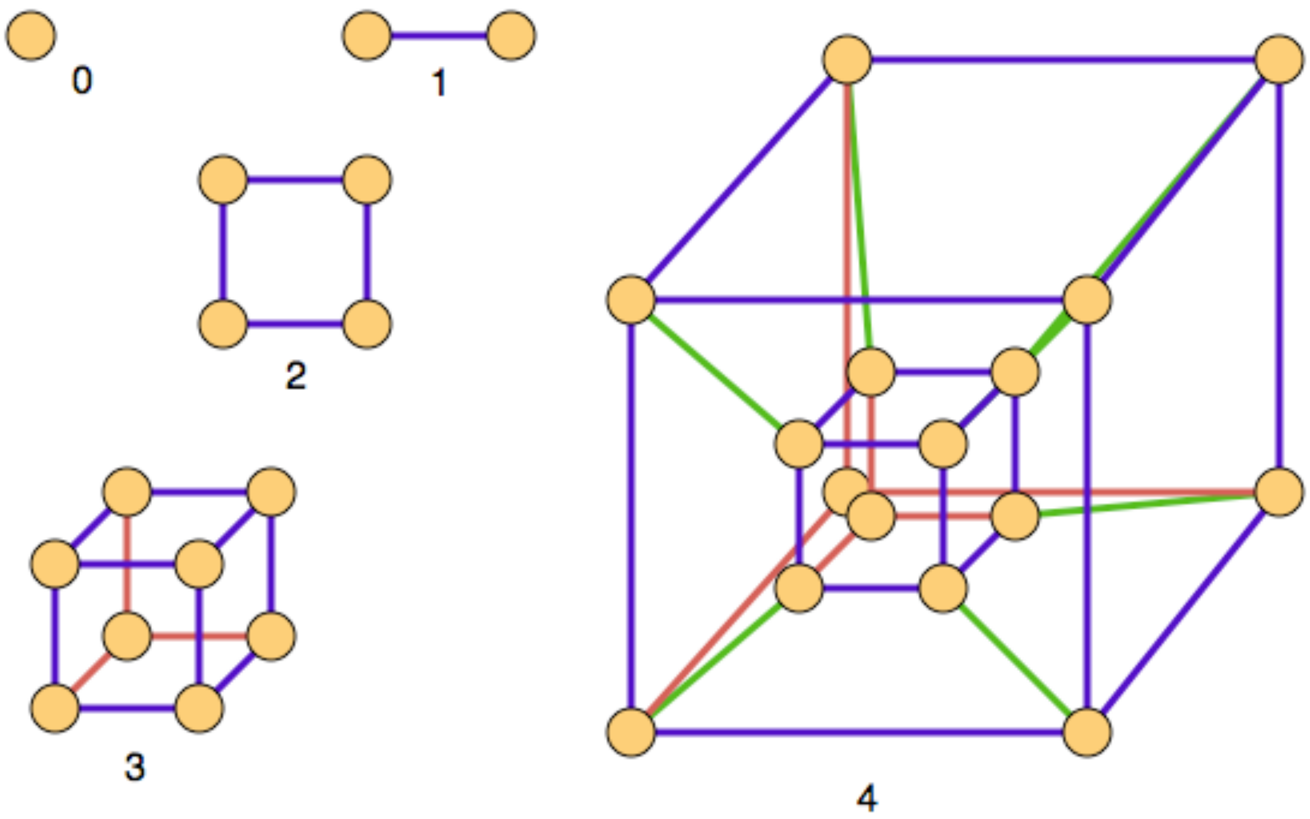
Each PE has 6 links to neighbors





Hypercube

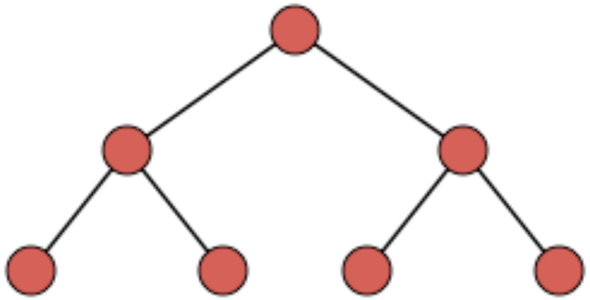
Dimensions 0-4, recursive definition

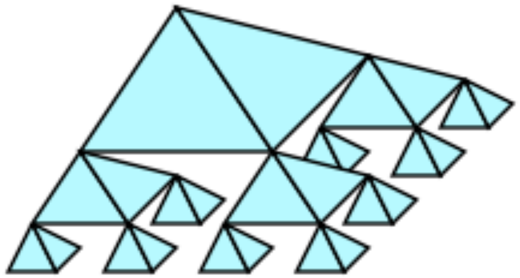


Binary tree, quadtree

Logarithmic cost

Problem of bottleneck at root node

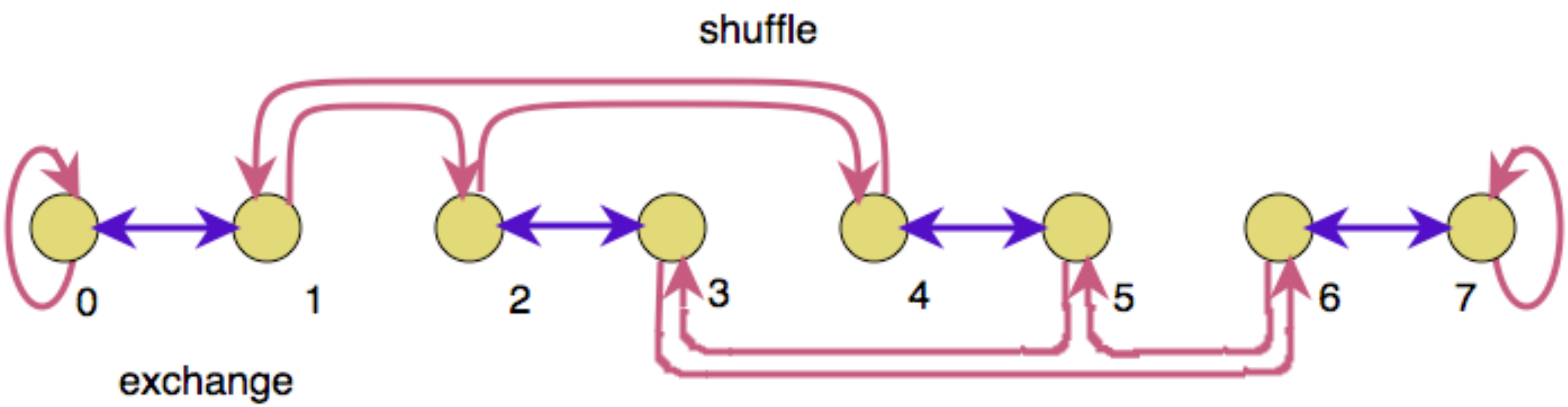




Shuffle-Exchange network

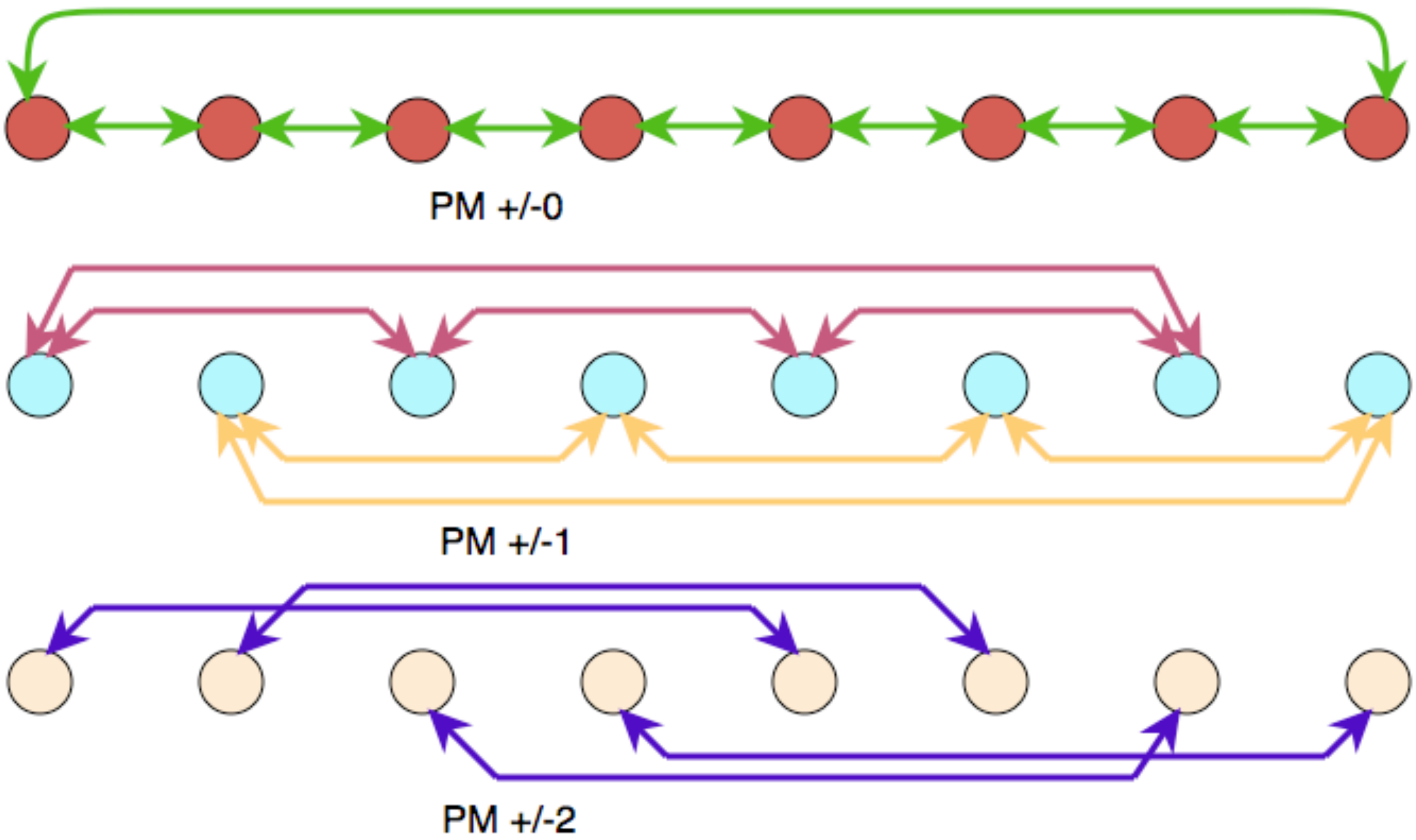
Logarithmic cost

Uni-directional shuffle network + bi-directional exchange network



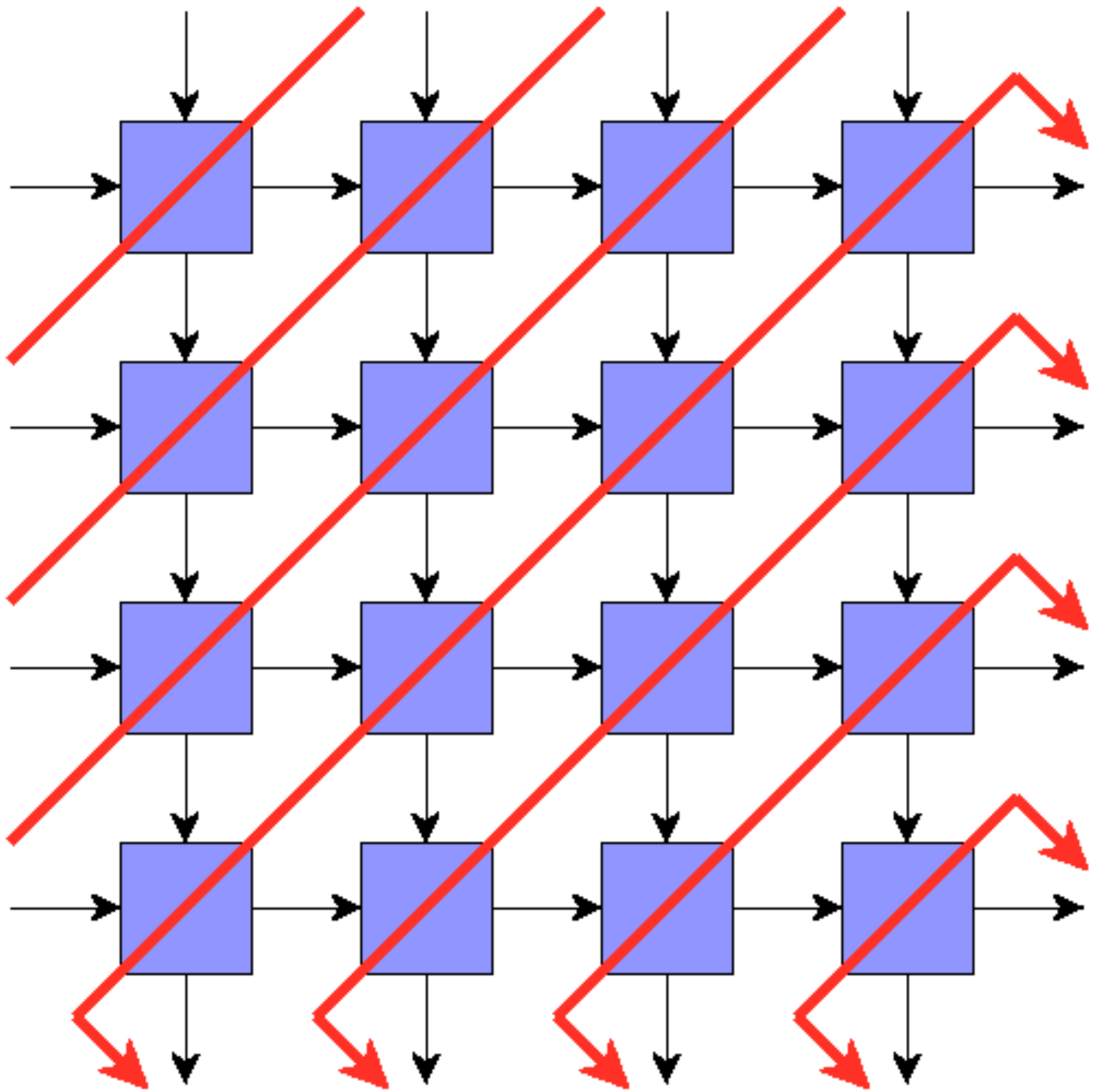
Plus-Minus-Network

PM 2i – 2*m-1 separate unidirectional interconnection networks



Systolic Arrays

- Data flow architecture
- Common clock
- Maximum signal path restricted by frequency
- Single faulty element breaks the complete array



Comparison

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

Comparison

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	$p - 1$

Comparison of networks

Netz1 simuliert Netz2 ->	Gitter(2D)	PM 2i	Shuffle-Exchange	Hypercube
Gitter(2D)	—	$\frac{\mathrm{sqrt}(2)}{2}$	sqrt(n)	sqrt(n)
PM 2i	1	—	$\log_2 n$	2
Shuffle-Exchange	$2 * \log_2 n$	$2 * \log_2 n$	—	$\log_2 n + 1$
Hypercube	$\log_2 n$	$\log_2 n$	$\log_2 n$	—