RUPRECHT-KARLS-
UNIVERSITÄT
HEIDELBERG

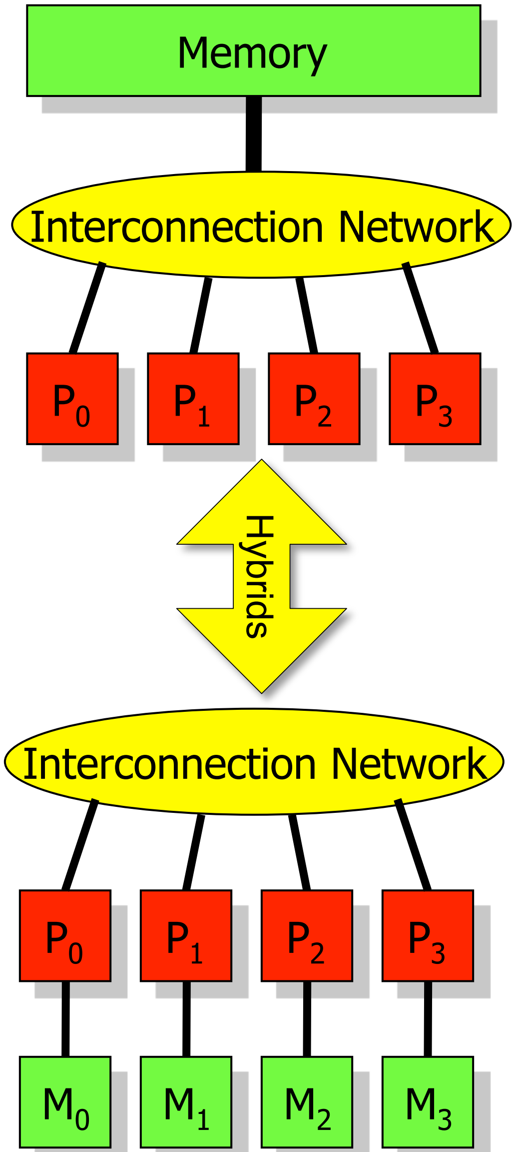
Introduction to High Performance Computing

Lecture 11 – Basics of Interconnection Networks I

Holger Fröning
Institut für Technische Informatik
Universität Heidelberg

Introduction

- Up to now: Interconnection Network (IN) as a black box
  - Turning into the **key component** of HPC systems
  - Exact behavior is crucial to overall performance
- INs are found everywhere
  - On-Chip Networks (different modules or cores)
  - Intra-Node (CPU, memory, graphics, devices)
  - Inter-Node (multiple nodes)
    - SAN, LAN, WAN
- Different requirements/workloads!
  - Here: focus on HPC and its demands



Types of INs in a computer system

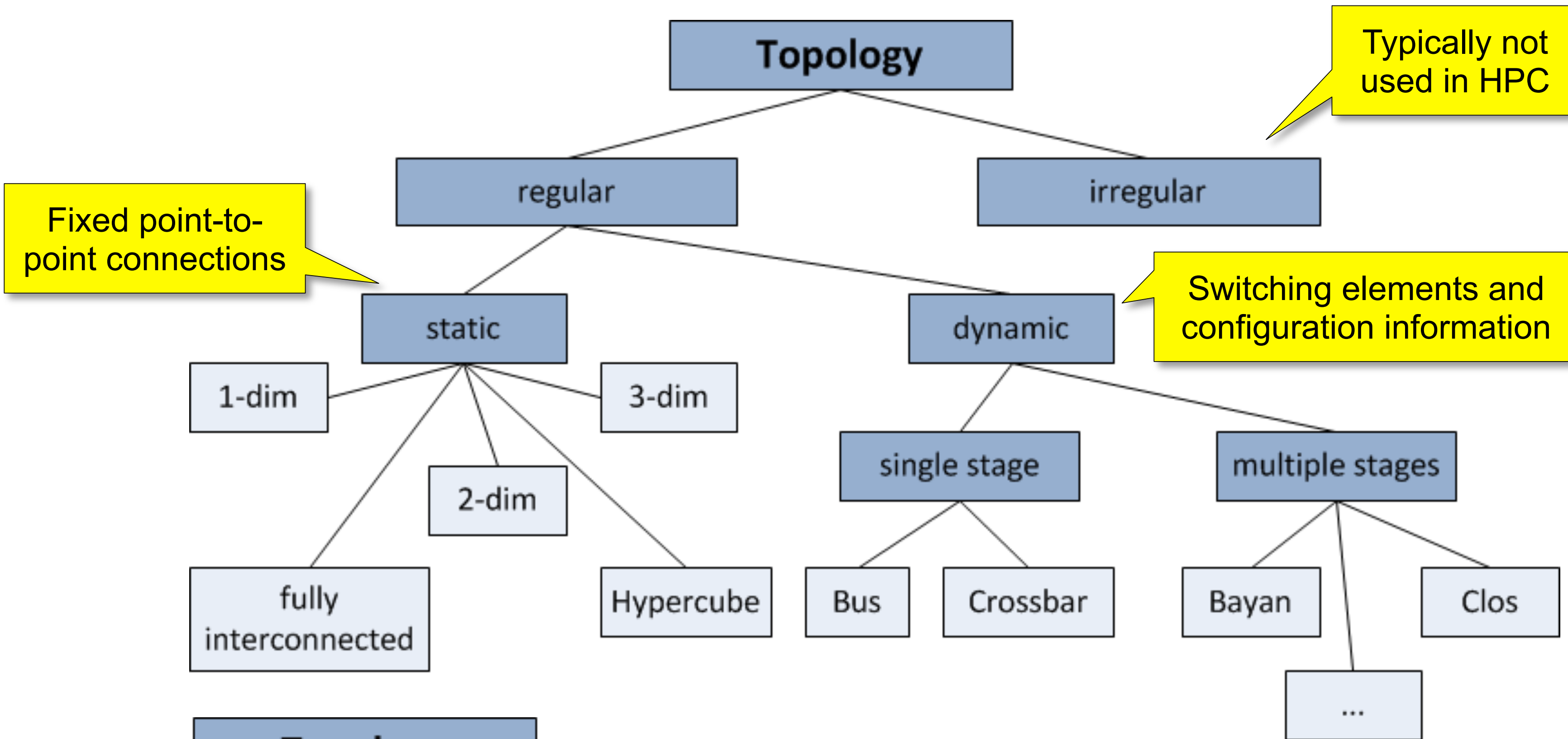
Type	Description	Length
Processor or system interconnect	Connections between processors, memory controllers, … (HyperTransport, QPI, FSB)	10..30cm
Memory network	Connections between memory controller and memory modules	10cm
I/O bus (better: interconnect)	Connection from device to system using connectors (PCI-Express)	30cm..1m
System-Area-Network	Connections within a cluster or parallel computer	5-25m
Storage-Area-Network (SAN)	Connection from processing nodes to storage modules	5-25m
Local-Area-Network (LAN)	Connection between loosely coupled workstations and/or servers (*Ethernet)	25-500m
Metropolitan-Area-Network (MAN)	Connections within the scope of city limits (ATM, FDDI)	~25km
Wide-Area-Network (WAN)	Connections without any length restrictions, worldwide, multiplexing of a large number of connections, typically using fiber optics (SONET)	unlimited

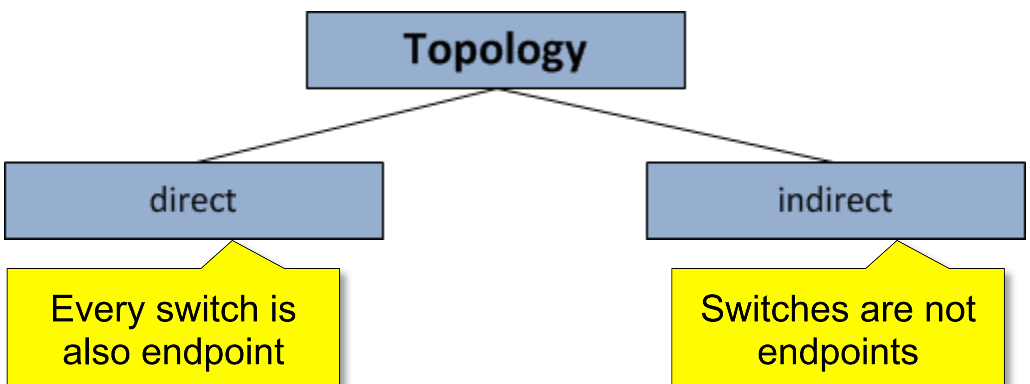
User requirements

- Costs
- Bandwidth
- Max. supported transmission length
- Scalability
- Latency
- Blocking behaviour
- Lossy/loss-less (reliability)
- In-order/out-of-order delivery

Order of importance application-dependent

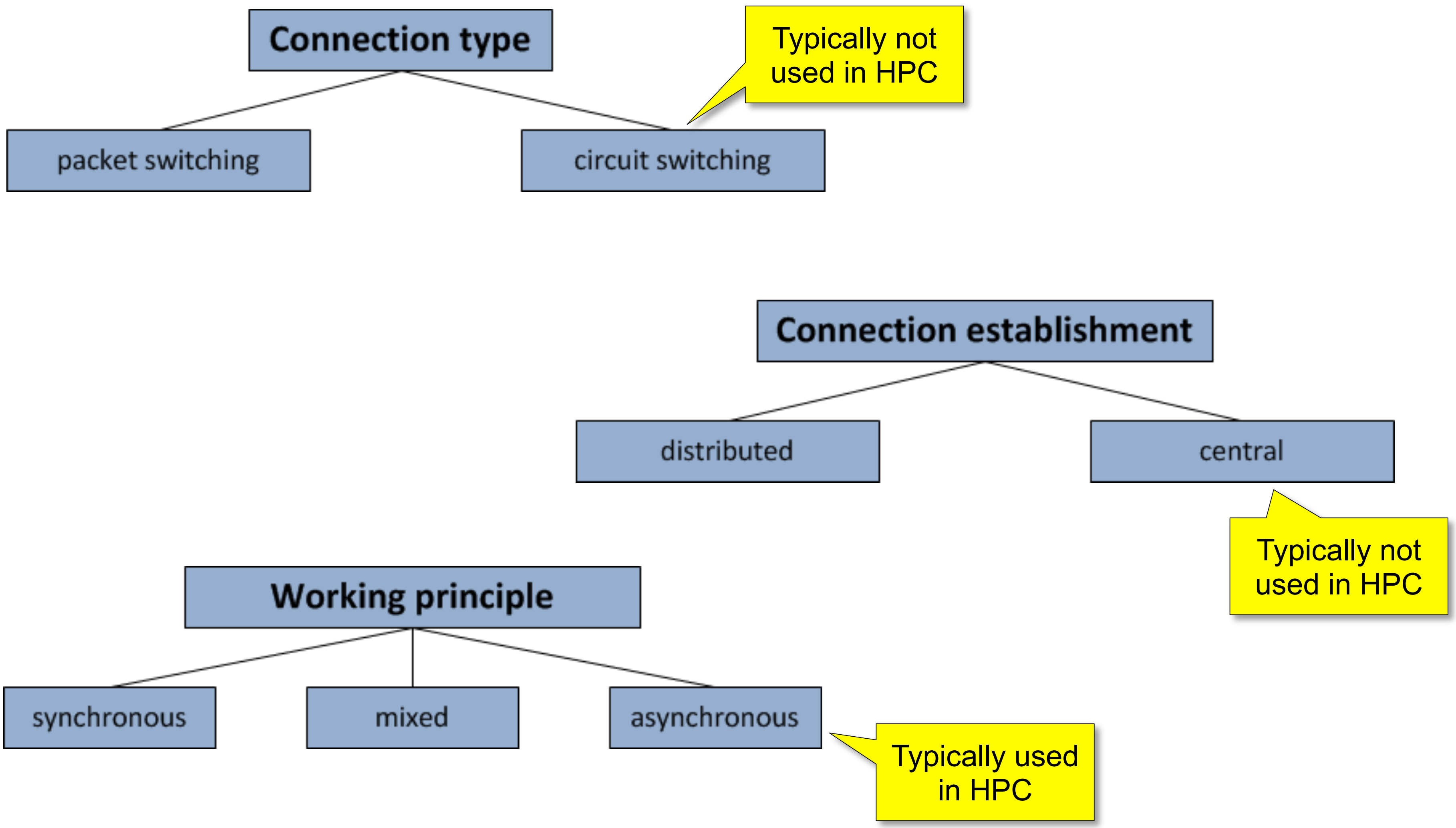
Classification





Topology most important:
resulting bandwidth translates into overall performance

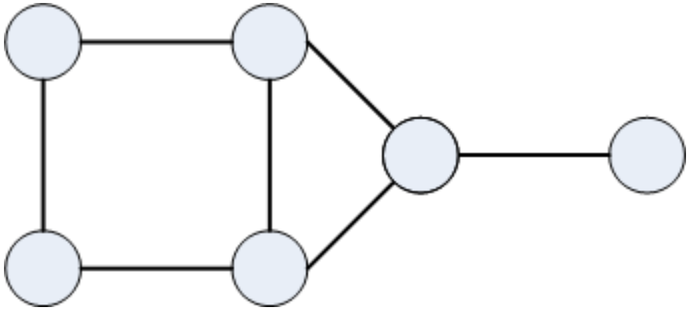
Classification



Static Topologies

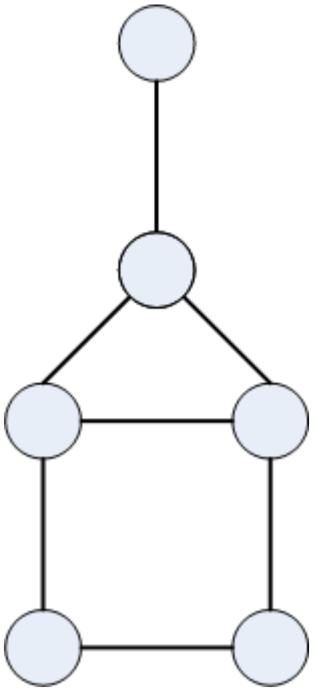
Static Topologies

- Mainly used in massively parallel processors (MPP)
  - Fixed communication structure
  - Based on point-to-point connections between processing nodes
    - Node, processor, …
- Representation
  - Node as node, connection as edge
  - Directed or undirected graph



Representation of a static IN as graph

Properties of (Static) Topologies



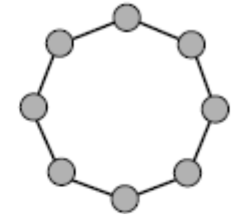
- Topological and functional properties
- **Node degree**: Number of connections per node
  - As few as possible due to costs
  - Fixed degree mandatory for scalability
- **Diameter**
  - Maximum distance in hops between any two node pairs
- **Symmetry**
  - IN is symmetric if the view of the IN from each node is identical
- More:
  - Scalability, blocking behaviour, costs, latency, fault-tolerance, max. expansion

Static Topologies - Examples

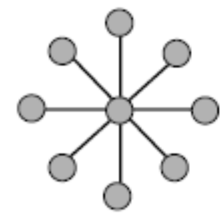
- Trivial ones
  - Chain
  - Ring
  - Star
  - (binary) Tree

Node degree?
Diameter?
Symmetry?

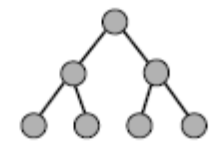
Chain



Ring

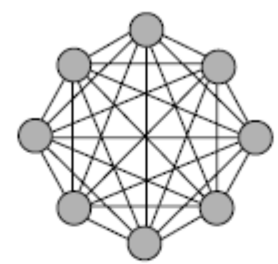


Star



(Binary) Tree

- Completely interconnected
  - Every node is connected to every other node
  - Not used in practise
  - Max. distance from one node to another = 1
    - „1 hop“

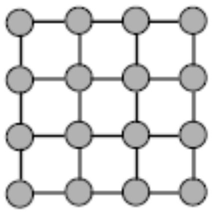


Completely interconnected

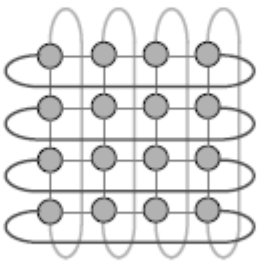
Static Topologies - Examples

Node degree?
Diameter?
Symmetry?

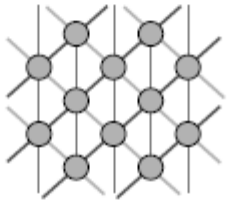
- Grid or Mesh
  - Nearest neighbor mesh
  - N-dimensional mesh suitable for n-dimensional problems
- Hexagonal grid
  - 2D, but maps nicely to 3D problems
  - Systolic algorithms
- Torus
  - Based on grid with wrap-around links
  - Better connectivity



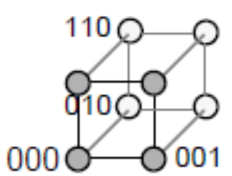
Grid



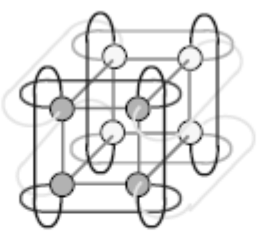
2D Torus



Hexagonal grid



Cube

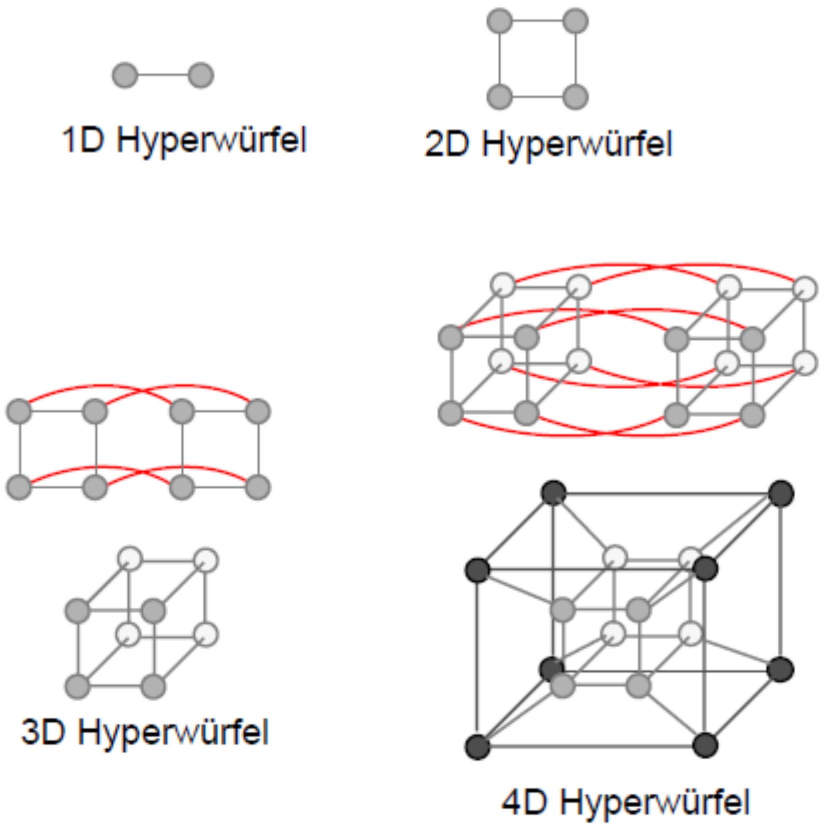


3D Torus

Static Topologies - Examples

Construction: double each node per additional dimension

- Hypercube
  - Given: n dimensions
  - 2^n nodes,
    $n * 2^n / 2$ connections,
    n connections per node
  - diameter = ***n***
- Better properties than a grid
  - Limited scalability (node degree)
- Mainly used in the beginning of MIMD-based parallel computing: nCube



Properties of Static Topologies

Translates into scalability

Only scalability with regard to topology

Topology	Node degree	Diameter	Number of connections	Scalable	Symmetric
1D grid (chain)	2	N-1	N-1	Yes	No
1D torus (ring)	2	(N-1)/2	N	Yes	Yes
2D grid	4	$2(N^{1/2}-1)$	$2N-2N^{1/2}$	Yes	No
2D torus	4	$N^{1/2}-1$	2N	Yes	Yes
3D grid	6	$3(N^{1/3}-1)$	$3N-3N^{1/3}$	Yes	No
3D torus	6	$3/2(N^{1/3}-1)$	3N	Yes	Yes
Hypercube	$\log_2 N$	$\log_2 N$	$N \log_2(N/2)$	No	Yes
Binary Tree	3	$2(\log_2 N-1)$	N-1	Yes	No
Completely interconnected	N-1	1	N(N-1)/2	No	Yes

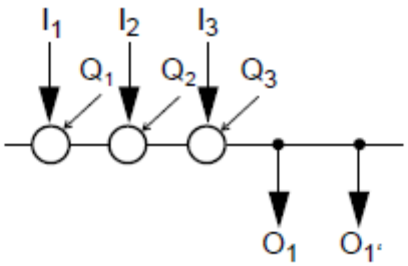
Properties of Static Topologies

- Filtering for scalability and symmetry: only **n-dimensional tori**
- Torus vs. Mesh
  - Basically only advantages for tori: highly reduced diameter, symmetric
  - Slightly higher connection count is not relevant in practise
- Side note: Binary tree
  - Disadvantage: root is bottleneck
  - Typically only used for specialized tasks: synchronization and collective communication (barrier, multi-/broad-cast)

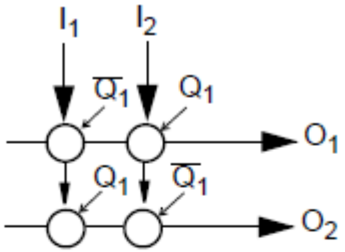
Dynamic Topologies

Dynamic Topologies

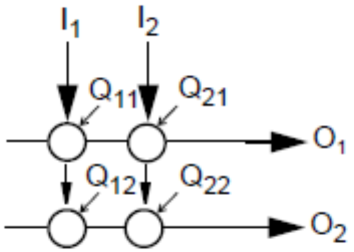
- Dynamic INs are based on **configurable switching elements**
  - Different number of stages
- Single stage
  - Building blocks
  - Shuffle, crossbar, bus
  - Representation as graph: switching elements are nodes, connections are edges
  - Control signals Q_i
  - Inputs I_i
  - Outputs O_i
- Today basically only crossbar used



Bus



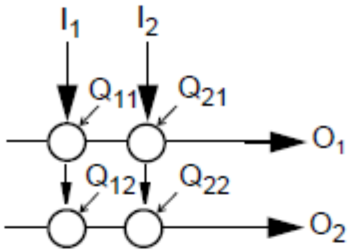
shuffle



2 x 2 Crossbar

Dynamic Topologies: Crossbar

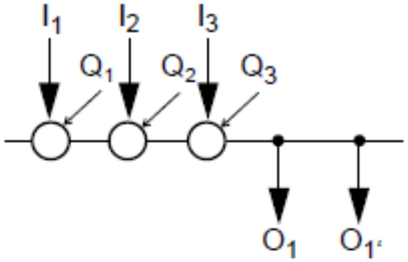
- Most universal element
- Can connect arbitrary combinations of inputs and outputs
  - Broadcast
- Conflicts avoided by arbiter
  - For all i: only one Q(i,y)=1
- Logical complexity is $O(N^2)$
  - For N inputs and N outputs
  - Due to VLSI technology basically no limitation
  - Most limiting today is pin count
    - Number of pins for a certain package
    - Pin is several orders of magnitude larger than a transistor!
      - Micrometers vs. nanometers
  - Pin limitation



2 x 2 Crossbar

Dynamic Topologies: Bus

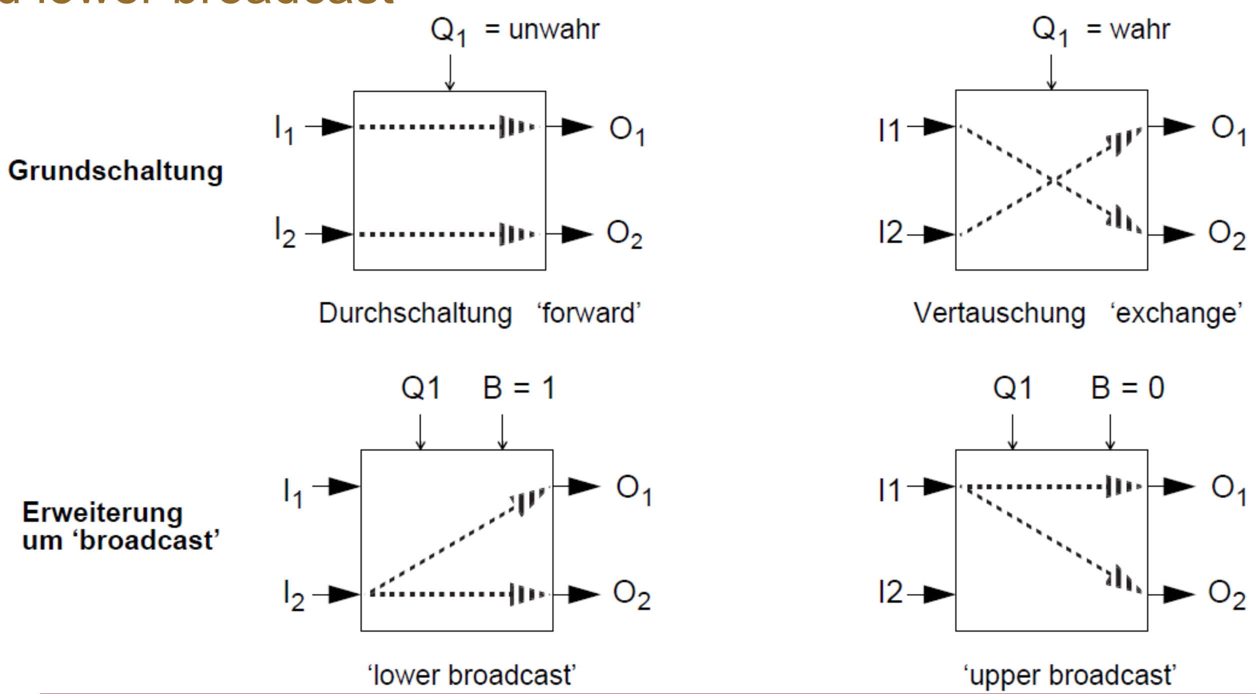
- A bus is basically a crossbar with a ***1 x m*** configuration
  - Only one driver at a time
    - High blocking potential
    - Arbiter required
  - Limited operation frequency
    - Length of connection, capacities, signal levels
  - Limited number of nodes
- Advantages: implicit broadcasts
  - Simplicity
  - Snooping protocols for cache coherency
- Today: almost vanished
  - Except human I/O and other peripherals



Bus

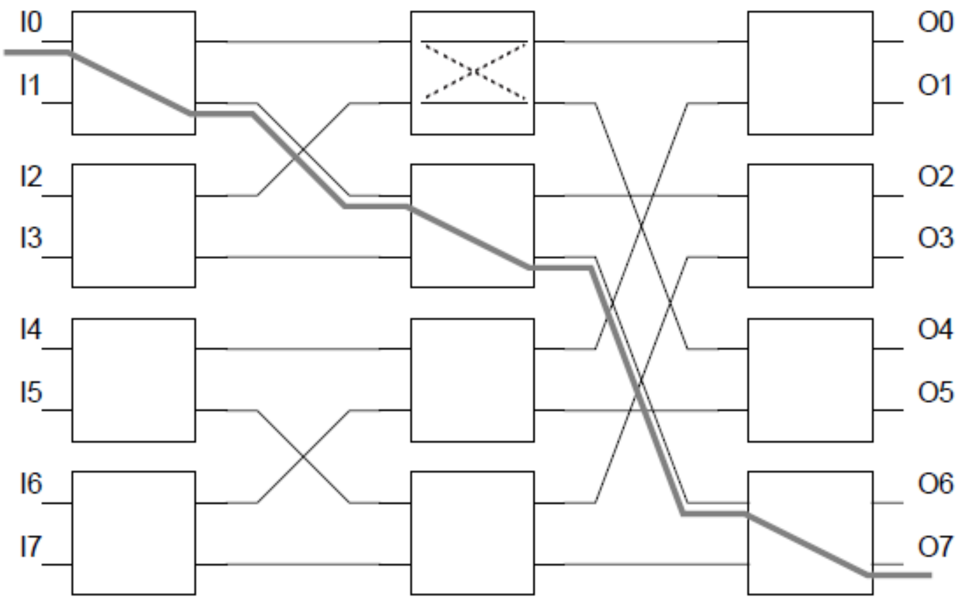
Dynamic Topologies: Shuffle

- A shuffle is basically a crossbar with a restricted set of configurations
  - „Forward“ or „exchange“
  - Possible extensions: upper and lower broadcast
- Shuffle only as 2x2 element available
  - Larger structures based on shuffle as building block



Dynamic Topologies: multi-stage

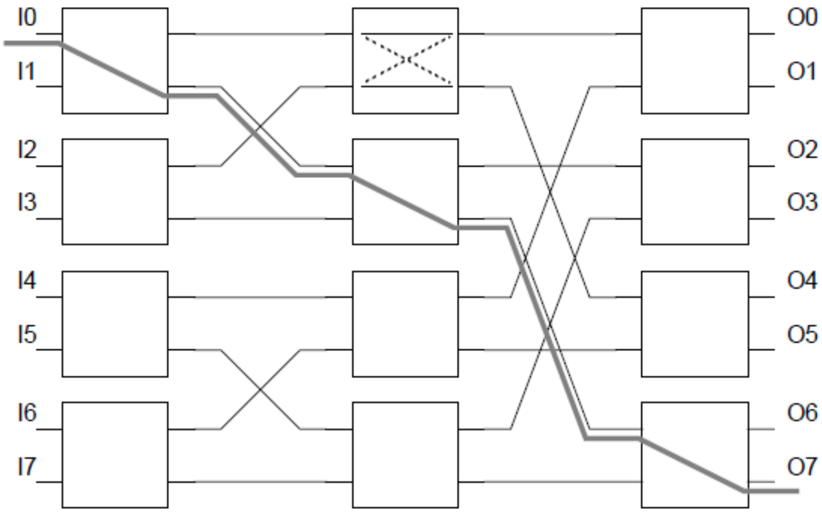
- **No single-stage element scales!**
- Multiple stages with one-staged elements as building blocks
  - (Bus,) crossbar, **shuffle**
- Examples
  - Banyan, Baseline, Cube, Delta, Flip, Indirect Cube, Omega
- Basically identical, only connectivity differs



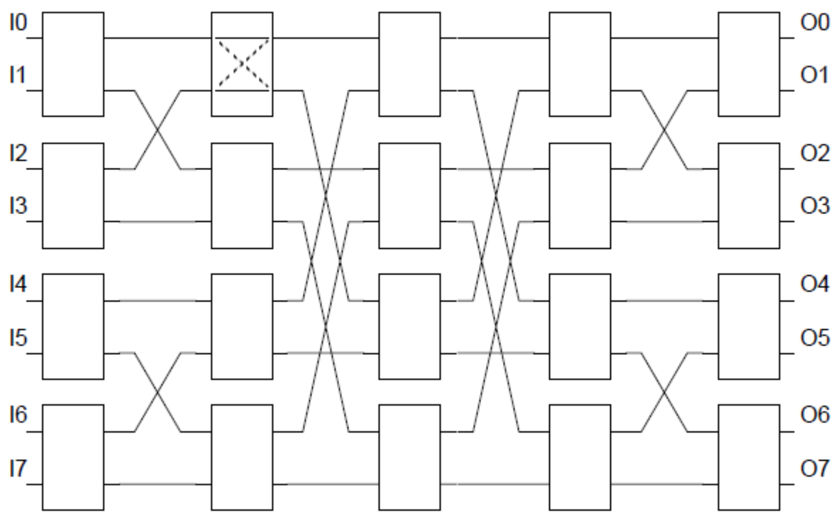
Banyan Network
Note that the shaded connection may block other connections

Dynamic Topologies: multi-stage

- Unidirectional
  - N inputs and N outputs
- Properties
  - $\log_2 N$ stages
  - $N/2 * \log_2 N$ shuffle elements
- **Blocking!**
  - Unlike crossbar
- Improved blocking behaviour
  - Additional stages
  - Benes network, composed of two banyan networks
  - ➔ Nonblocking



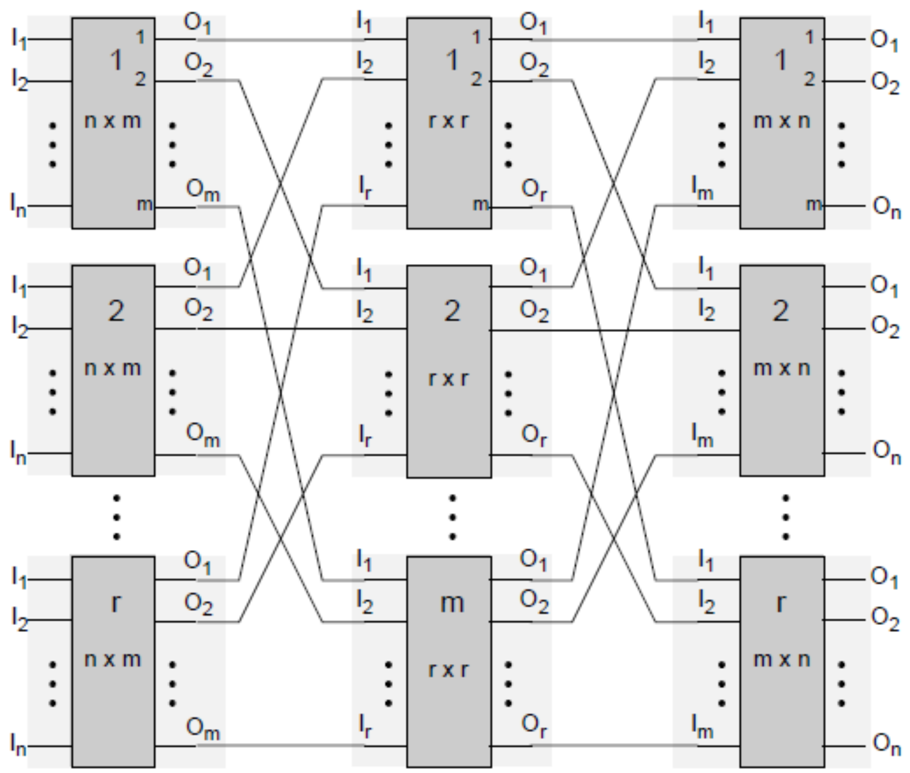
Banyan Network



Benes Network

Dynamic Topologies: multi-stage

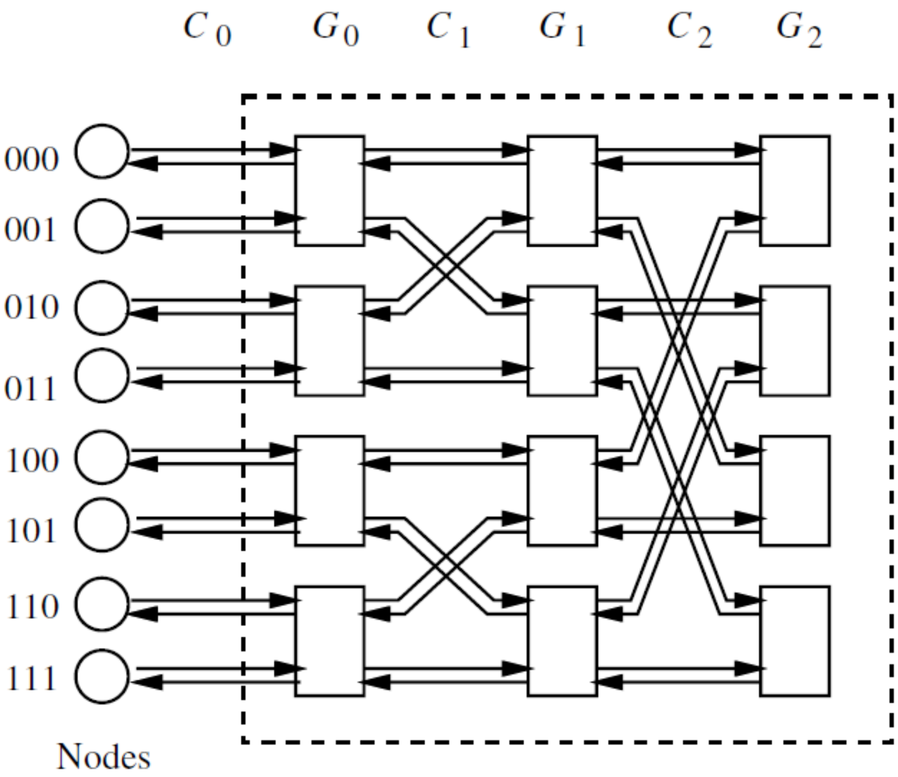
- Use of crossbars instead of shuffles: CLOS network
  - Advantages of crossbars (no blocking) and of multi-staged INs (reduced complexity)
- 1-stage CLOS: identical to crossbar
- 2-stage CLOS: blocking
- 3-stage CLOS: nonblocking (see Banyan-Benes)
- Each CLOS can be seen as a crossbar with higher degree



CLOS Network
Notice the 3 different types of XBARs used
Assuming n=m=r and a 16x16 building block: 256x256 CLOS

Dynamic Topologies: BMIN

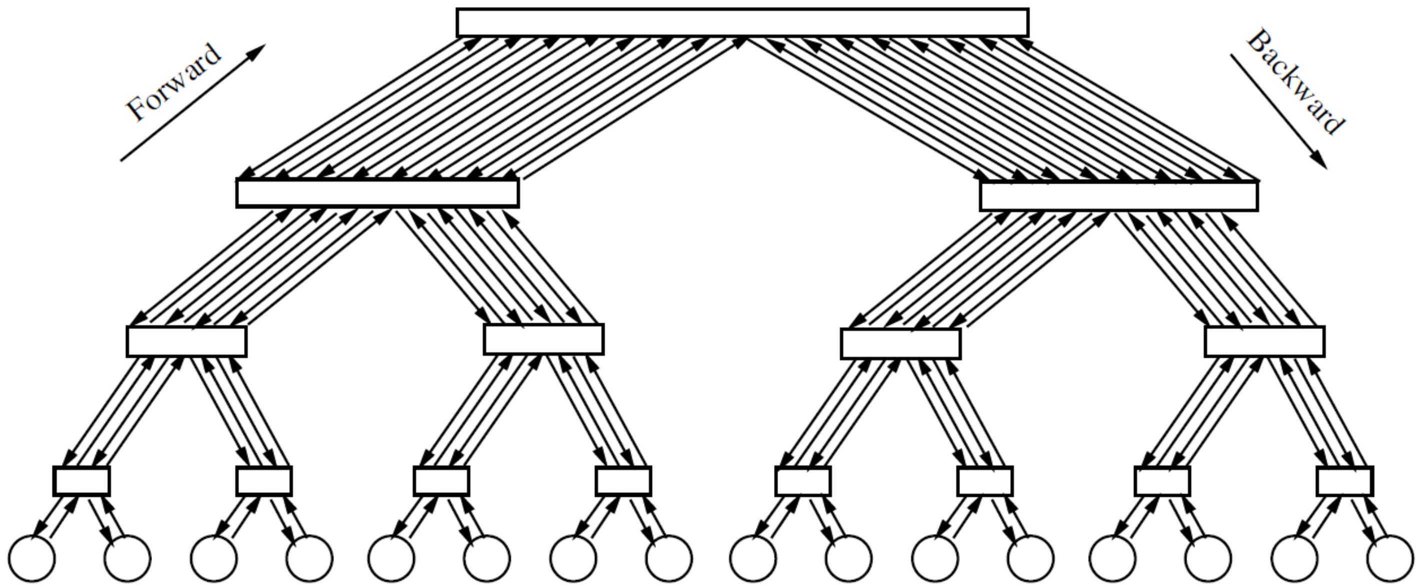
- BMIN: bi-directional multi-stage IN
  - Similar to before, but inputs/ outputs all on the left
- Switching elements are extended
  - Forward
  - Backward
  - Turnaround
- Alternate paths
  - Path diversity



8-node butterfly BMIN
[Duato et al., Interconnection Networks, 2003]

Dynamic Topologies: BMIN

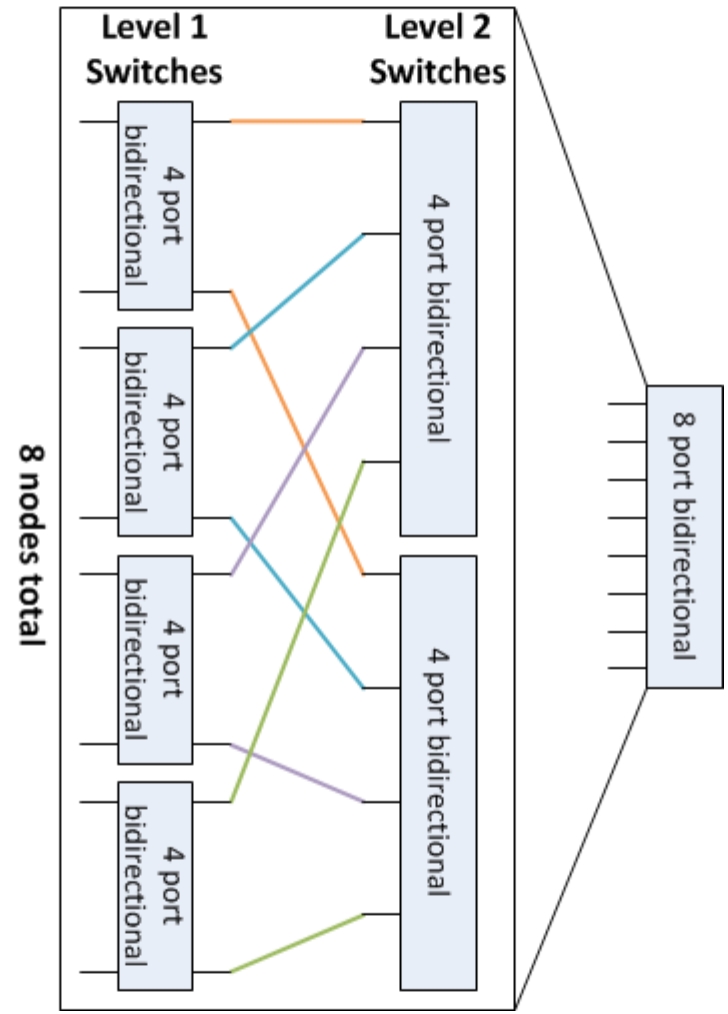
- Remember the nice scalability of binary trees
  - Replace graph nodes with switching elements and increase number of connections accordingly
- Fat Tree
  - Bottleneck at root is avoided by appropriate provisioning
  - Typically uses crossbars
  - Main disadvantage: heterogeneity, different elements required



BMIN with turnaround viewed as **Fat Tree** – switching elements are typically crossbars or CLOS

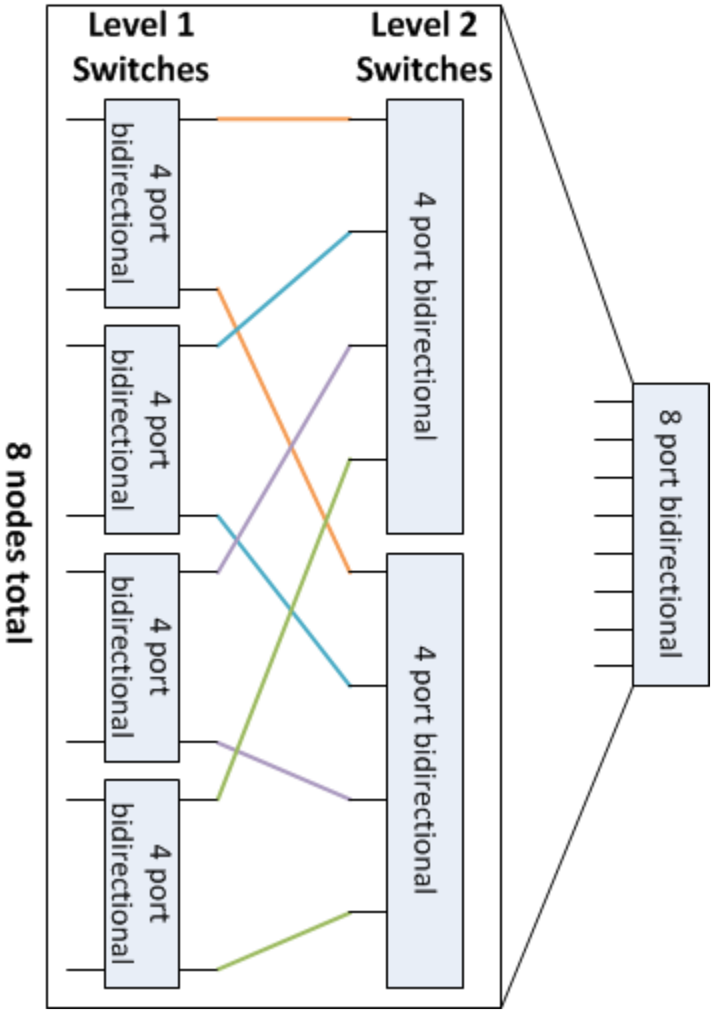
Dynamic Topologies: BMIN

- Typically complete Fat Tree is based on one building block
- ***n***-port crossbar switch (single chip)
- „Fatter“ switches constructed out of this switch in a CLOS fashion
  - 2 stages: max. ($\boldsymbol{n^2}/2$) end points
  - 3 stages: max. ($\boldsymbol{n^3}/4$) end points
- User point of view:
  - Non-blocking fat crossbar
  - But number of internal stages may increase ➔ hop latency increases!



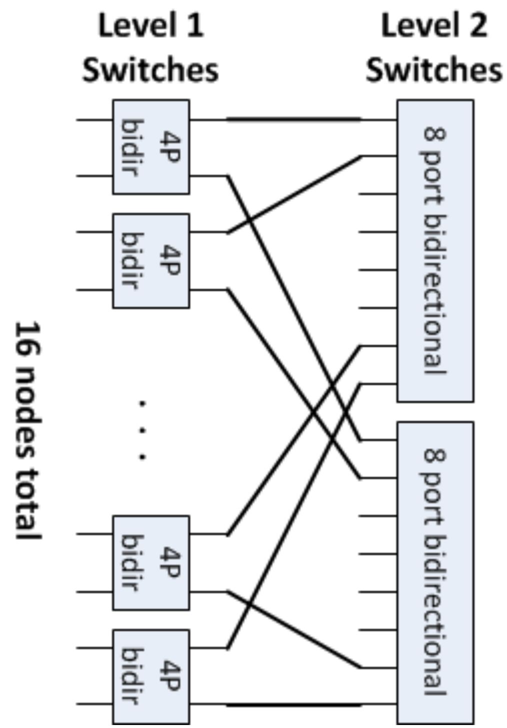
Constructing a 8 port BMIN with 4-port building blocks

Dynamic Topologies: BMIN



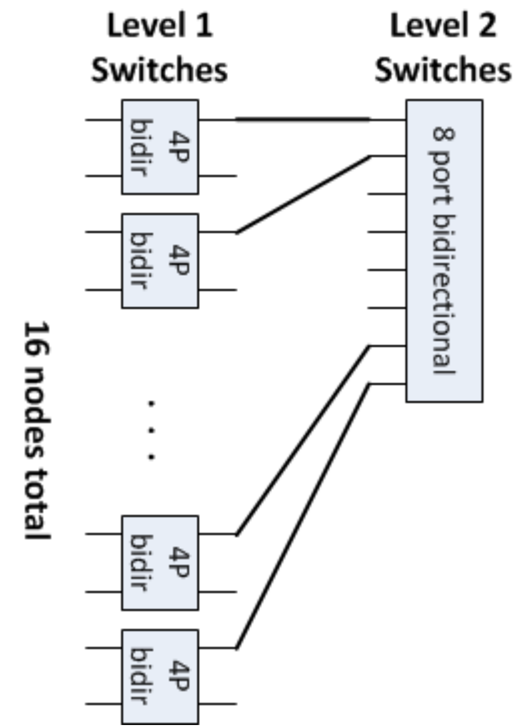
Constructing a 8 port BMIN with 4-port building blocks

3 stages - nonblocking



Larger configuration based on element on the left
Full bisection configuration

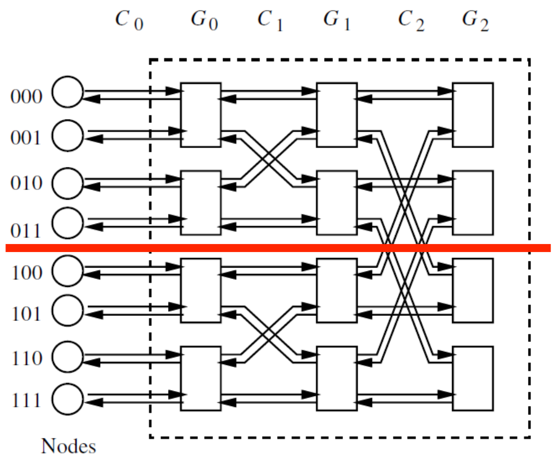
3 stages – blocking

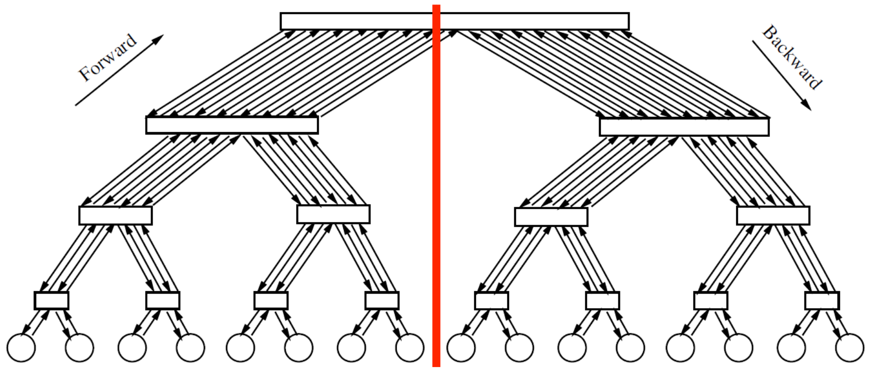


Larger configuration based on element on the left
Reduced bisection configuration

Bisection bandwidth

- Bisection: Segmentation of an IN into two equal parts
  - As few cuts as possible
- Bisection BW: sum of the data rate of all cutted links
- The higher the bisection BW is, the lower is the blocking potential
  - Uniform traffic: ½ of traffic crosses bisection





Overview of Properties

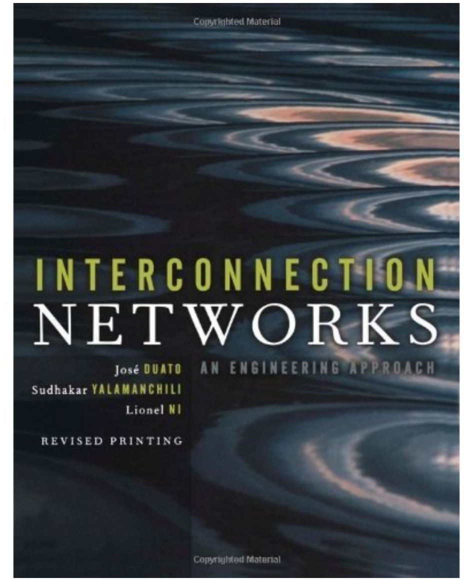
Topology	Node degree	Diameter	Number of connections	Scalable	Symmetric	Bisection
2D grid	4	$2(N^{1/2}-1)$	$2N-2N^{1/2}$	Yes	No	$N^{1/2}$
2D torus	4	$N^{1/2}-1$	2N	Yes	Yes	$2N^{1/2}$
3D grid	6	$3(N^{1/3}-1)$	$3N-3N^{1/3}$	Yes	No	$N^{2/3}$
3D torus	6	$3/2(N^{1/3}-1)$	3N	Yes	Yes	$2N^{2/3}$
Hypercube	$\log_2 N$	$\log_2 N$	$N \log_2(N/2)$	No	Yes	$2^{(\log_2 N)-1}$
Crossbar	1	1	N^2	No	Yes	N/2
CLOS	1	3	r(2n+2m) (4N² for r=n=m)	Yes	Yes	N/2
Fat Tree, *S is number of stages*	1	2(S-1); *S=O(logN)*	N*S	Yes	Yes	N/2

Summary

- Interconnection networks as key in HPC
  - INs are pervasive today: from smartphones to microcontrollers to large computing facilities
- Topologies and their properties
  - Direct vs. indirect
  - Static vs. dynamic
  - Node degree, diameter, number of connections, symmetry, scalable, (non-)blocking
- Bisection bandwidth
- Many more topologies possible
  - Regular but hierarchical
  - Irregular

Credits & Further Reading

- Duato, Yalamanchili, Ni: Interconnection Networks - An Engineering Approach. 2002



- Dally, Towles: Principles and Practices of Interconnection Networks. 2003