Interconnection Networks

2010-08-25

Michael Schliephake

KTH PDC - michs@kth.se

Agenda

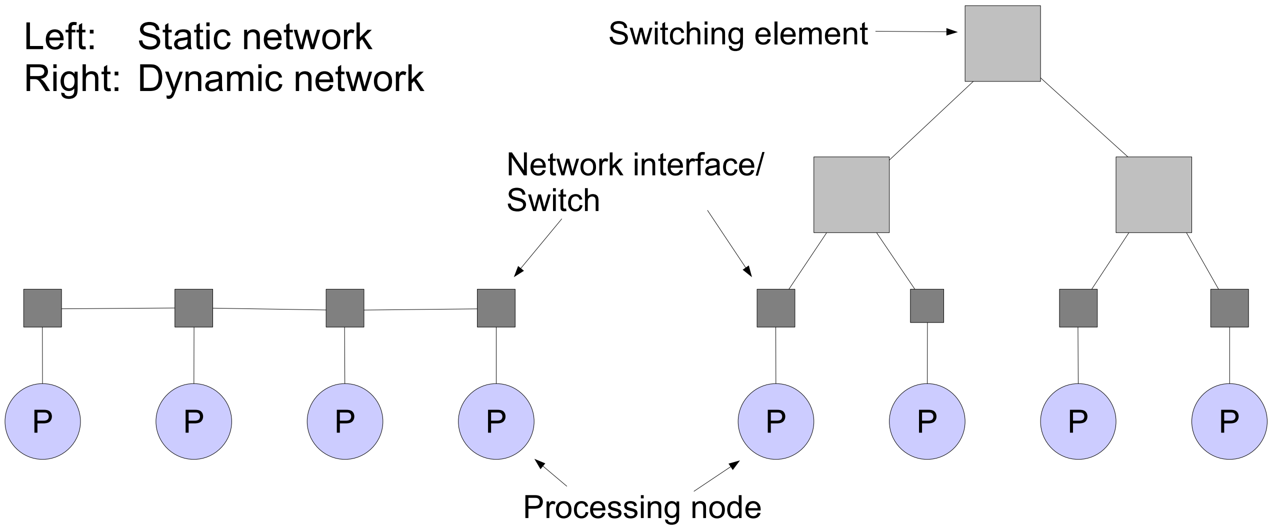
1. Introduction
2. Static Networks
3. Dynamic Networks
4. Routing and Switching
5. Communication Operations

1. Introduction

Interconnection Networks I

- Data transfer between
  - Processing nodes
  - Processors and memory
- Abstract view: n inputs, m outputs
- Typical components: links, switches, interfaces

Left: Static network
Right: Dynamic network



Interconnection Networks II

- Topology – describes the geometric structure
  - Graph – switches, processors, memory as nodes; connection links as edges
  - Static networks = Direct or Point-to-point networks
  - Dynamic networks = Indirect networks
- Routing technology – defines how and along which path messages are transported
  - Routing = routing algorithm selects a path
  - Switching strategy = defines segmentation of messages, mapping to a path, handling by switches and processors

Topology and Routing technology determine decisive the performance of the communication

Criteria for Networks I

Diameter

Maximum distance between any two processing nodes

- Distance: Shortest Path (number of links) between two nodes
- Measure for the time to transfer messages between arbitrary nodes

Criteria for Networks I

Degree

Maximum degree of all processing nodes

- Degree of a node is the number of in- and outgoing links of a node
- Measure for the number of simultaneously active communication connections
- Measure for hardware efforts

2. Static Networks

Criteria for Networks II

Connectivity

Measure of the multiplicity of pathes between arbitrary processing nodes

- High connectivity lowers contention, increases reliability
- Arc (node) connectivity = number of links (nodes) to remove to separate the network in two disconnected networks

2. Static Networks

Criteria for Networks III

Bisection Width, Bisection Bandwidth

Bisection width = minimum number of links to remove to get two equal halves,

Bisection bandwidth = minimum volume of communication between the halves

- Bisection bandwidth
  = bisection width x channel bandwidth
- Measure of loading capacity: simultaneously transmission of „bisection width + 1“ messages may saturate the network

2. Static Networks

Criteria for Networks IV

Cost

- Many criteria possible
  - Number of communication links, number of wires
  - Bisection bandwidth
- Technology
- Additional equipment

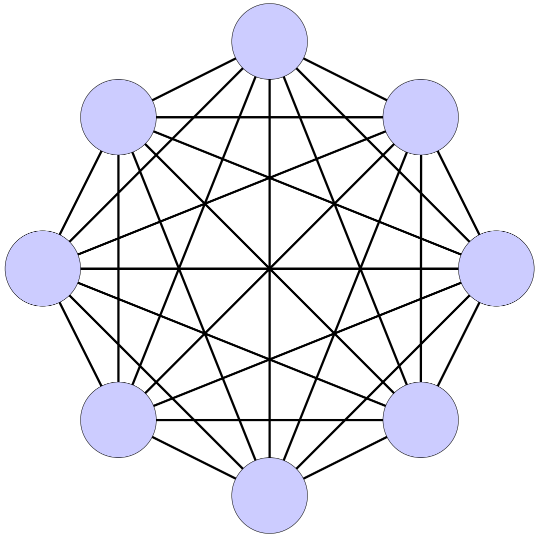
2. Static Networks

Criteria for Networks IV

General Requirements

- Small diameter
  - ⇒ short distances for transmissions
- Small node degree
  - □ reduction of the hardware efforts
- High bisection bandwidth
  - □ high throughput
- High connectivity
  - □ high reliability
- Good extensibility

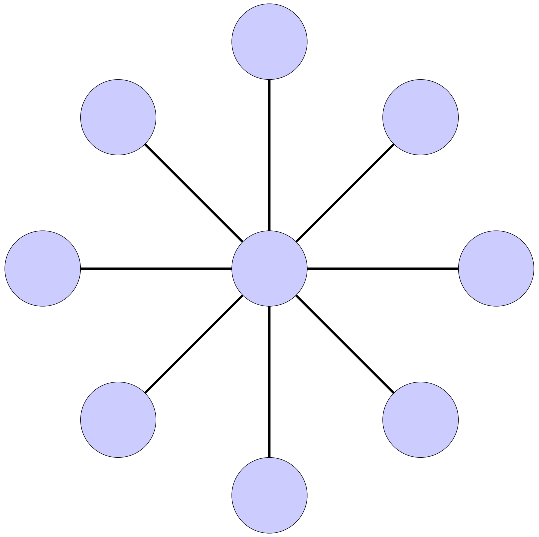
Completely connected



Diameter	1
Degree	$p-1$
Arc connectivity	$p-1$
Bisection bandwidth	$(\frac{p}{2})^2$
Cost (# links)	$\frac{p\cdot(p-1)}{2}$

Topologies II

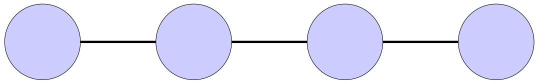
Star



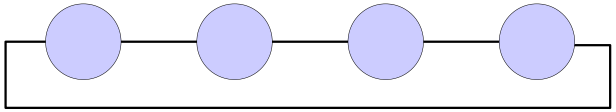
Diameter	2
Degree	$p-1$
Arc connectivity	1
Bisection bandwidth	1
Cost (# links)	$p-1$

2. Static Networks
Topologies III

Linear Array



Linear Array (no wraparound)



Ring

	Lin. Array	Ring
Diameter	$p-1$	$\lfloor \frac{p}{2} \rfloor$
Degree	2	2
Arc connectivity	1	2
Bisection bandwidth	1	2
Cost (# links)	$p-1$	p

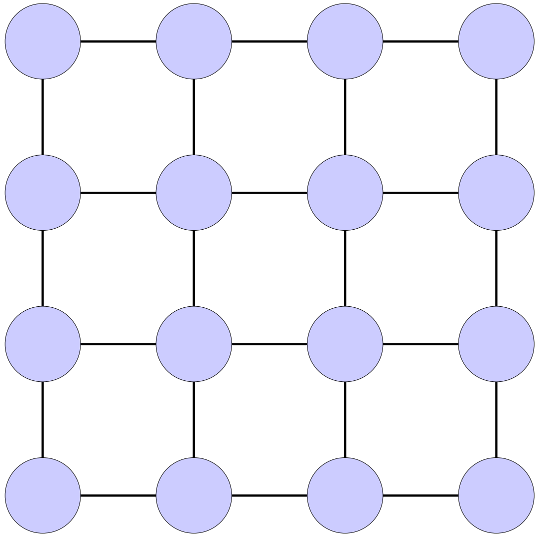
2. Static Networks
Topologies IV

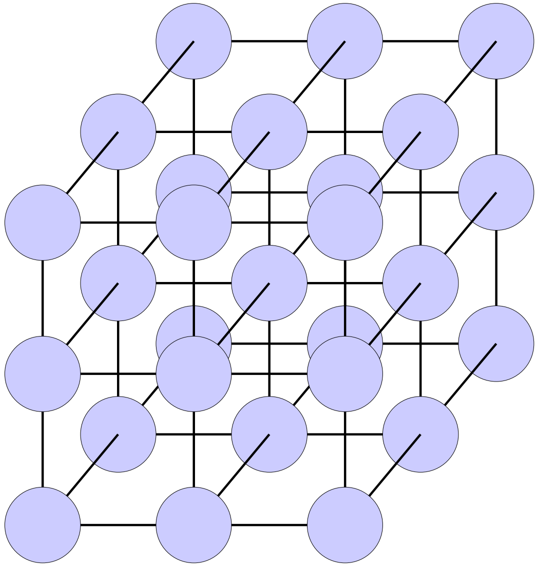
Mesh

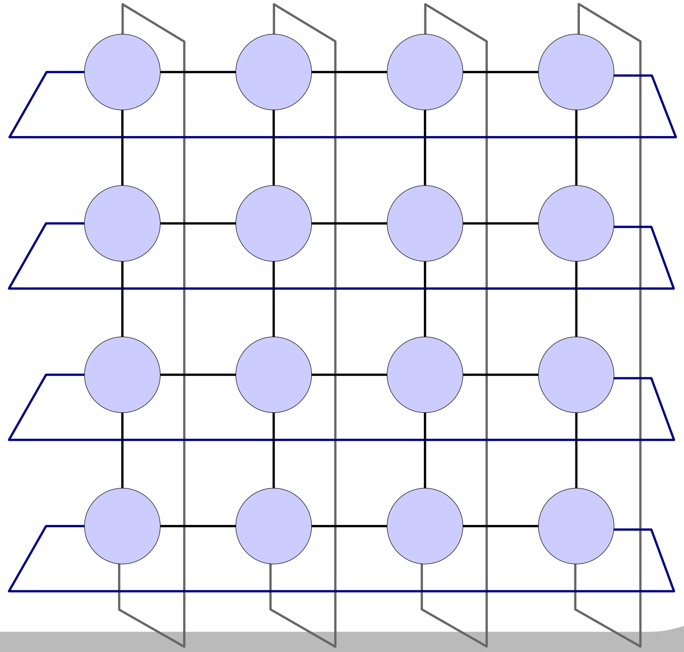
- Mesh (d dimensions)
- Torus (d dimensions)

$$p=r^d$$

	Mesh	Torus
Diameter	$d\cdot(\sqrt[d]{p}-1)$	$d\cdot\llbracket\frac{\sqrt[d]{p}}{2}\rrbracket$
Degree	$2\cdot d$	$2\cdot d$
Arc connectivity	d	$2\cdot d$
Bisection bandwidth	$p^{\frac{d-1}{d}}$	$2\cdot p^{\frac{d-1}{d}}$
Cost (# links)	$d\cdot(p-\sqrt[d]{p})$	$d\cdot p$





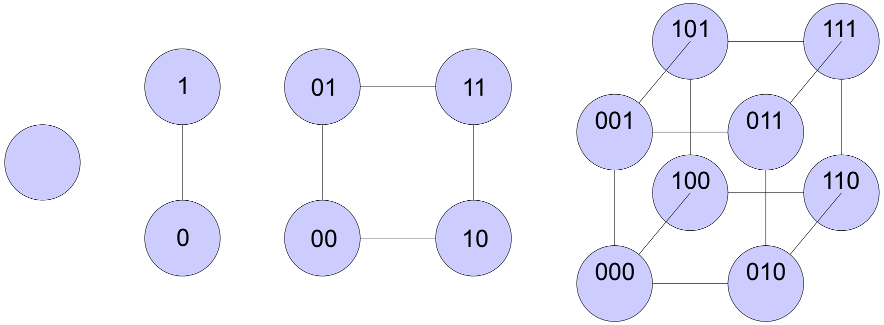


2. Static Networks

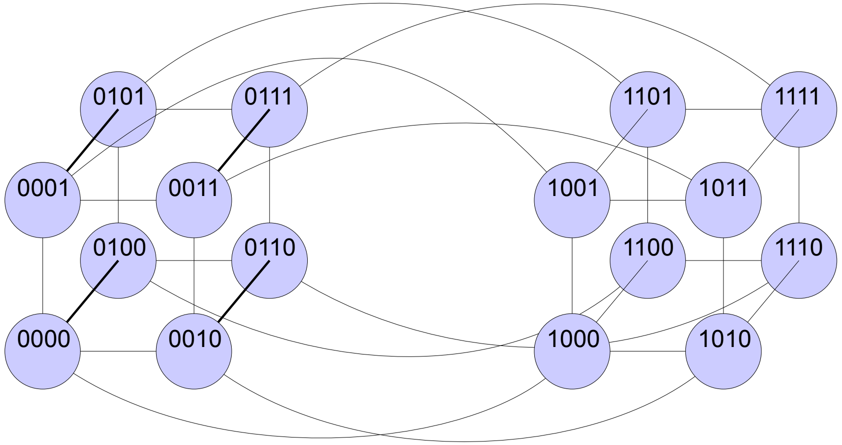
Topologies V

Hypercube

$$p = 2^d$$



0-d 1-d 2-d 3-d



4-d

Diameter	$\log(p)$
Degree	$\log(p)$
Arc connectivity	$\log(p)$
Bisection bandwidth	$\frac{p}{2}$
Cost (# links)	$\frac{p \cdot \log(p)}{2}$

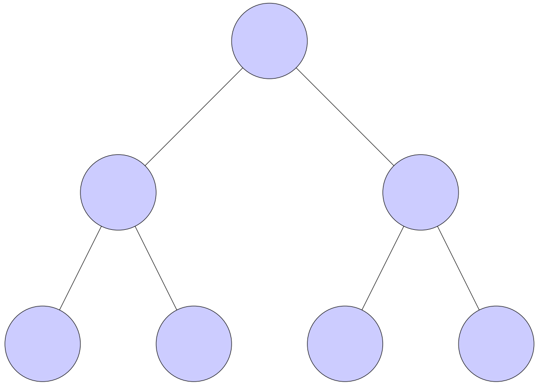
Topologies VI

(Static) Tree

$$p=2^k-1$$

Compl. binary tree

Diameter	$2\cdot\log(\frac{n+1}{2})$
Degree	3
Arc connectivity	1
Bisection bandwidth	1
Cost (# links)	$p-1$



Complete binary tree

2. Static Networks
Topologies VII

Summary table

	Compl. Connect.	Star	Lin. Array	Ring	Mesh	Torus	Hyper-cube	Binary Tree
Diameter	1	2	$p-1$	$\llbracket \frac{p}{2} \rrbracket$	$d\cdot(\sqrt[d]{p}-1)$	$d\cdot\llbracket \frac{\sqrt[d]{p}}{2} \rrbracket$	$\log(p)$	$2\cdot\log(\frac{n+1}{2})$
Degree	$p-1$	$p-1$	2	2	$2\cdot d$	$2\cdot d$	$\log(p)$	3
Arc connectivity	$p-1$	1	1	2	d	$2\cdot d$	$\log(p)$	1
Bisection bandwidth	$(\frac{p}{2})^2$	1	1	2	$p^{\frac{d-1}{d}}$	$2\cdot p^{\frac{d-1}{d}}$	$\frac{p}{2}$	1
Cost (# links)	$\frac{p\cdot(p-1)}{2}$	$p-1$	$p-1$	p	$d\cdot(p-\sqrt[d]{p})$	$d\cdot p$	$\frac{p\cdot\log(p)}{2}$	$p-1$

4. Dynamic Networks

Criteria for Networks I

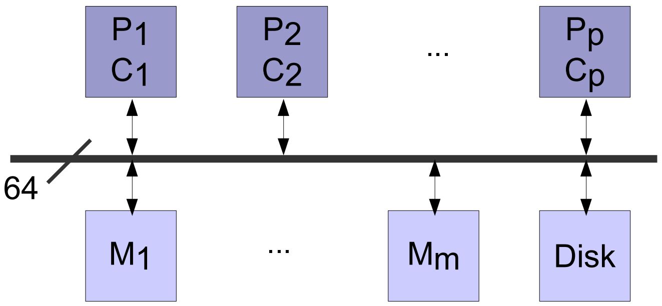
- Similar to static networks
  - Processing in switches costs time ⇒ seen as nodes
- Diameter = maximum distance between any node (in practice processing nodes)
- Node and edge connectivity = number of nodes or edges to remove to get two networks
- Bisection bandwidth = any possible partitioning of processing nodes into two equal parts to consider □ induces partitioning of switching nodes with minimized number of crossed edges

4. Dynamic Networks

Bus, Crossbar

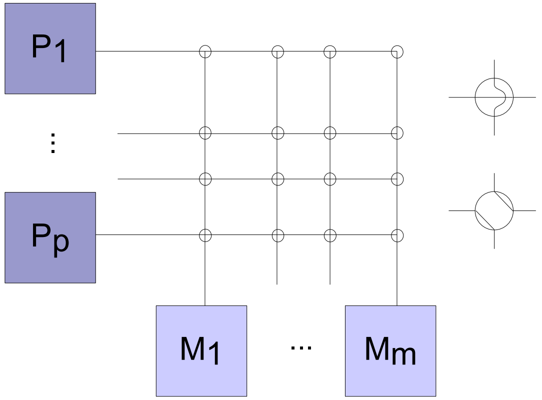
Bus

- Simple, cheap
- Constant distance
- Good for Broadcasts
- Scaling limited by performance



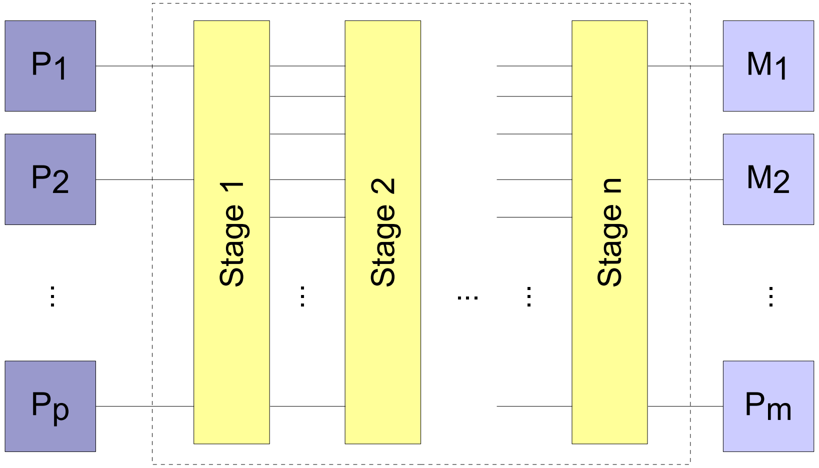
Crossbar

- Complex, expensive
- Non-Blocking
- Realization hard for large p and high speed
- Scaling limited by cost

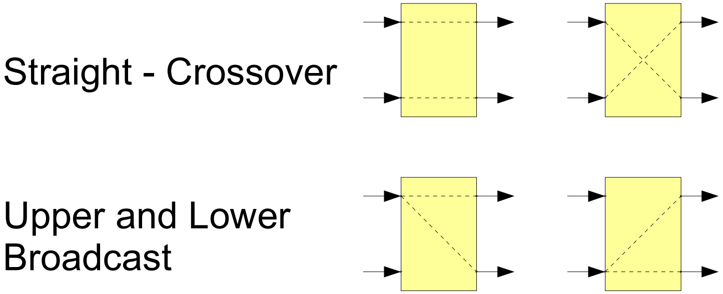


4. Dynamic Networks
Multistage Interconnection Networks I

- Intermediate features between bus and crossbar
- Characteristics
  - constuction rule
  - degree of switching nodes
- Examples
  - omega network
  - baseline network
  - butterfly network
  - benes network



Schematic view of a multistage interconnect network



Switch positions for a 2x2 switch

4. Dynamic Networks

Multistage Interconnection Networks II

Omega network

- Example of a blocking network
- log(p) stages
- Construction rule:

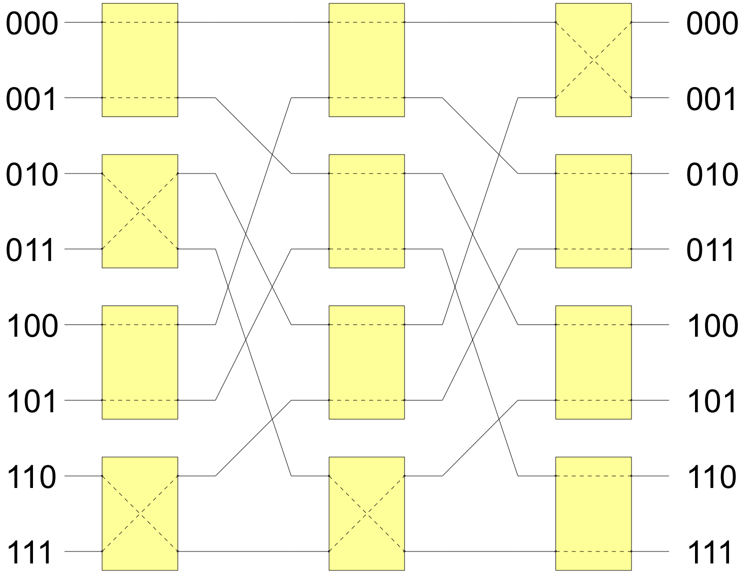
$$j = \begin{cases} 2i, & 0 \le i \le \frac{p}{2} \\ 2i+1-p, & \frac{p}{2} \le i \le p-1 \end{cases}$$

(perfect shuffle)

- Number of switching nodes

$$\frac{p}{2} \cdot \log(p)$$

$$\pi^8 = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 4 & 7 & 0 & 2 & 6 & 5 & 3 \end{pmatrix}$$

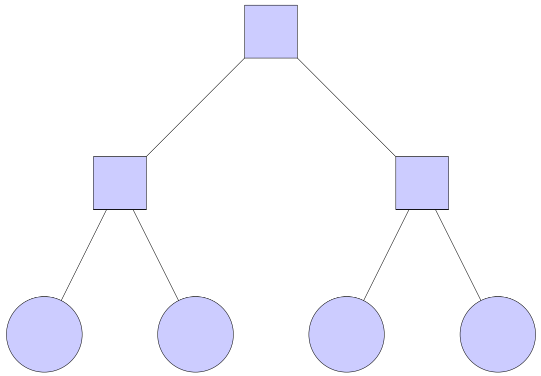


Realisation of an omega network with 2 x 2 switches

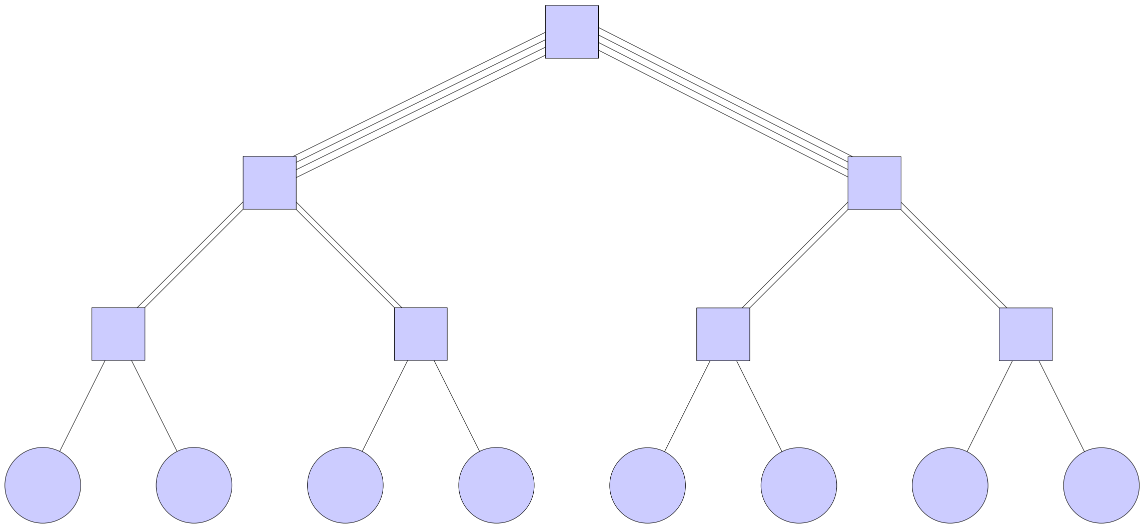
3. Dynamic Networks
Tree-based networks

(Dynamic) Tree

- Nodes at intermediate levels are switches, processing nodes are leafs
- Communication bottleneck at higher levels



Tree with dynamic network



Fat tree

Properties

Summary table

	Crossbar	Omega Network	Dynamic Tree
Diameter	1	$\log(p)$	$2\cdot\log(p)$
Arc connectivity	1	2	2
Bisection bandwidth	p	$\frac{p}{2}$	1
Cost (# links)	p^2	$\frac{p}{2}$	$p-1$

Routing

- Routing algorithm defines a path to send messages between nodes
- Requirements
  - Deadlock-free
  - Consideration of the topology
  - Avoid Contention
  - Avoid Congestion
- Types of algorithms
  - minimal, non-minimal
  - deterministic, adaptive

4. Routing and Switching

Switching

- Switching strategy defines how a message travels along a routing path
  - Segmentation
  - Allocation type of the path
  - How messages are processed in switching nodes

⇒ Strong influence on the transfer time of messages between nodes

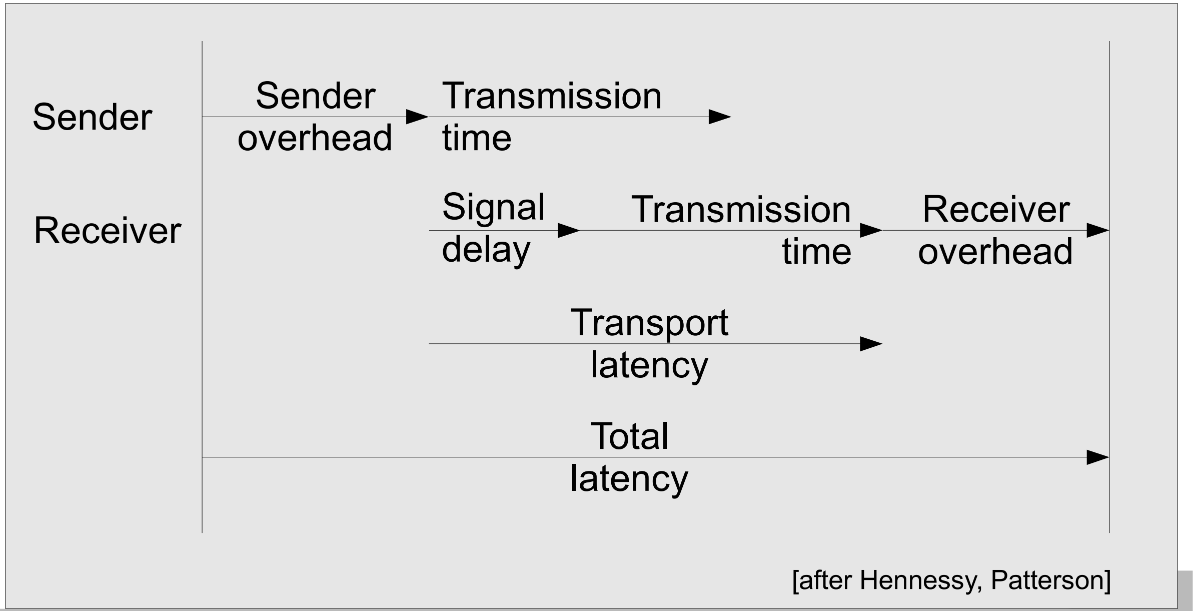
5. Communication Operations
Message Passing Costs

- Message size $m[B]$
- Bandwidth $[B\cdot s^{-1}]$
- Byte transfer time $t_B = \frac{1}{bandwidth}$
- Transfer time $\frac{m}{bandwidth} = t_B m$
- Hop time t_h
- Signal Delay
- Transport Latency
- Sender overhead
- Receiver overhead

- Latency

$$\text{Latency} = \text{Overhead} + \text{Transfer time}$$

$$t(m) = t_s + t_B m$$



[after Hennessy, Patterson]

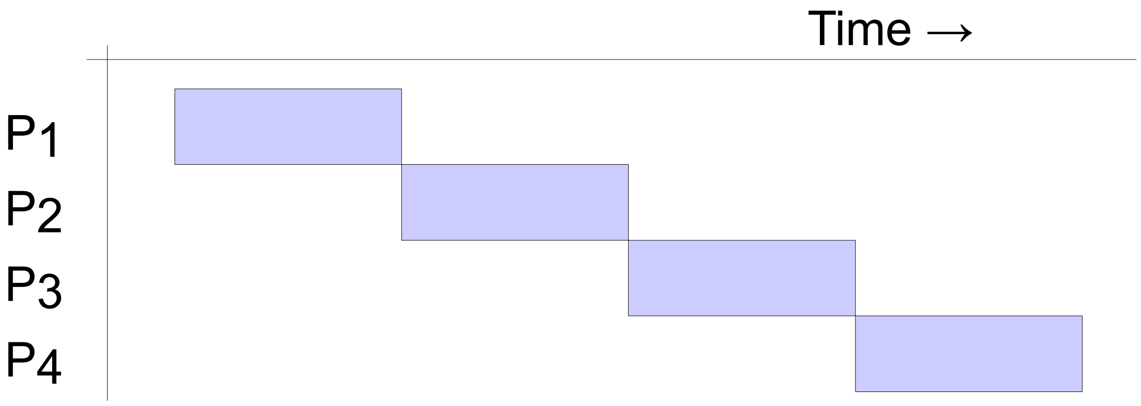
4. Routing and Switching
Store-and-Forward Routing

- Message is transferred completely between nodes on the path
- Communication time for path of length l:

$$t_{comm} = t_s + (m\,t_B + t_h)\,l$$

simplified for modern equipment:

$$t_{comm} = t_s + m\,l\,t_B$$



Store-and-Forward Routing with 4 nodes

Packet Switching

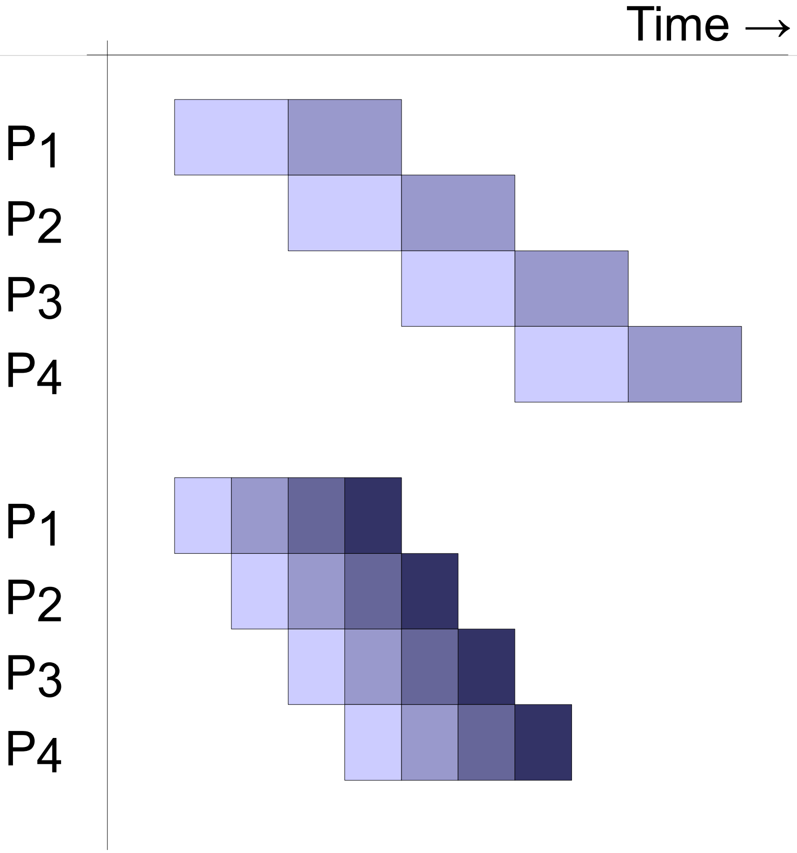
- Divide mesage in packages to reduces transfer time
- Other advantages
  - Packet losses cheaper
  - Packages can use different pathes
  - Better error correction possibilities
- Communication time (static route)

$$t_{comm} = t_s + t_h l + t_{B'} m$$

$$t_{B'} = t_p + t_B \left(1 + \frac{s}{r} \right)$$

$t_p \ldots$ effort to packetize message,

$r \ldots$ message length in packet, $s \ldots$ header size of packet



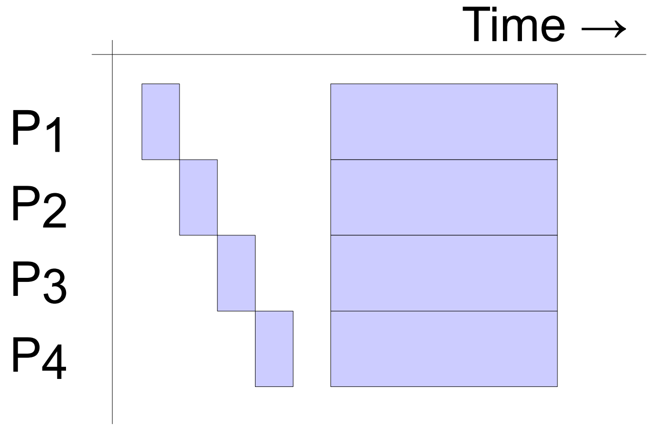
Packet Routing with 4 nodes
for a message divided in packages

4. Routing and Switching
Circuit Switching

- Path established through the sending of a control message, exists until the the communication ends
- All nodes active at the same time to transfer the message
- Communication time for path of length l:

$$t_{comm} = t_s + t_B (m_c l + m)$$

$m_c \ldots$ length of control message



Circuit switching with 4 nodes

4. Routing and Switching

Cut-Through Routing

- Virtual Cut-Through Routing
  - Packages are subdivided and transported in a pipelined manner after evaluation of the header
    - Message is divided into „flow control units“ (flits) – smaller than packets
  - Collection at nodes where route is blocked
- Variant Wormhole Routing
  - Flits are blocked at their current position
- Advantages:
  - Safe of intermediate stores and sends
  - Reduced buffer size

Communication Time: $t_{comm} = t_s + l\, t_h + t_B m$

4. Routing and Switching

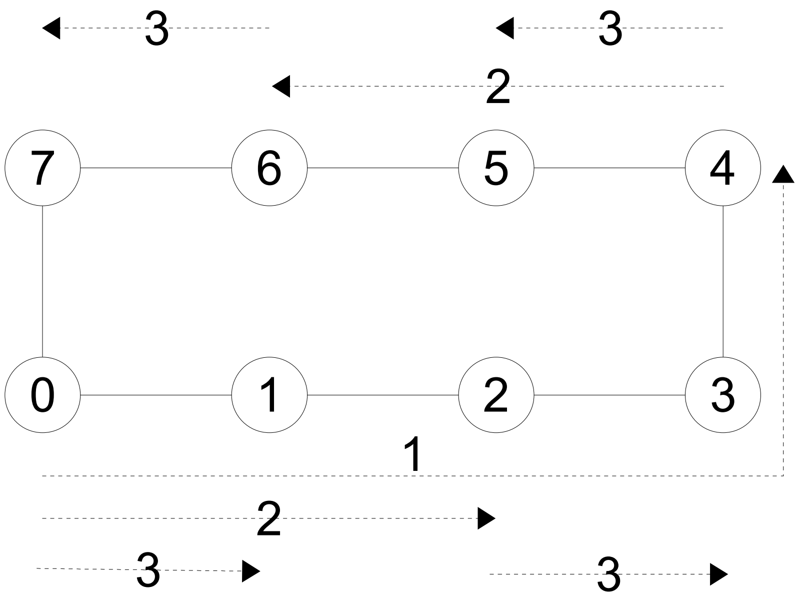
Simplified Cost Model

- Cut-Through-Routing $t_{comm} = t_s + l t_h + t_B m$
  - Prefer communication in bulk
  - Minimization of the transfer distance
  - Minimization of the data volume
- Reality allows simplification $t_{comm} = t_s + t_B m$
  - Limited influence on process mapping
  - Often randomized routing
  - Per-Hop time can be ignored often
- Consequences for programmer
  - Assumes the same time between arbitrary nodes (= assume completely connected network)
  - Accuracy loss: Only valid in networks without congestion
    - Topologies are sensible to congestion in different grade
    - Communication patterns produce different congestion

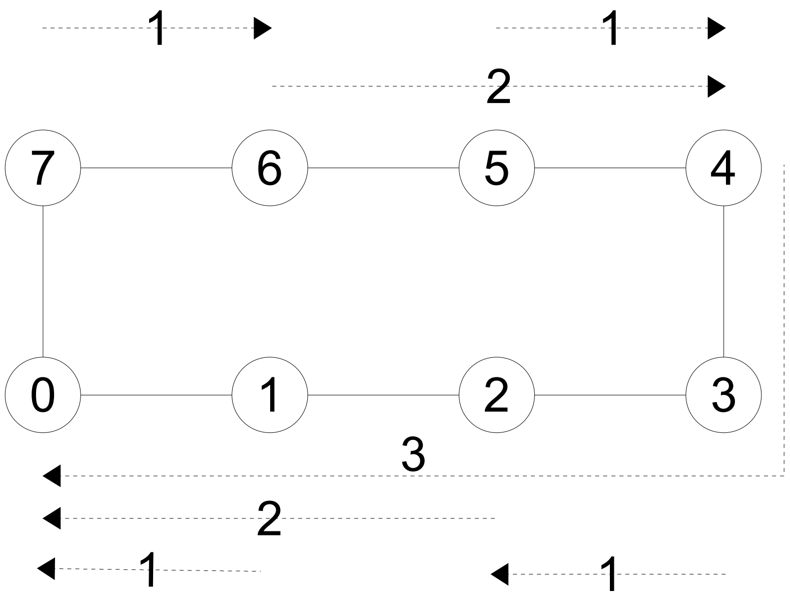
5. Communication Operations

- Communication influences the efficiency of a prallel program essentially
- Examples presented here should give an impression how important good implementations are (what will be done for most of us by the developers of libraries like MPI)
- Assumptions for the following
  - Cut-through routing
  - Bidirectional links
  - Single-port communication model

5. Communication Operations
Linear Array, Ring



One-to-All Broadcast on a ring with eight nodes (MPI_Bcast)



All-to-one Reduction on a ring with eight nodes (MPI_Reduce)

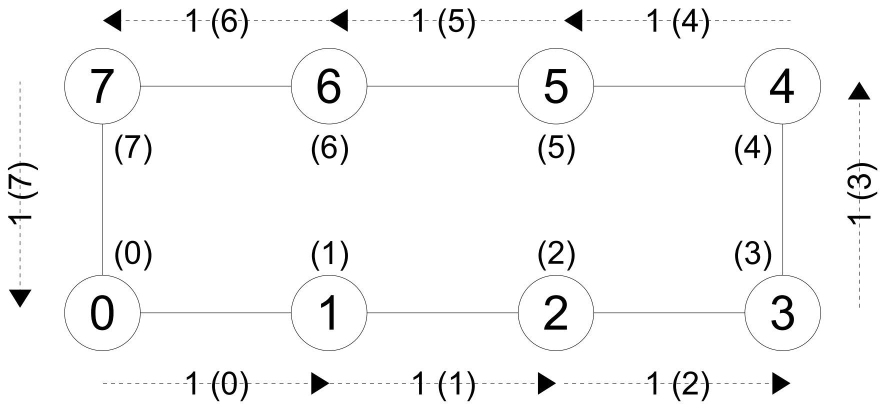
Dual operations

[after Grama et al.]

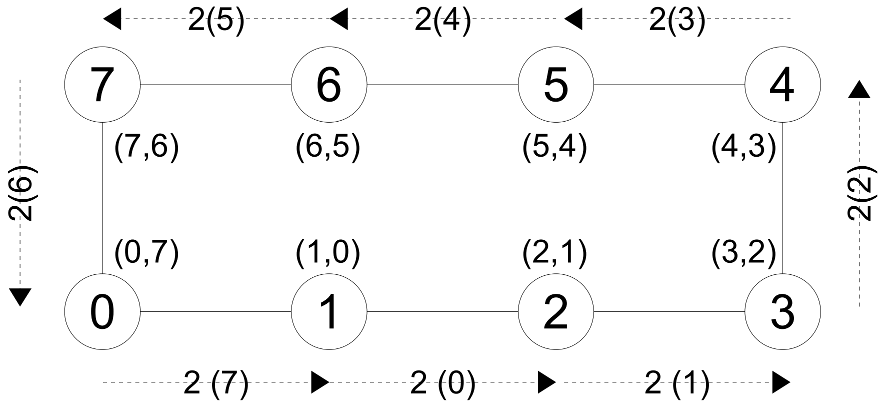
5. Communication Operations

Linear Array, Ring

1st Step

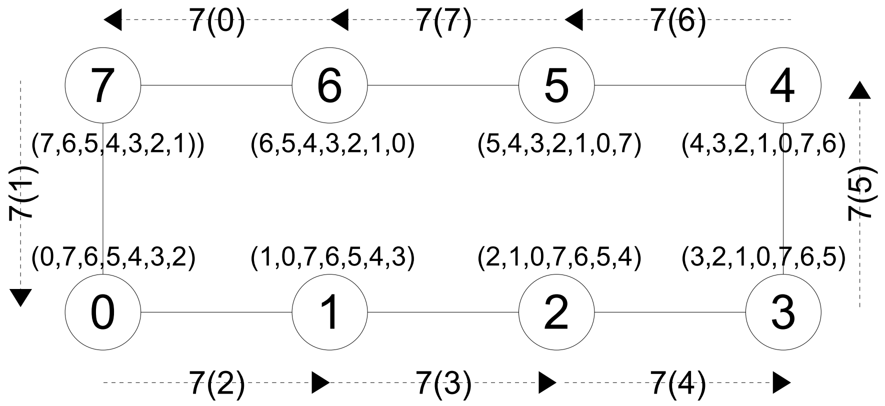


2nd Step



...

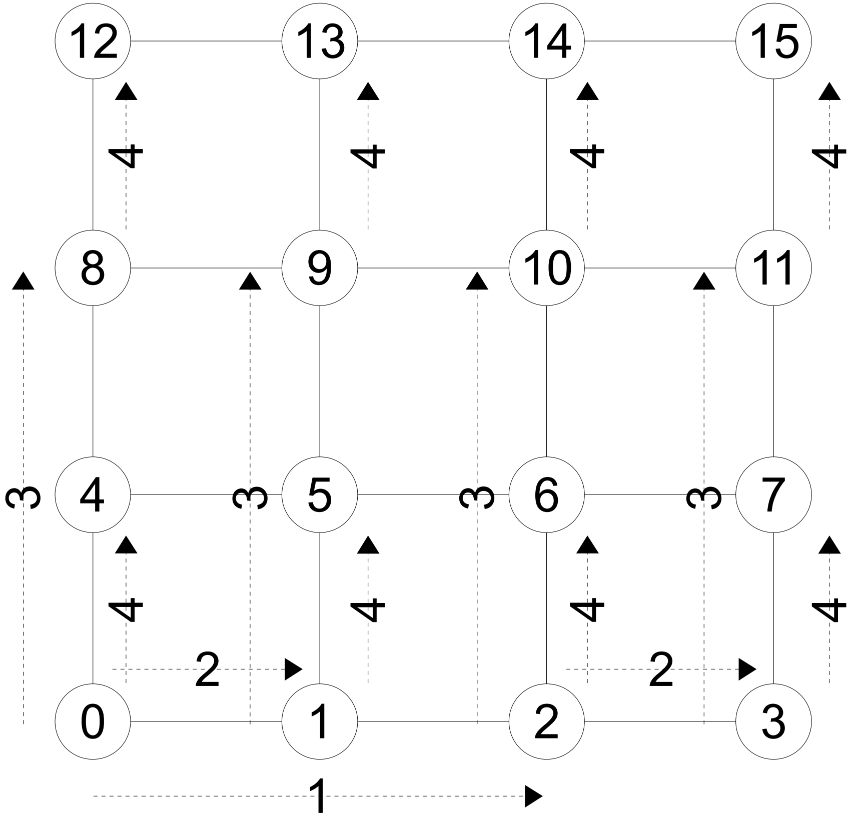
7th Step



All-to-All Broadcast on a ring with eight nodes (MPI_Allgather)

[after Grama et al.]

5. Communication Operations
Linear Array, Ring



One-to-All Broadcast on a mesh (MPI_Bcast)

[after Grama et al.]

Literature

1) (Used for the lecture)
Introduction to parallel computing / Ananth Grama ... [et al.].
Harlow, England ; New York : Addison-Wesley, 2003.
ISBN: 0201648652

2) Parallel programming : for multicore and cluster Systems / Thomas Rauber and Gudula Rünger
Berlin, Heidelberg: Springer, 2010.
ISBN: 978-3-642-04817-3

3) (Predecessor of 2) in German language and used for the lecture)
Parallele Programmierung / Thomas Rauber ; Gudula Rünger
Berlin ; Heidelberg ; New York : Springer, 2007.
ISBN: 978-3-540-46549-2