Computer Architectures

D. Pleiter

Jülich Supercomputing Centre and University of Regensburg

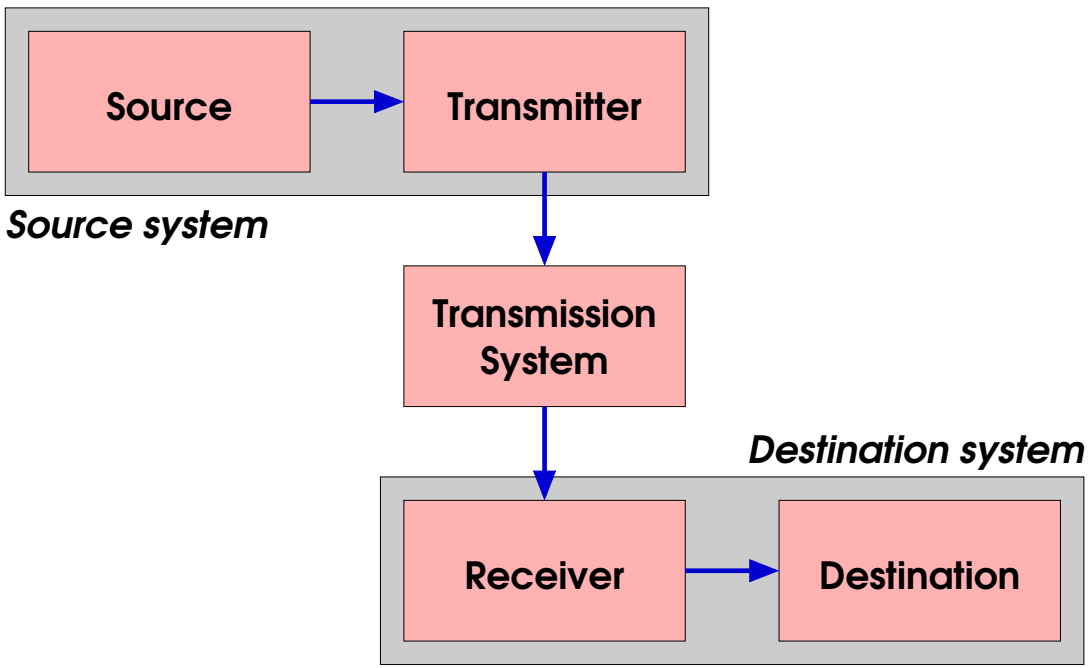
January 2016

Overview

Content

Computer Network

- **Computer network** = interconnection between computing nodes
- Simplistic view:

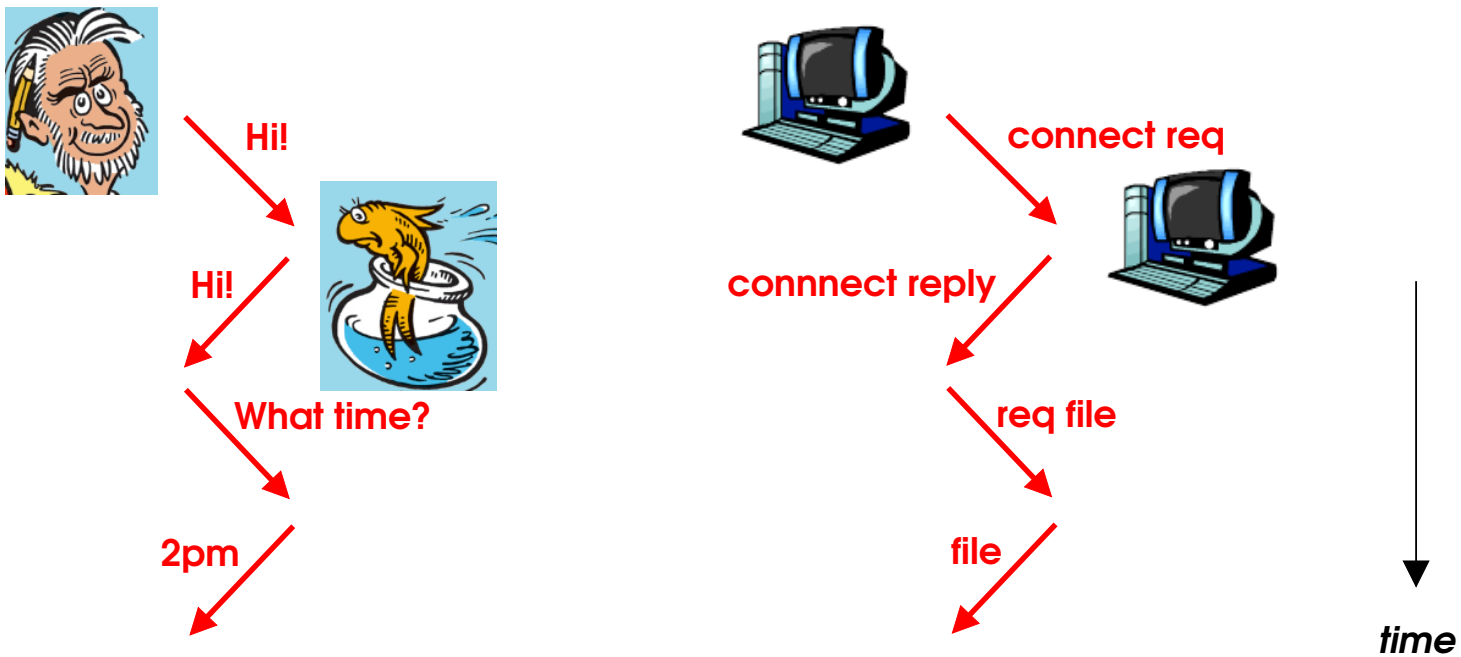


Protocol Architecture

- **Protocol** = Agreement about communication between two or more entities
- Key elements
  - Syntax
    - Data formats
    - Signal levels
  - Semantics
    - Control information
    - Error handling
  - Timing
    - Speed matching
    - Sequencing

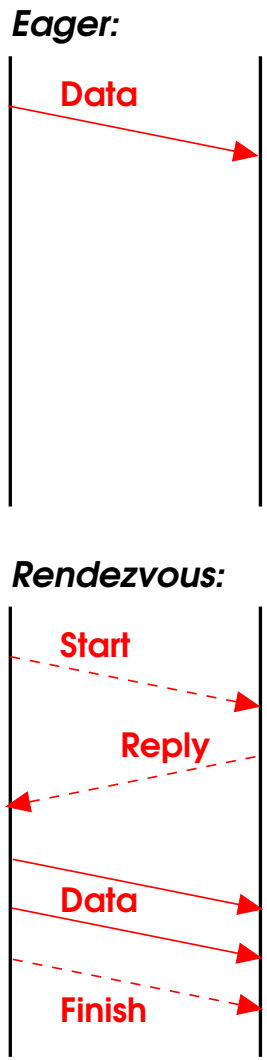
Protocol Architecture

Human protocol vs. computer network protocol



Message Communication Protocols

- **Eager protocol**:
  - Sender pushes message regardless of receiver state
  - Low overhead
  - Useful for small messages
- **Rendezvous protocol**:
  - Handshake happens between the sender and the receiver before sending data
  - High overhead due handshake involving control messages
  - Useful for large messages

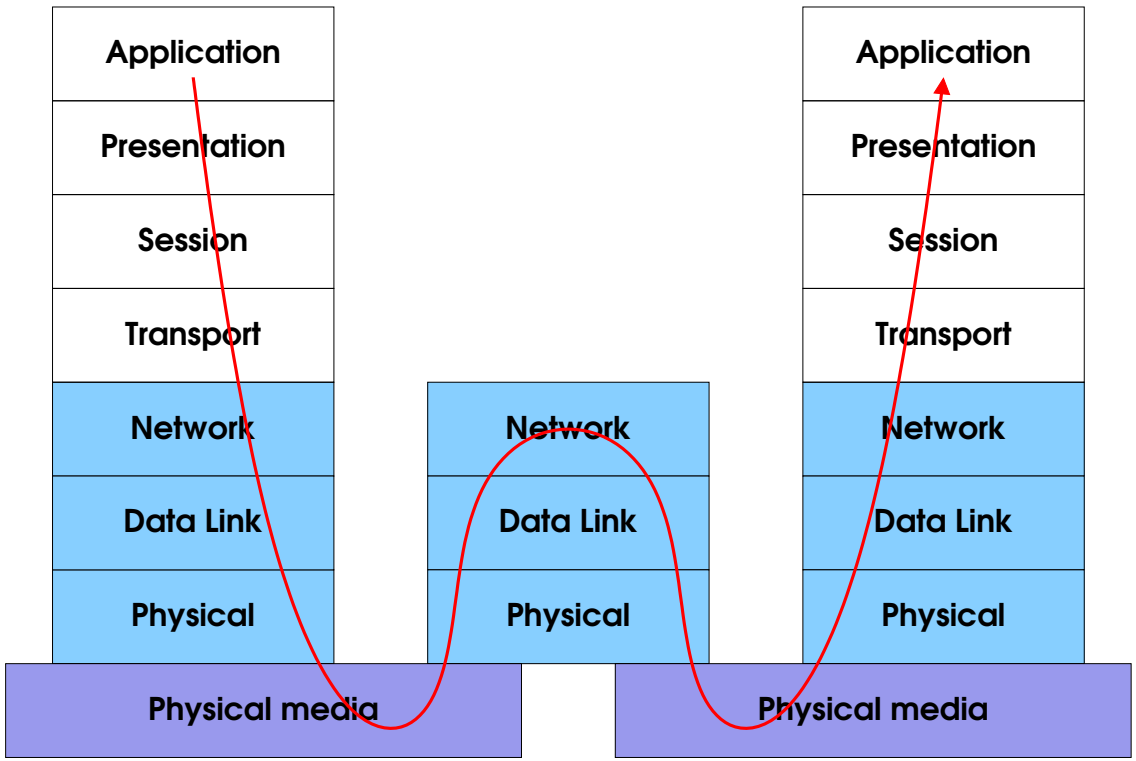


Layering of Networks

- Most networks are organised as a series of layers
  - Physical communication takes place at lowest layer
  - At higher layers: virtual communication
  - Lower levels are implemented in hardware, while top levels are implemented in software
- API between layers are tightly defined
- Advantages of layering:
  - Design becomes less complex as problem is broken down
  - Changes typically affect only one layer
    - Example: new physical layer, new protocols
  - For different layers solutions from different providers can be used

OSI Network Layer Model

- **OSI** = Open Systems Interconnection
  - Standard defined by the International Organization for Standardization (ISO)
  - 2nd edition published in 1994



OSI model: Physical Layer

- **OSI Standard**: “The Physical Layer provides the (...) means to activate, maintain, and de-activate physical-connections for bit transmission between data-link-entities.”
  - Transmission of bits, not, e.g., packets
- Within a given network architecture a large range of physical media may be supported
  - Example: Gigabit Ethernet physical layers (as defined in IEEE 802.3 standard)

Name	Description
1000BASE-T	CAT5 copper cables
1000BASE-SX	Short-wave laser
1000BASE-LX	Long-wave laser

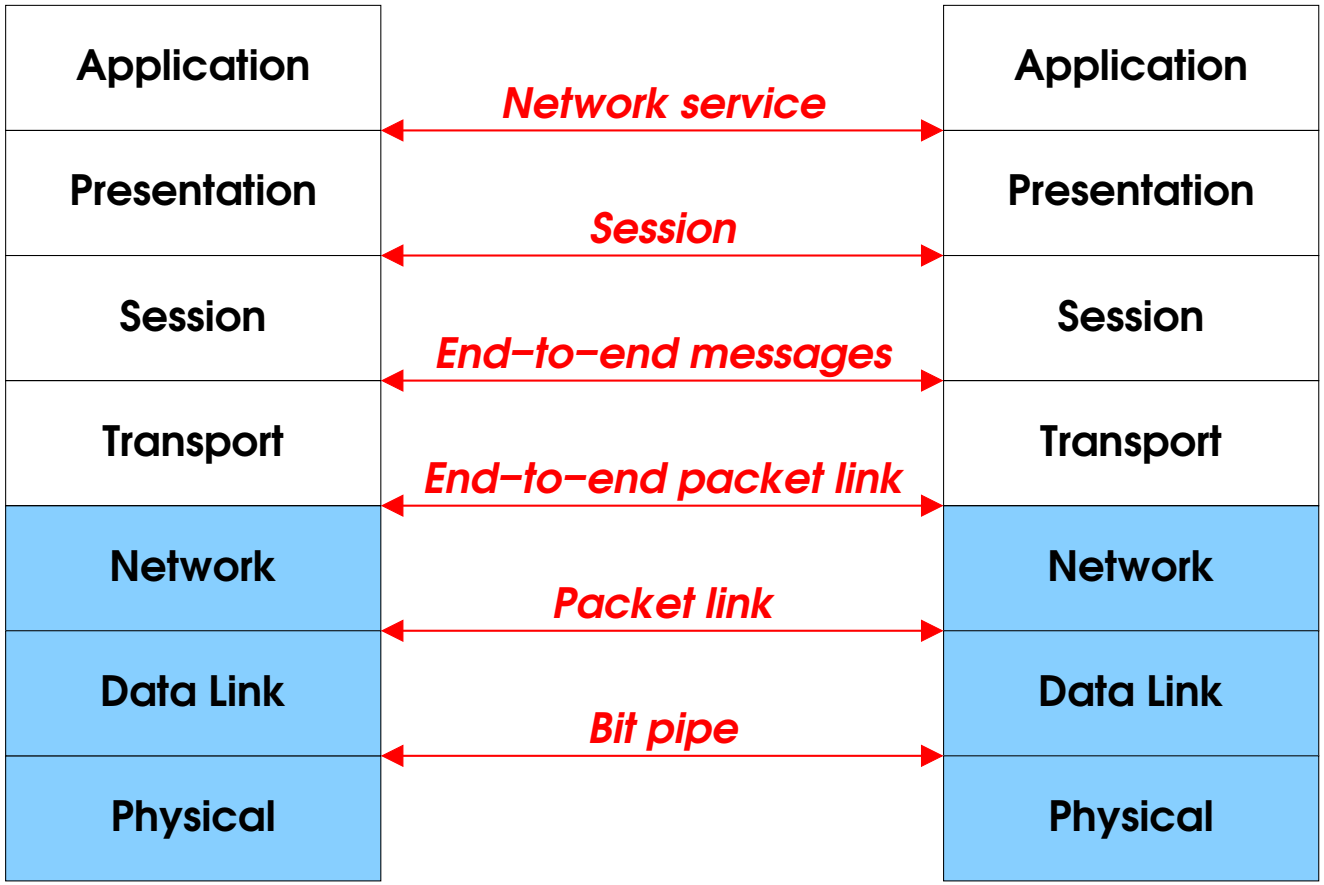
OSI Model: Data Link Layer

- Purpose as defined in the OSI standard:
  - Establish link connections between transport-entities
  - Detects and possibly corrects errors which may occur in the physical layer
  - Connects the network layer to the physical layer
- Layer may implement flow control
  - **Flow control** = Mechanism to control data transmission rate
  - Required to avoid transceiver from outrunning receiver
  - Example: Credit mechanism
    - Receiver provides credits to transmitter depending on available buffer space ☞ give credit
    - Transmitter only sends data when credits available ☞ consume credit

OSI Model: Network Layer

- **OSI Standard**: “It provides to the transport-entities independence from routing and relay considerations associated with the establishment and operation of a given network-connection.”
- The standard defines that transport-entities are known to this layer by mean of network-addresses
- Routing can be implemented at this layer
  - Layer can receive packets via the interface to the data link layer of one link and forward the packets to another link depending on the network-address

OSI Model: Virtual View



Digital Signal Encoding

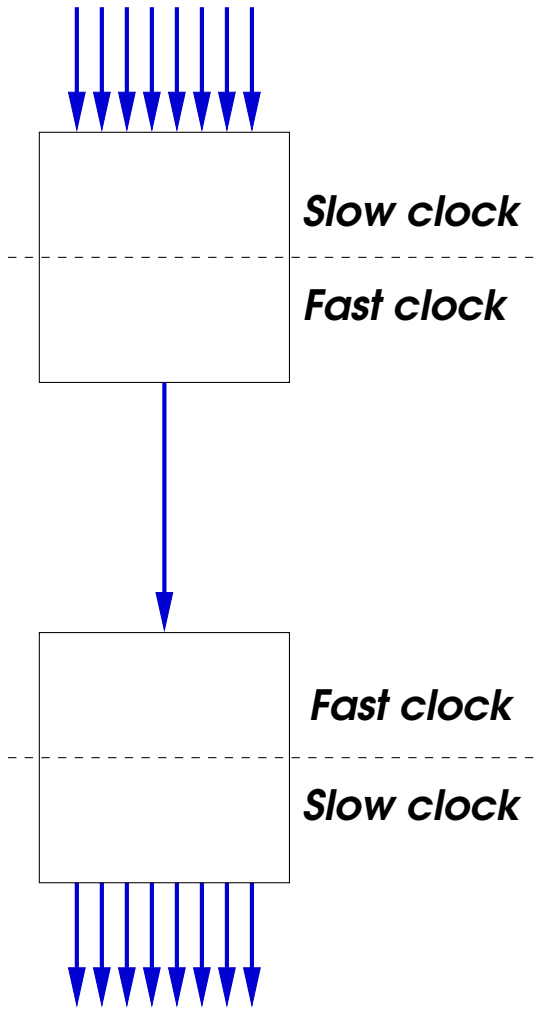
- **Digital signal** = discrete, discontinuous voltage pulses
  - In practice, signal may be only approximatively digital
- Binary data encoded into signal elements
- Signal encoding schemes (selection)
  - Return to Zero (RZ), Non-Return to Zero (NRZ)
  - Non-Return to Zero Inverted (NRZI)
  - Manchester
- Pre-requisites for receiving data
  - Matching of clock
  - Voltage level matching

Digital Signal Encoding: Rates

- **Symbol rate** = Number of times a signal in a communication channel changes state or varies per time unit
  - Unit: baud
- **Bit rate** = Number of bits transmitted in a given time unit
- Symbol rate $\neq$ bit rate
  - One change of state in the physical link may correspond to one bit or more or less than one bit
- Bit rate [bits/s] = baud $\times$ bits per baud

Physical Layer: High-speed Transceivers

- Standard technology: Serialisation/Deserialisation (**SerDes**)
  - Transmitter:
    - Wide parallel input
    - High-speed serial output
  - Receiver: vice versa
- Example:
  - Parallel interface: 16 bit at 250 MHz
  - Serial interface: 1 bit at 5 GHz
  - Higher output data rate due to encoding (here: 8b/10b)
- Clock is typically not transmitted
  ☞ need **clock recovery**



Link Errors

- Errors in network will occur!
  - Error detection
  - Error recovery
- Error types
  - Temporary errors (transient or intermittent)
    - Data corruption, e.g. bit flips in data characters due to random noise
    - Data loss, e.g. packet loss due to control character corruption
    - Shift in sampling point
  - Permanent errors
    - Link failure
- It may not be possible to guarantee detection of all errors
  - Strategy: Make probablity of undetected errors sufficiently small

Link Errors (2)

- **BER** = Bit Error Rate

$$\frac{\text{Number of flipped bits}}{\text{Number of received bits}}$$

- **PER** = Packet Error Rate
  - Assume packets of equal length of N bits and assume bit errors to be independent (assuming BER is small):

$$\text{PER} = 1 - (1 - \text{BER})^N \simeq N \cdot \text{BER}$$

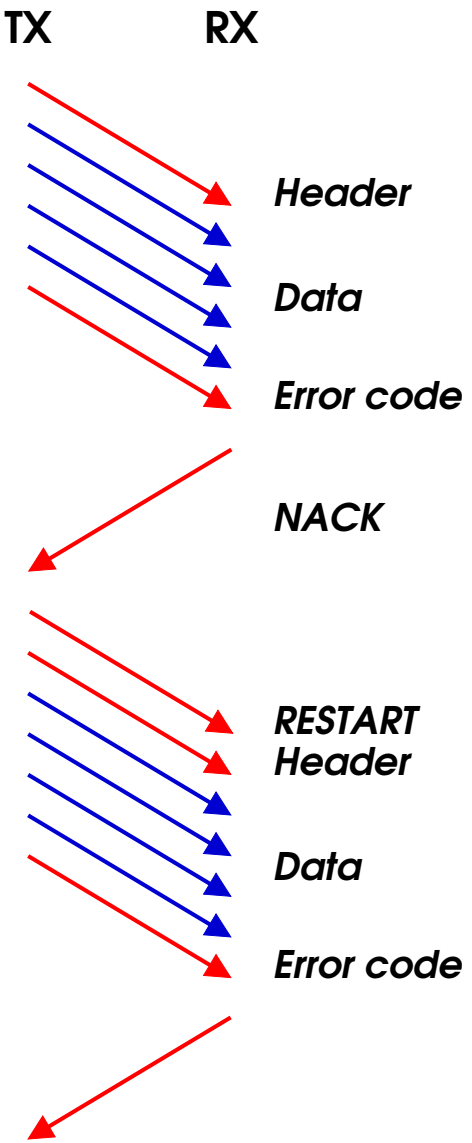
  - Warning: - In practice, bit errors are not independent
    - Encoding results in non-deterministic relation between BER and PER
- BER targets
  - 10 Gbit Ethernet: 10^{-12} (IEEE 802.3-2008, clause 44)
  - Infiniband: 10^{-12} (IBA, rev. 1.2.1, C6-11, C8-4)

Network Error Detection

- **Detection of data corruption**
  - Transmitter adds error-detecting code
  - Code is recalculated and checked by receiver
  - Simple example: parity bit
    - Set parity bit if number of ’1’ is odd (odd parity) or even (even parity)
    - Even number of bit flips remains undetected
  - More robust error codes: Cyclic Redundancy Check
- **Detection of data loss**
  - Robust protocols where each packet is acknowledged

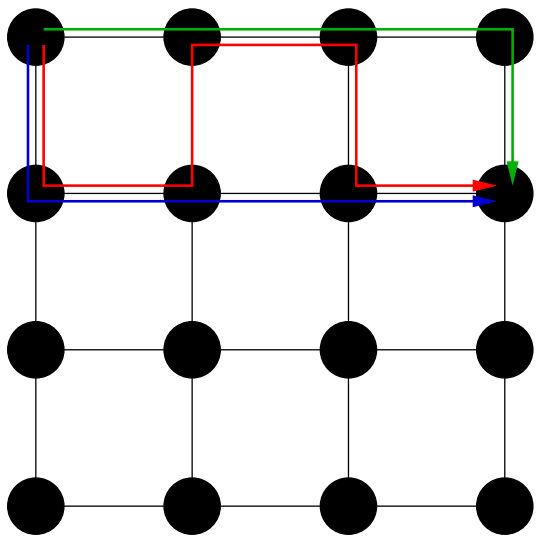
Error Recovery: Packet Retransmission

- Error case: Packet payload is corrupted
- Transmitter adds error-code to packet and keeps copy of packet
- Receiver checks error-code and returns feedback to transmitter
  - `ACK` if code matched
  - `NACK` if mismatch detected
- Transmitter action depending on feedback:
  - `ACK`: Delete copy
  - `NACK`: Retransmit packet



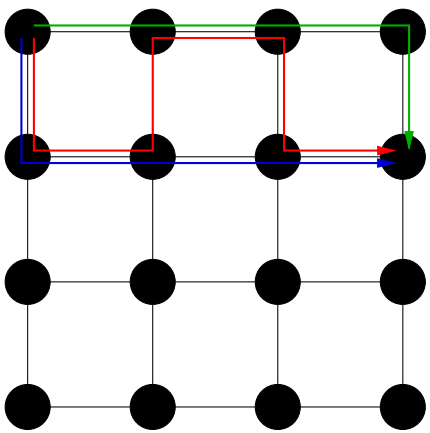
Message Routing Introduction

- **Routing** = Mechanism to allow messages to find a path from source to destination
  - May be implemented in hardware or software
- Features relevant for HPC network:
  - Performance
    - High bandwidth
    - Low latency
  - Scalability
    - Algorithm must be implemented in a distributed way



Network Routing Ingredients

- Network topology
  - Restricts number of routes from source to destination
- Switching techniques
  - **Switching** = Mechanism that a router uses to move a packet from its input to output ports
- **Routing algorithm** = The algorithm used to determine the path that a message will take to go from the source to destination



Classification of Routing Algorithms

- Static vs. dynamic routing algorithms
  - **Deterministic/oblivious routing**
    - Routes based only on information that is available before the algorithm begins
  - **Adaptive routing**
    - Decisions are made at run-time
  - **Quasi-static routing**
    - Similar to deterministic routing, but allow for routing tables to be updated periodically
- Path length
  - **Minimal routing**
    - Route packets along minimal path, i.e. do not allow packets to move away from destination
  - **Non-minimal routing**
    - Allow packets to move away from destination

Classification of Routing Algorithms (2)

- Other performance relevant features
  - Congestion management
    - **Congestion** = Situation where a link or node is carrying so much data that its quality of service deteriorates
- Worst-case characteristics
  - **Deadlock** = State where packets cannot progress in network due to cyclic wait dependency
  - **Livelock** = State where packets injected into the network are constantly routed and never reach destination

Network Topology

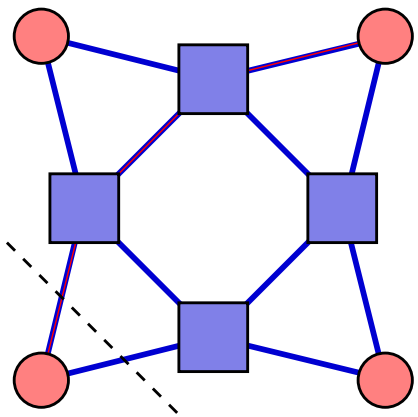
- **Network topology** refers to the way in which the network of computing nodes is connected
- Physical topology = topology of the physical network
- Logical topology = logical view on the way network is connected
  - Not explored in this lecture
- Simplest topology: point-to-point topology
  - Permanent link between two **endpoints**

Network Evaluation Metric

- Diameter
  - Maximum minimal distance between 2 nodes
  - Smaller is better
- Connectivity
  - The minimum number of arcs that must be removed to break it into two disconnected networks
  - Larger is better
- Bi-section width
  - The minimum number of arcs that must be removed to partition the network into two equal halves
  - Larger is better
- Bi-section bandwidth
  - Aggregate bandwidth of all links connecting two equal halves
  - Larger is better
- Costs
  - Smaller is better

Network Evaluation Metric: Diameter and Connectivity

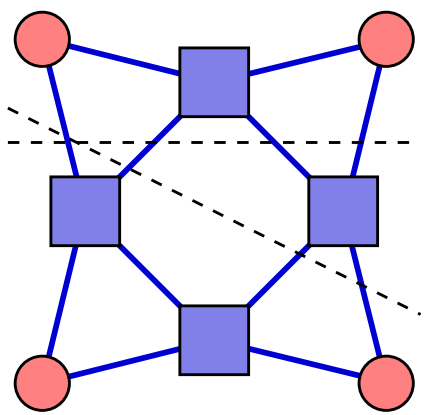
- Diameter is the maximum minimal distance between 2 nodes
- Connectivity gives the minimum number of edges to be removed for breaking network in disconnected pieces
- Example:



  - Diameter = 3
  - Connectivity = 2

Network Evaluation Metric: Bi-section (Band-)Width

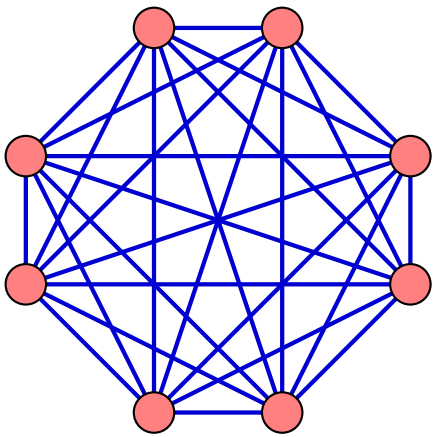
- Procedure to determine bi-section width and bandwidth:
  1. Determine all possible ways to create 2 equi-partitions
  2. Select the partitioning with minimal number of edges removed
  3. Compute aggregate bandwidth of removed links
- Example:



  - Bi-section width = 4

All-connected and Star Topology

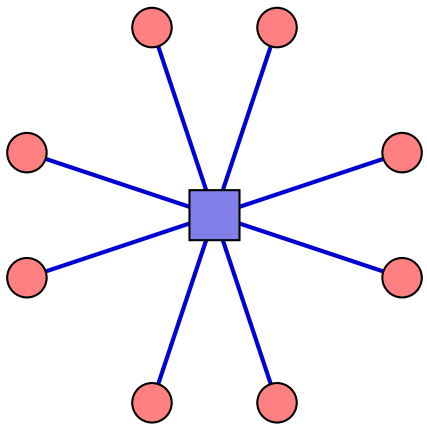
All-connected network



Evaluation:

- Diameter: 1
- Connectivity: $P - 1$
- Bi-section width: $(P/2)^2$
- Costs: $O(P^2)$

Star topology

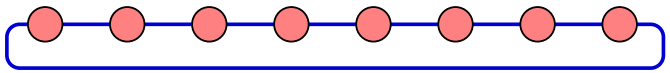


Evaluation:

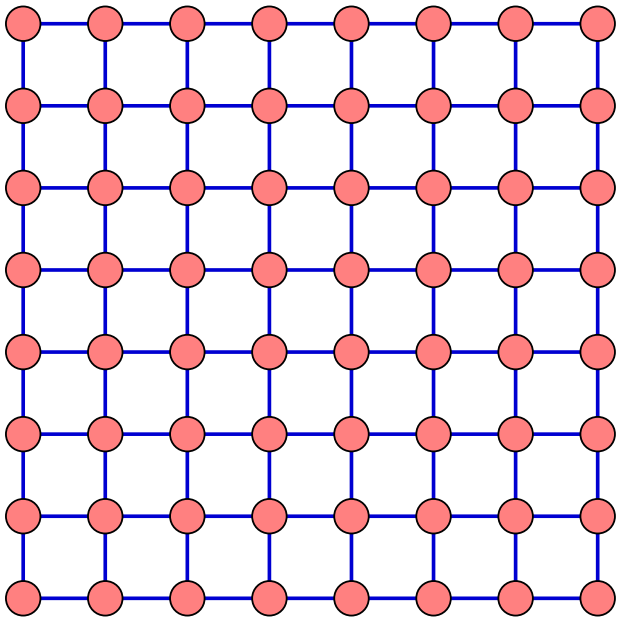
- Diameter: 2
- Connectivity: 1
- Bi-section width: —
- Costs: $O(P)$

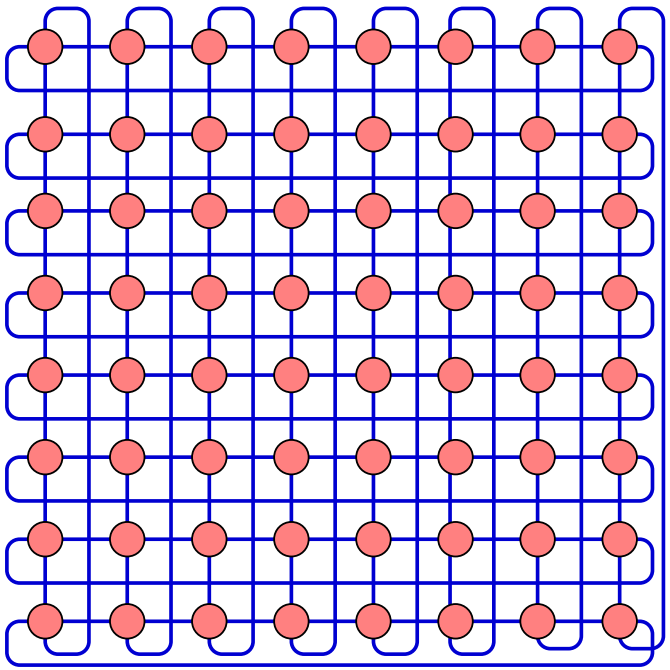
Cartesian Network Topologies

- 1-dimensional linear array and ring:



- 2-dimensional mesh and torus:





Cartesian Network Topologies (2)

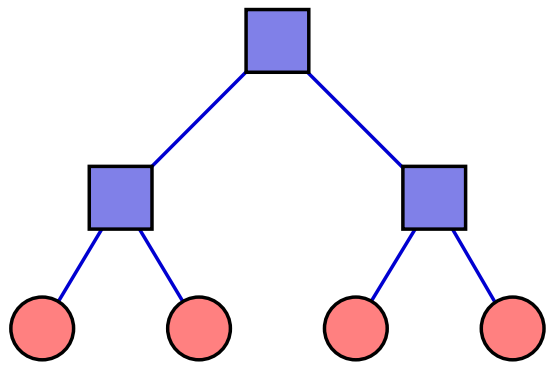
- Evaluation:

	Mesh	Torus
Diameter	$d(P^{1/d} - 1)$	$d(P^{1/d}/2)$
Connectivity	d	$2d$
Bi-section width	$P^{(d-1)/d}$	$2P^{(d-1)/d}$

Network Topology: Tree

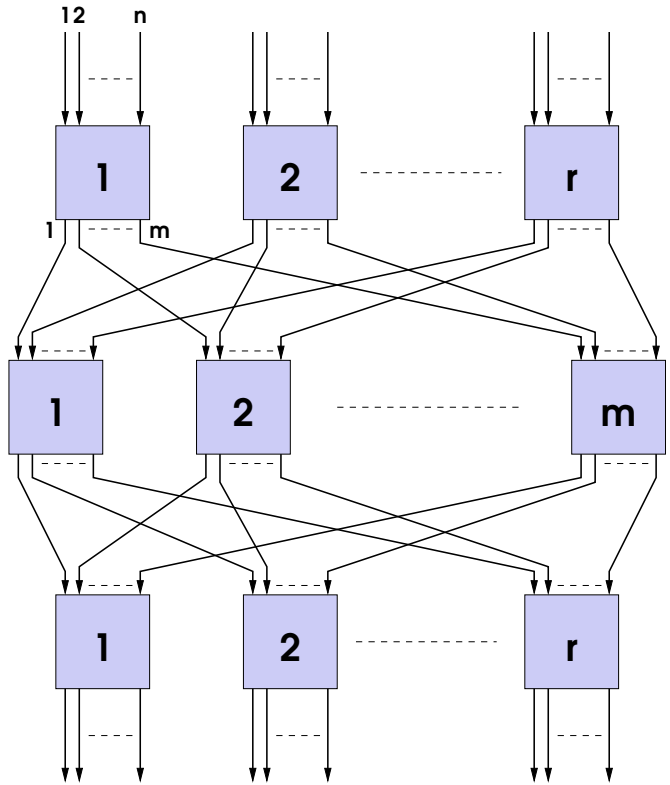
- Switch-based network
- Evaluation binary tree:
  - Diameter: $O(\log(P))$
  - Connectivity: 1
  - Bi-section width: 1
  - Costs: $O(P)$
- Number of links between upper levels smaller

  ☞ **Blocking network topology**



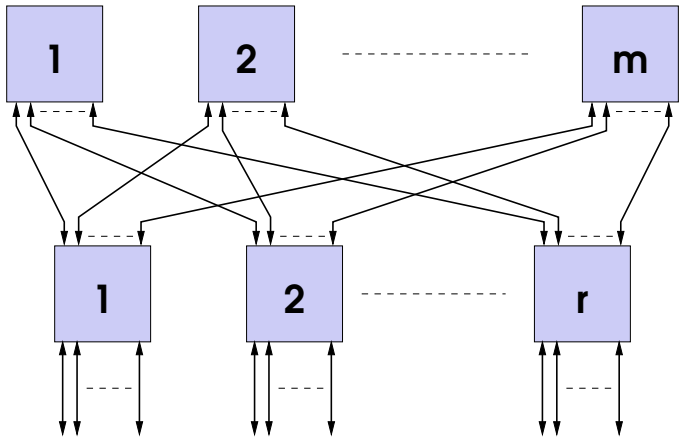
Network Topology: Clos Network

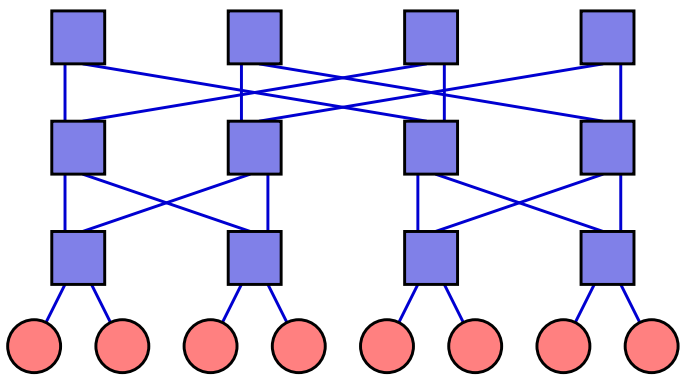
- N-stage switch-based network
  - Here: $N = 3$
- Network is characterised by
  - r: number of ingress stage switches
  - n: number of ingress ports at ingress stage
  - m: number of egress ports at ingress stage
- Evaluation:
  - Diameter: $N + 1$
  - Connectivity: m (at switch-level)
  - Bi-section width: $r\frac{m}{2}$
- A Clos network can be made **non-blocking**
  - Non-blocking for $m \geq n$ if one allows for rearrangements
  - Strictly non-blocking if $m \geq 2n - 1$



Network Topology: Folded Clos Network

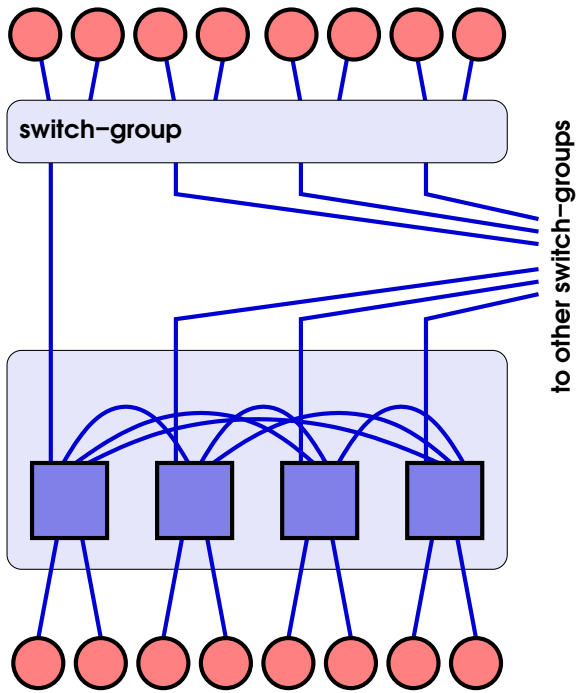
- Topology obtained by folding ingress and egress stage
  - Example: folded 3-stage Clos
- Evaluation of Clos and Folded Clos network:
  - Diameter = $N + 1$
  - Connectivity = m
  - Bi-section width = $r\frac{m}{2}$
- Closely related to **fat-tree** topology
  - Example: folded 5-stage Clos





Network Topology: Dragonfly

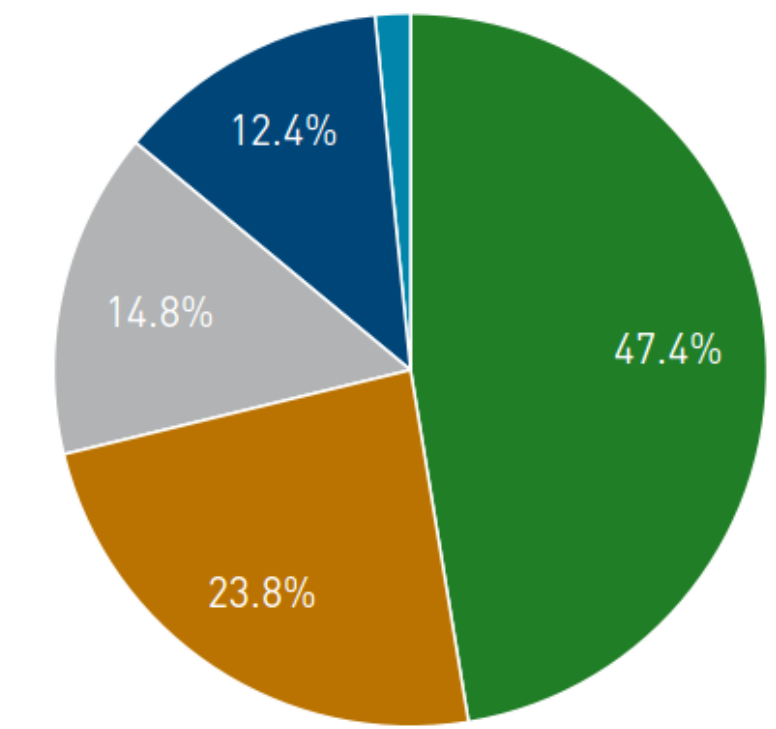
- Multi-layer network
  - Switch layer
  - Switch-group layer
  - System layer
- All-to-all connectivity within layer
- Network parameters
  - p: number of nodes per switch
  - a: number of switches per group
  - h: number of links per switch to other groups
- Evaluation:
  - Diameter: 5
  - Connectivity: $ah - 1$ (at system layer)
  - Bi-section width: $\left(\frac{a\,h+1}{2}\right)^2$



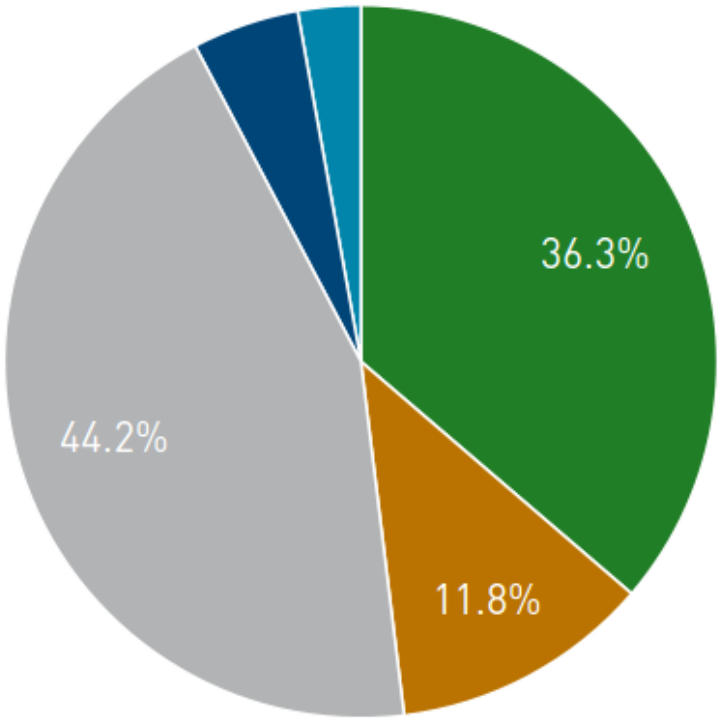
Network Solutions for Top500 Systems

[www.top500.org, 11/2015]

System share



Performance share



- Infiniband
- 10G
- Custom Interconnect
- Gigabit Ethernet
- Proprietary Network

Network Solutions: Ethernet

- Most popular data centre network technology
- Large class of technologies including several technology families like
  - 1 Gigabit Ethernet
  - 10 Gigabit Ethernet
  - 40 and 100 Gigabit Ethernet
- Limited uptake for compute node interconnect
  - Relatively low bandwidth, high latencies
  - High costs of large high-bandwidth switch fabrics
- Growing interest in high-speed Ethernet $\rightarrow$ new protocols
  - Internet Wide-Area RDMA Protocol (iWARP)
  - RDMA over Converged (Enhanced) Ethernet (RoCE)

Network Solutions: Infiniband

- Currently most popular network technology
- Fast evolution of (nominal) bandwidth

Technology	QDR	FDR	EDR
First Top500	06/2009	11/2011	06/2015
Data rate [Gbits/s]	8	13.6	24.2

- Low latency of $\mathcal{O}(1\ \mu\text{s})$
- Key features
  - OS kernel by-pass
  - Function offloading
    - Data transport managed by HCA
    - Offload of collective operations

Top10 Network Architectures

[www.top500.org, 11/2015]

	System	Technology	Topology
1	Tianhe-2	Custom	fat tree
2	Titan	Cray Gemini	3-d torus
3	Sequoia	Custom	5-d torus
4	K-Computer	Tofu	3-d toroidal
5	Mira	Custom	5-d torus
6	Trinity	Cray Aries	dragonfly
7	Piz Daint	Cray Aries	dragonfly
8	Hazel Hen	Cray Aries	dragonfly
9	Shaheen II	Cray Aries	dragonfly
10	Stampede	Infiniband FDR	fat tree