The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2494-6

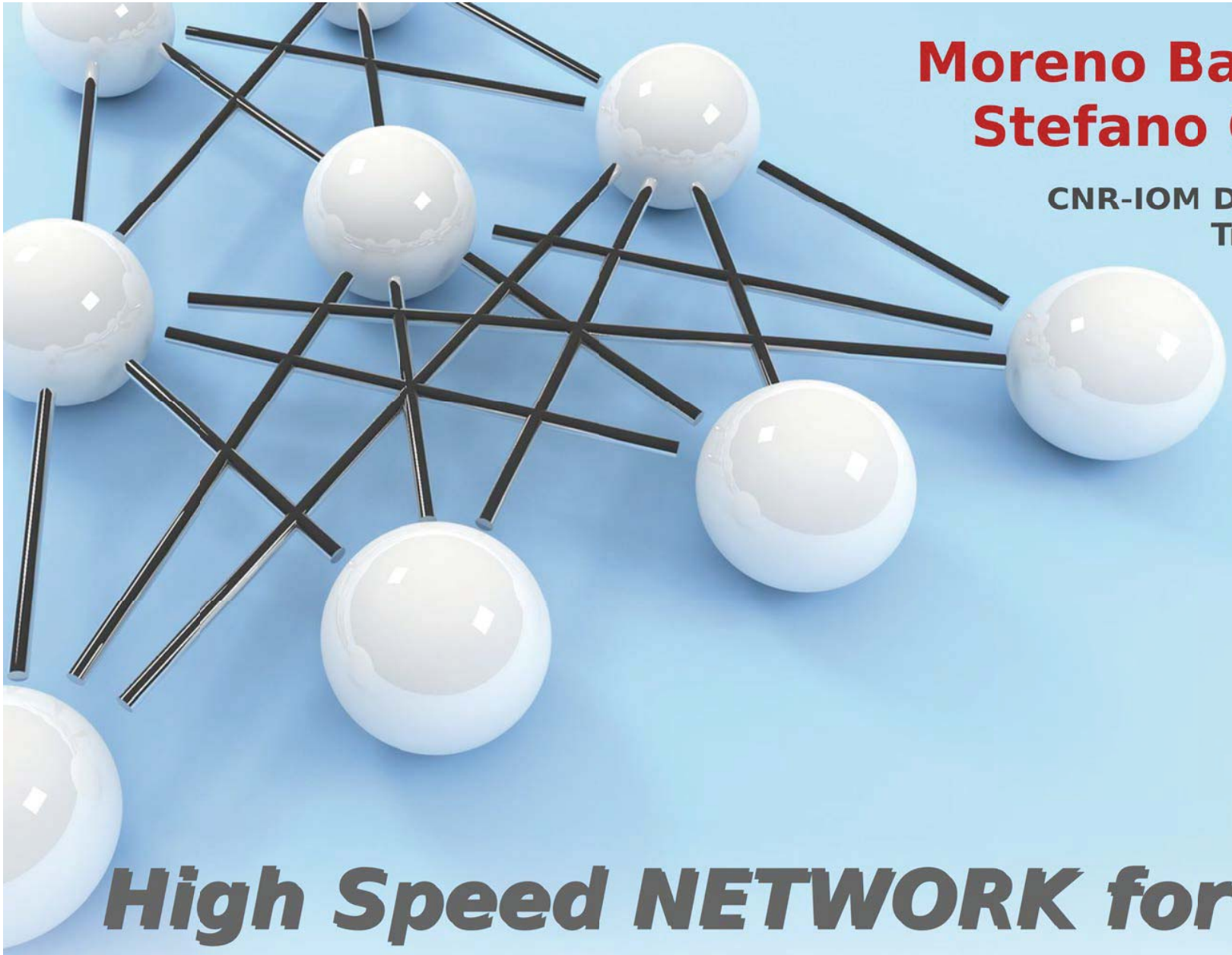
Workshop on High Performance Computing (HPC) Architecture and Applications in the ICTP

14 - 25 October 2013

High Speed Network for HPC

Moreno Baricevic & Stefano Cozzini
CNR-IOM DEMOCRITOS
Trieste

Strada Costiera, 11 - 34151 - Trieste - Italy • Tel. +39 0402240111 • Fax. +39 040224163 • sci_info@ictp.it • www.ictp.it
ICTP is governed by UNESCO, IAEA, and Italy, and it is a UNESCO Category 1 Institute



Moreno Baricevic
Stefano Cozzini

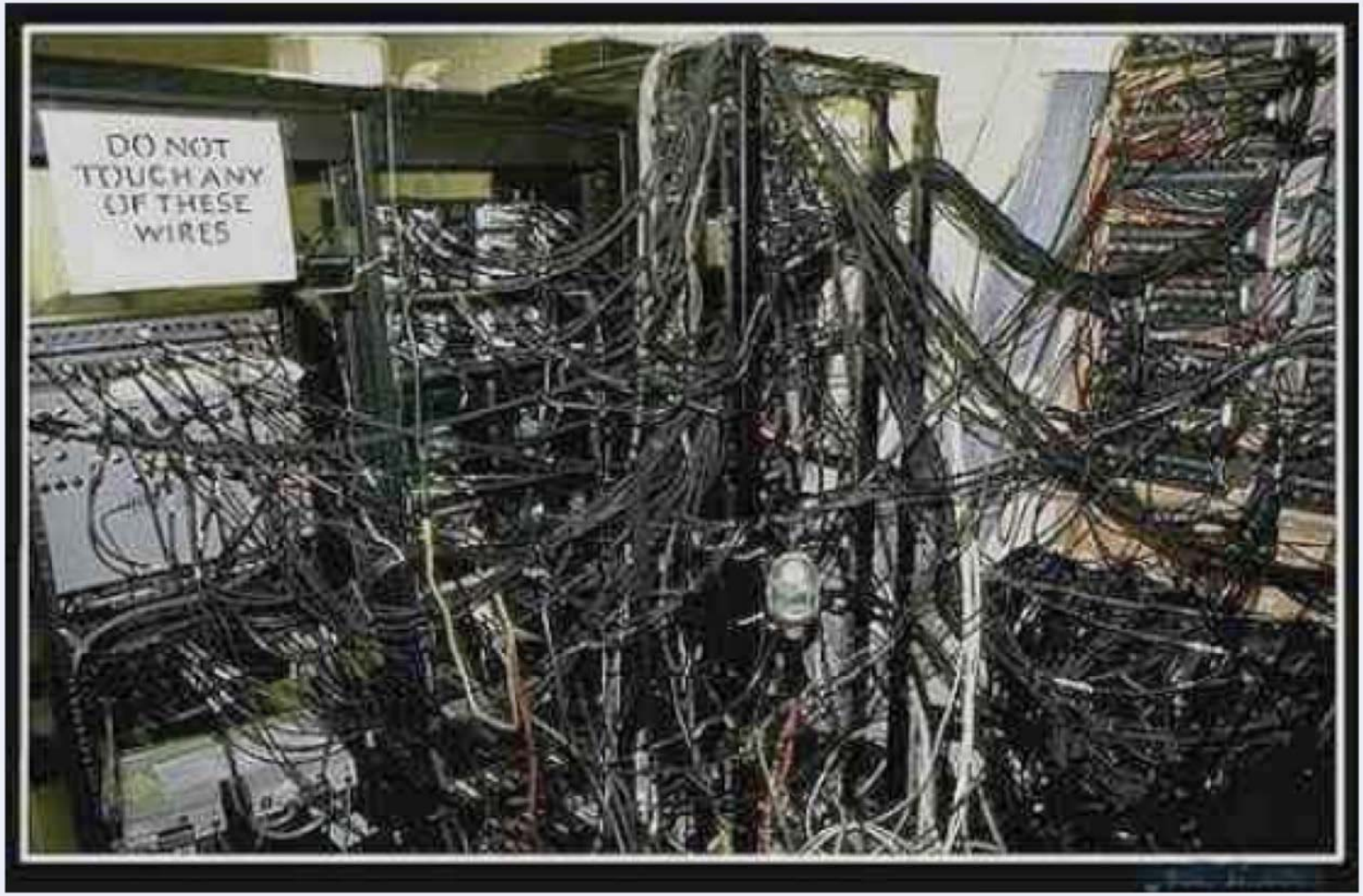
CNR-IOM DEMOCRITOS
Trieste, ITALY

High Speed NETWORK for HPC

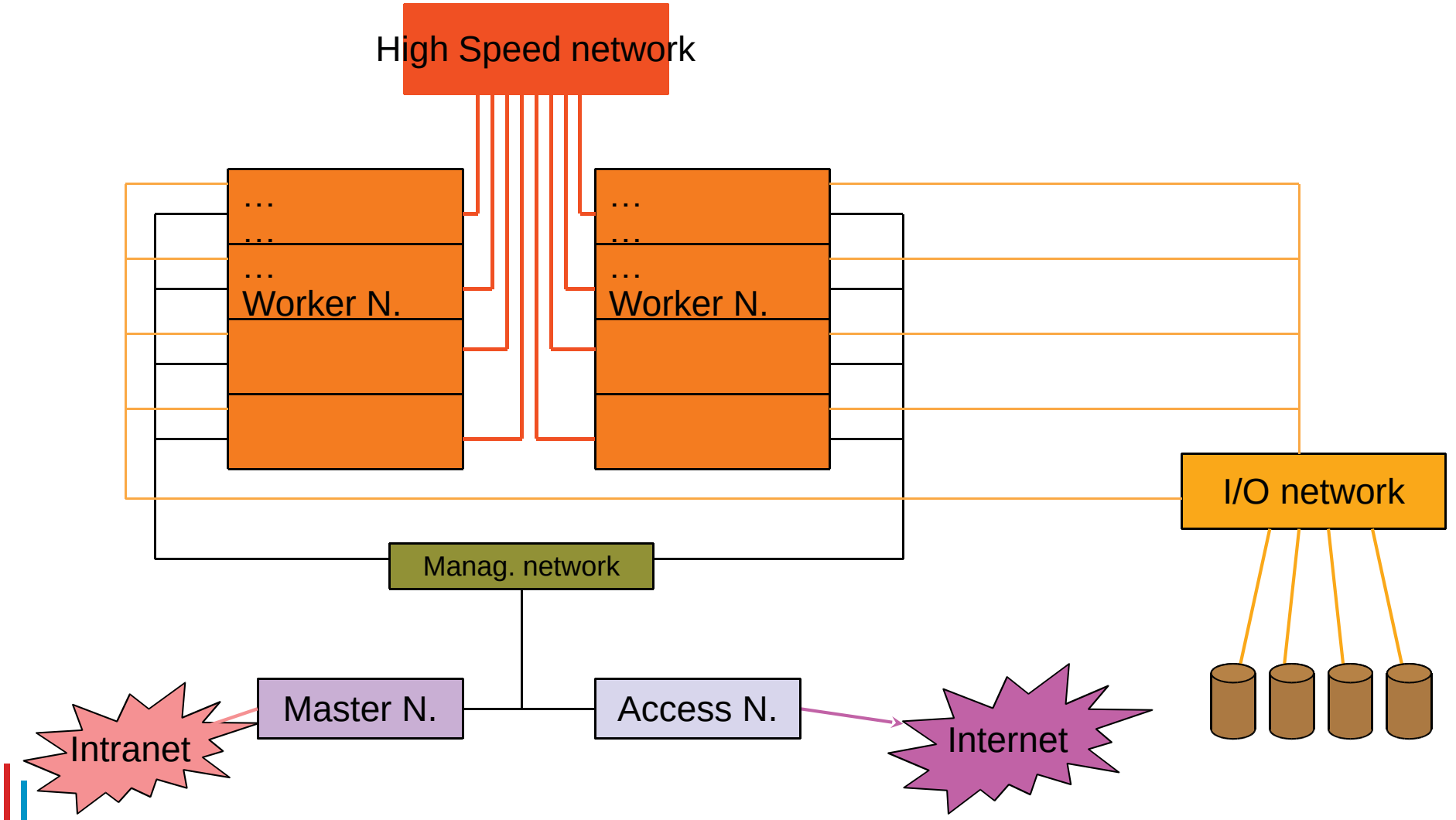
Agenda

- About network for clusters:
  - Basic Concepts of Networking
  - Topology Bandwidth, Latency, Throughput
- High Speed (and Low Latency) Networks
- Infiniband
- Which network for your cluster ?

About the network for clusters..



HPC cluster logical structure



Network Clusters classification

- HIGH SPEED NETWORK
  - parallel computation
    - low latency /high bandwidth
    - Usual choices: Infiniband...
- I/O NETWORK
  - I/O requests (NFS and/or parallel FS)
    - latency not fundamental/ good bandwidth
    - GIGABIT could be ok
- Management network
  - management traffic
    - any standard network (fast ethernet OK)

Characteristics of a network

- Topology
  - Diameter
  - Nodal Degree
  - Bisection bandwidth
- Performance
  - Latency
  - Link bandwidth

Topology

- How the components are connected.
- Important properties
  - Diameter: maximum distance between any two nodes in the network (hop count, or # of links).
  - Nodal degree: how many links connect to each node.
  - Bisection bandwidth: The smallest bandwidth between half of the nodes to another half of the nodes.
- A good topology: small diameter, small nodal degree, large bisection bandwidth

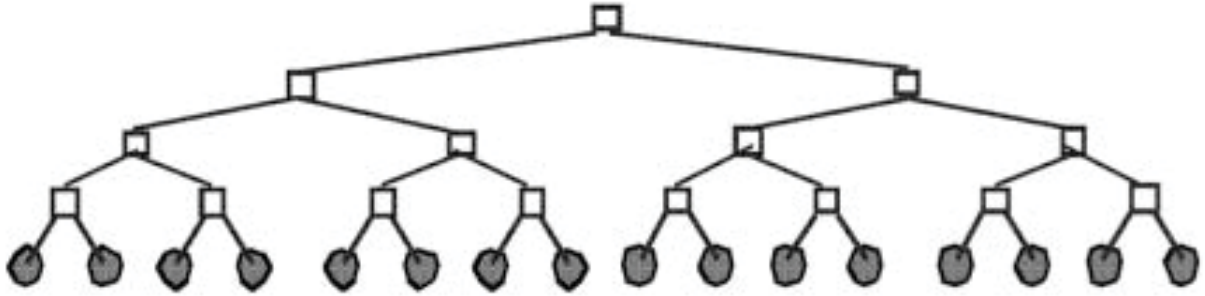
Bisection bandwidth

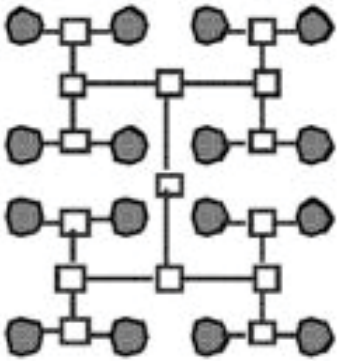
- Split N nodes into two groups of N/2 nodes such that the bandwidth between these two groups is minimum: that is the bisection bandwidth

Why is Bisection Bandwith relevant ?

- if traffic is completely random, the probability of a message going across the two halves is ½
- if all nodes send a message, the bisection bandwidth will have to be N/2
- The concept of bisection bandwidth confirms that some network topology network is not suited for random traffic patterns
- your worst case scenario of HPC workload is to have random traffic patterns..

Tree Topology



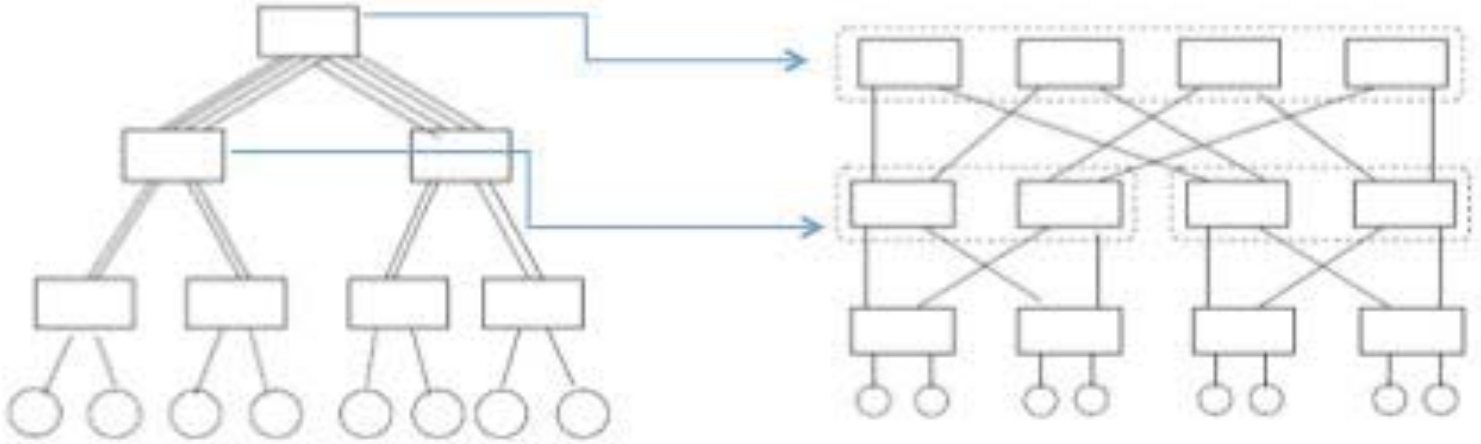


- Fixed degree, log(N) diameter, O(1) bisection bandwidth.
- Routing: up to the common ancestor than go down.

Fat tree topology

- Fatter links (really more of them) as you go up so bisection BW scales with N
- Not practical. Root is NXN switch

Practical fat tree topology



- Use smaller switches to approximate large switches.
- Most commodity large clusters use this topology.
- Also call constant bisection bandwidth network (CBB)

Latency in Networking

Latency is the delay between the time a frame begins to leave the source device and when the first part of the frame reaches its destination. A variety of conditions can cause delays:

- Media delays may be caused by the finite speed that signals can travel through the physical media.
- Circuit delays may be caused by the electronics that process the signal along the path.
- Software delays may be caused by the decisions that software must make to implement switching and protocols.

Latency in HPC

The **one-way latency** may be also meant as **the period of time that a 0-sized message spends traveling from its source to its destination**,

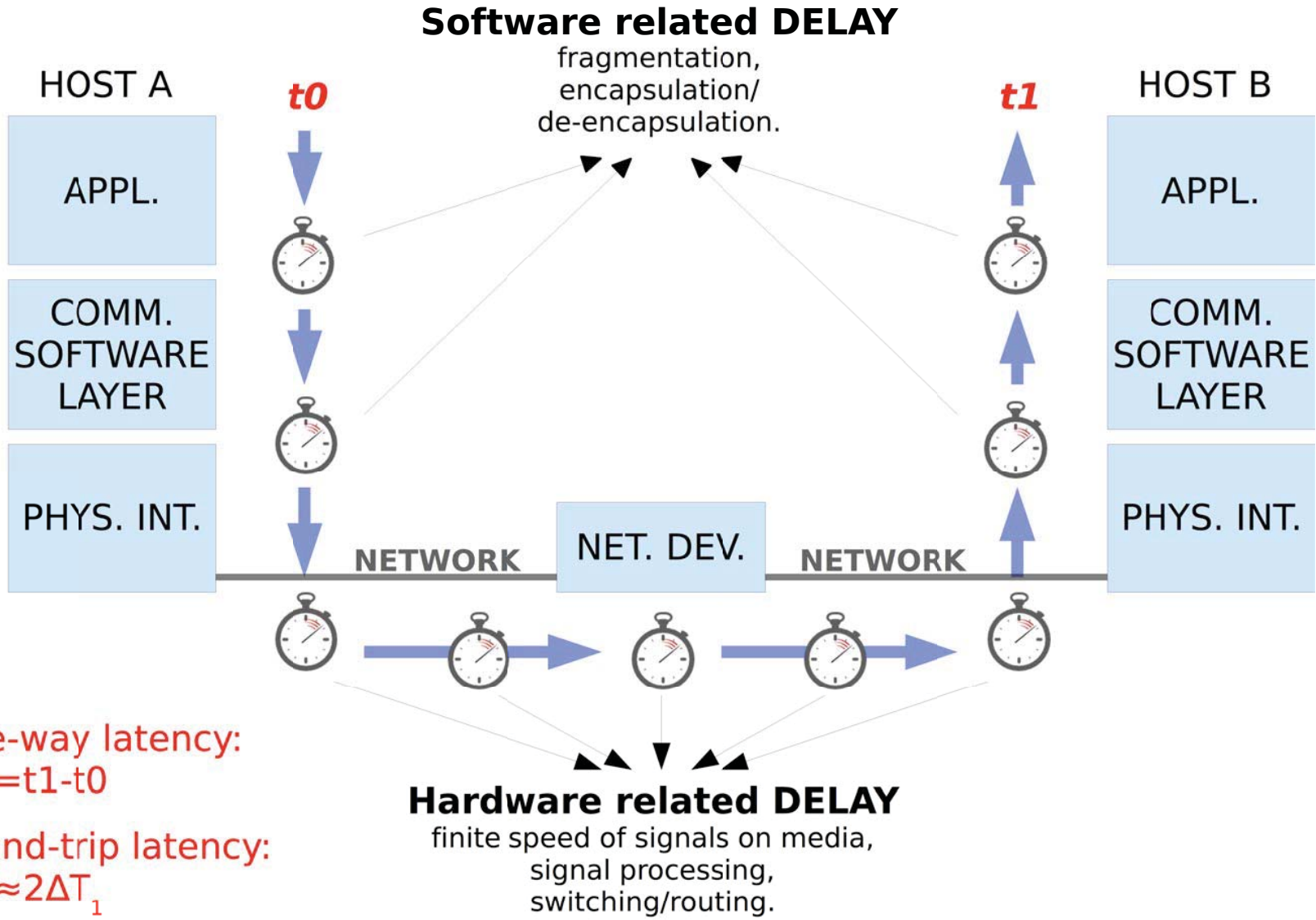
It involves the time needed to:

- encode,
- send the packet,
- receive the packet,
- decode

The **round-trip latency** includes also the travel back to the source of an **acknowledge message**.

- **How much does it take to open the channel ?**

Latency



One-way latency:
$\Delta T_1 = t1 - t0$

Round-trip latency:
$\Delta T_2 \approx 2\Delta T_1$

Bandwidth and Speed

Bandwidth
the measure of the **amount of information that can move through the network in a given period of time**.

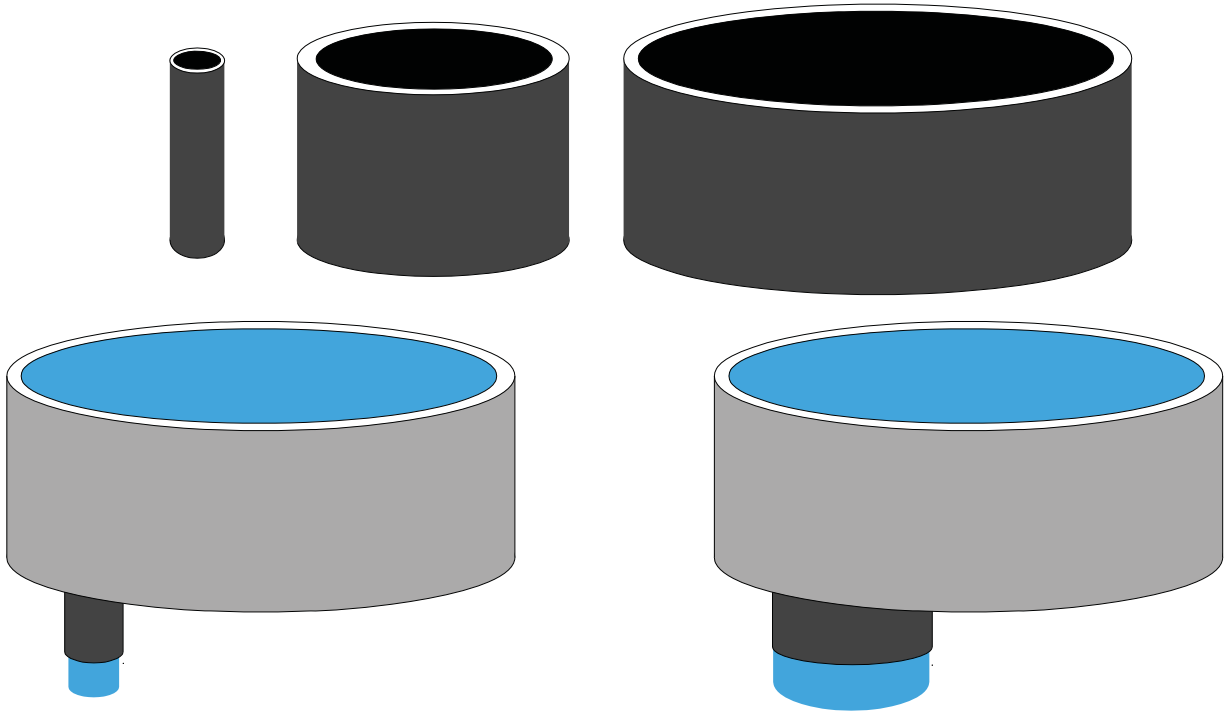
How wide is my channel ?

Warning:

Speed is often used interchangeably with bandwidth, but a large-bandwidth device will carry data at roughly the same speed of a small-bandwidth device if only a small amount of their data-carrying capacity is being used.

Bandwidth and Speed

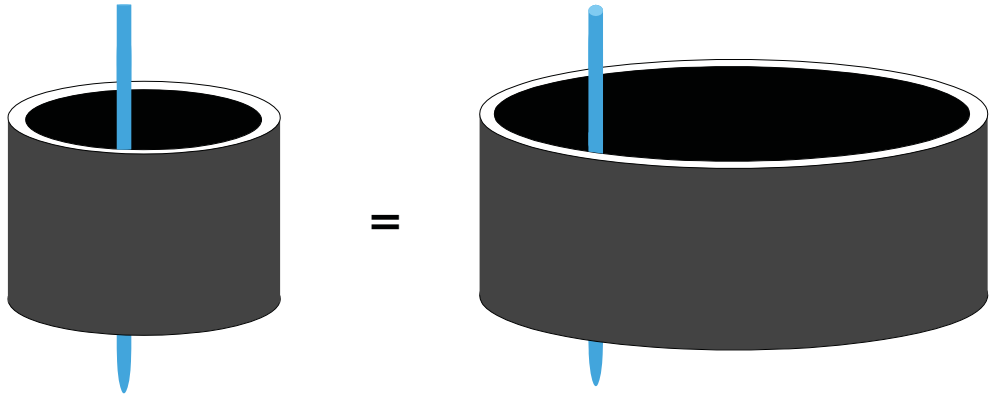
The larger the bandwidth the larger the amount of data that can pass through



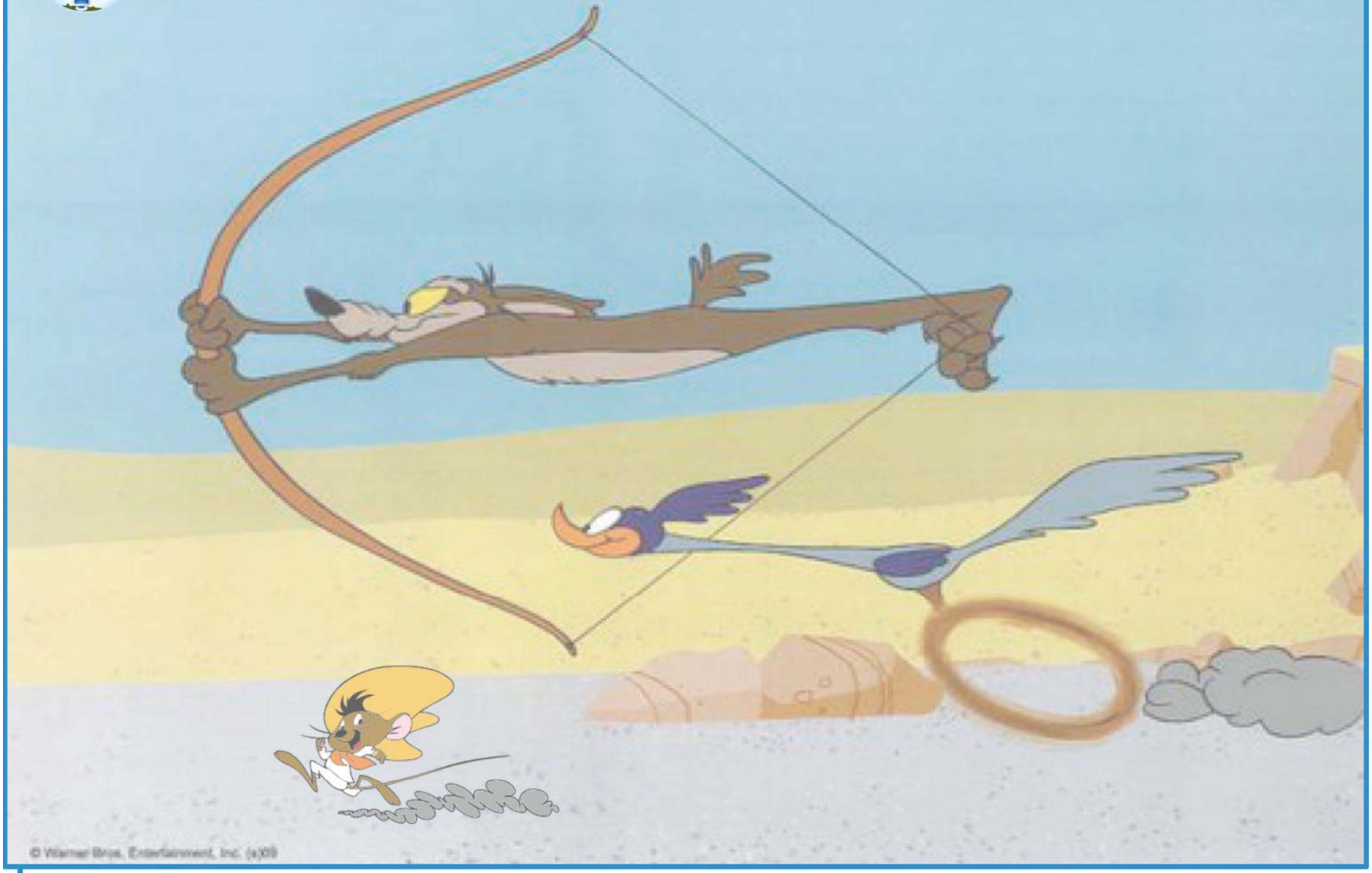
BUT

for amount of data significantly smaller than the actual capacity

BANDWIDTH ≠ SPEED



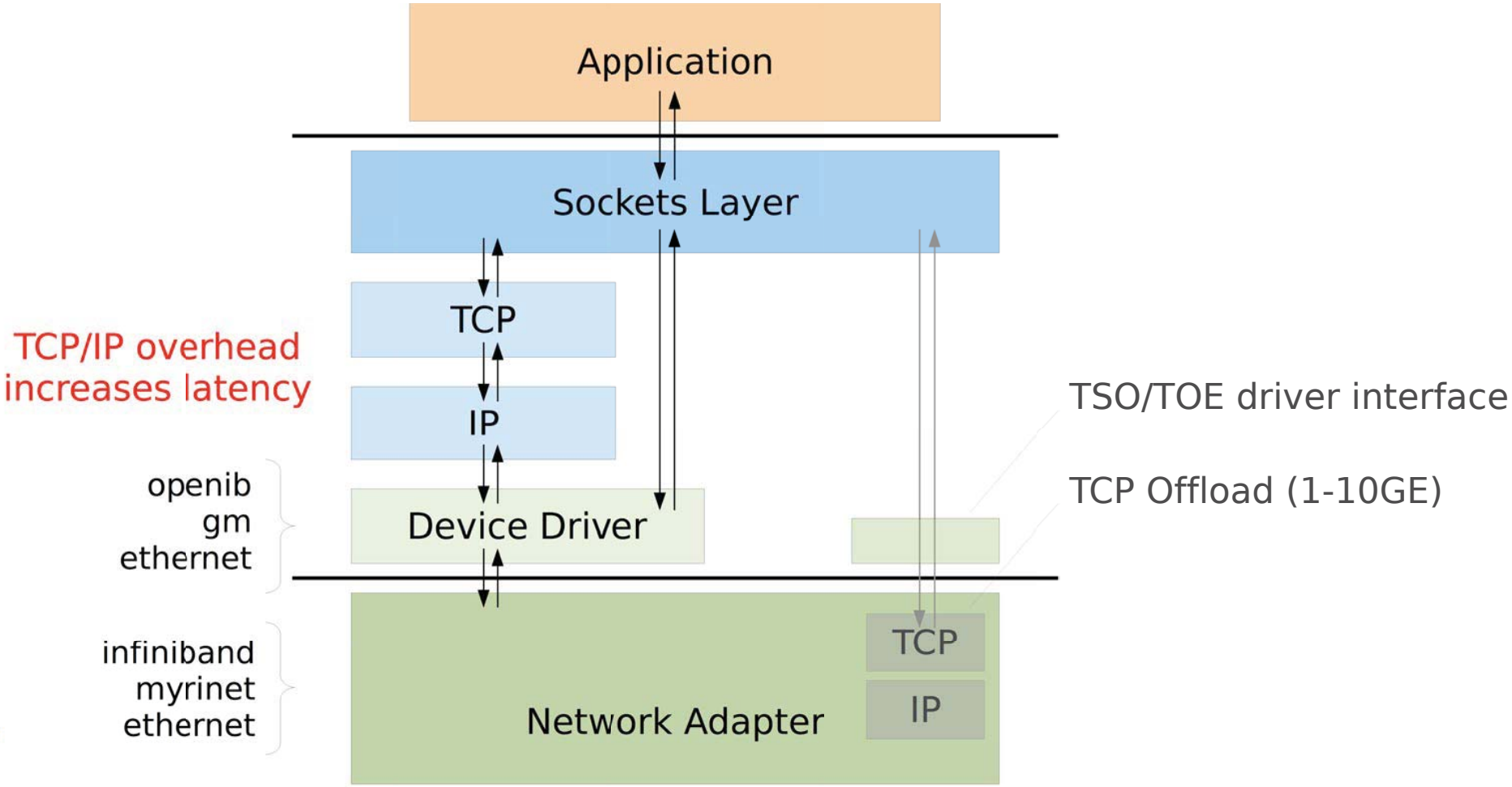
High Speed Networks



What we need from High Speed Networks ?

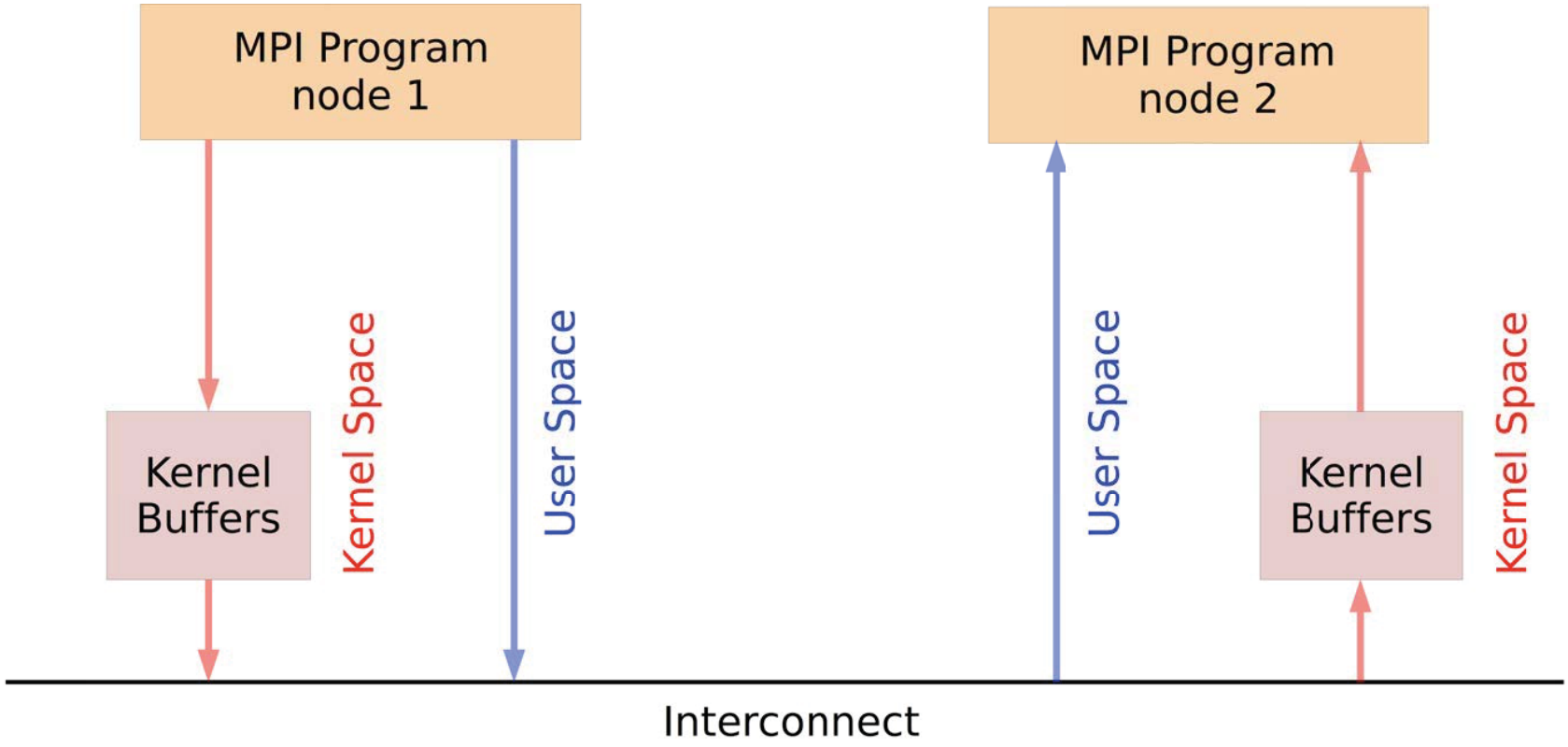
- Intelligent Network Interface Cards
  - Support entire protocol processing completely in hardware (hardware protocol offload engines)
- Provide a rich communication interface to applications
  - User-level communication capability
  - Gets rid of intermediate data buffering requirements
- No software signaling between communication layers
  - All layers are implemented on a dedicated hardware unit, and not on a shared host CPU

Low Latency Networks
TCP/IP vs Native Protocols



Low Latency Networks
Kernel Space vs User Space

Sending a message
(block of memory)
from node1 to node2



Previous High Speed network..

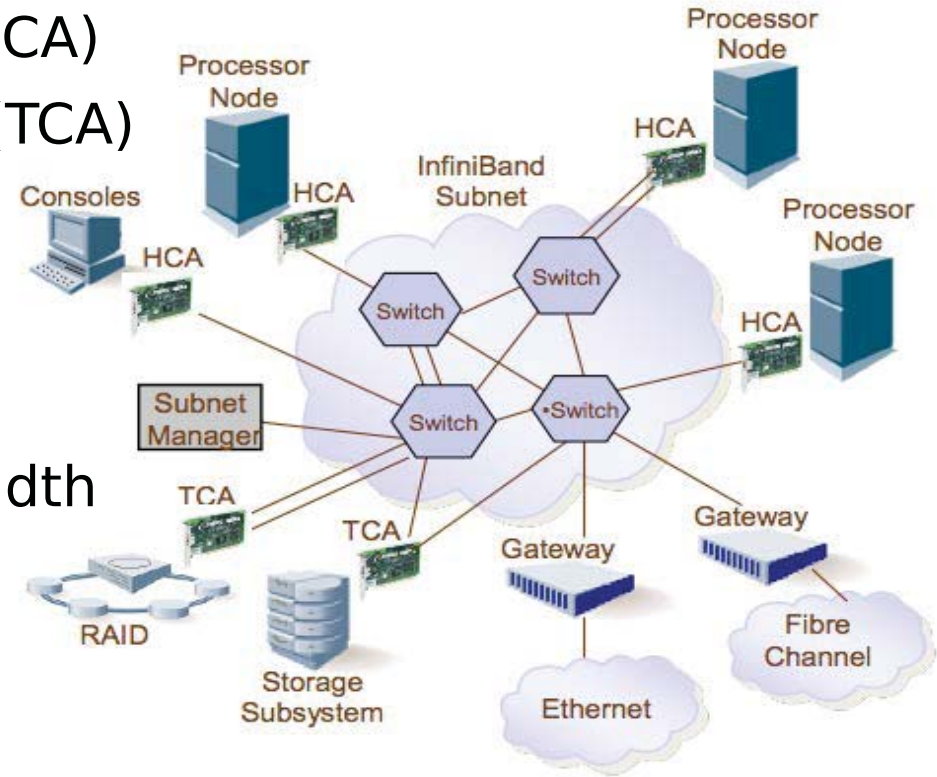
- VIA: Virtual Interface Architecture
  - Standardized by Intel, Compaq, Microsoft
- Fast Messages (FM)
  - Developed by University of Illinois
- Myricom GM
  - Proprietary protocol stack from Myricom
- These network stacks set the trend for high-performance communication requirements
  - Hardware offloaded protocol stack
  - Support for fast and secure user-level access to the protocol stack

What is InfiniBand?

- **Industry standard** defined by the InfiniBand Trade Association – Originated in 1999
- **InfiniBand specification** defines an input/output **architecture** used to interconnect servers, communications infrastructure equipment, storage and embedded systems
- InfiniBand is a pervasive, **low-latency, high-bandwidth interconnect** which requires low processing overhead and is ideal to carry multiple traffic types (clustering, communications, storage, management) over a single connection.
- InfiniBand is now used in thousands of high-performance compute clusters and beyond that scale from small scale to large scale: **de-facto standard**

InfiniBand Architecture

- Defines System Area Network architecture
  - Comprehensive specification: from physical to applications Processor
- Architecture supports
  - Host Channel Adapters (HCA)
  - Target Channel Adapters (TCA)
  - Switches
  - Routers
- Facilitated HW design for
  - Low latency / high bandwidth
  - Transport offload



Infiniband components

Host Channel Adapter (HCA)

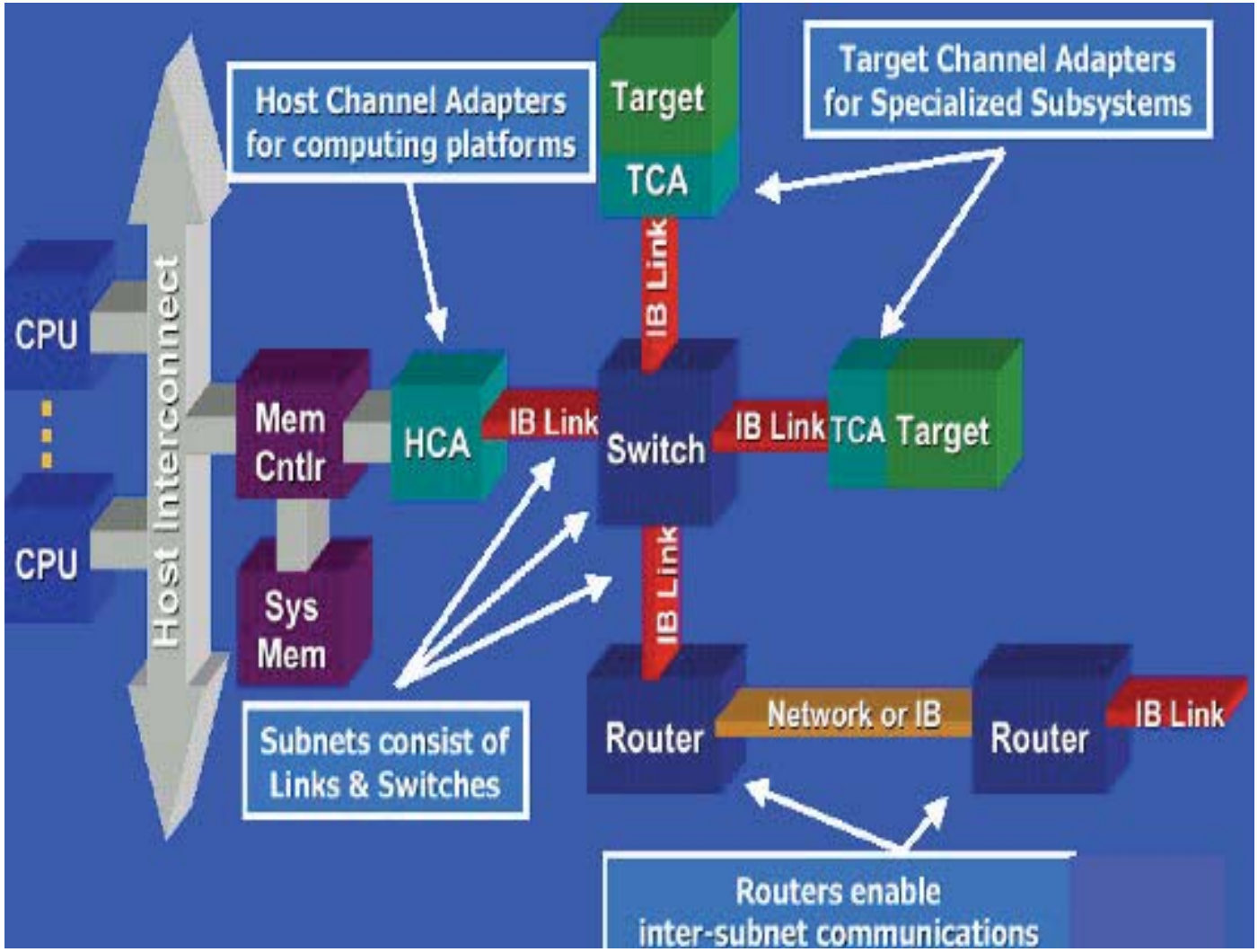
Device that terminates an IB link and executes transport-level functions and support the verbs interface

Switch

A device that routes packets from one link to another of the same IB Subnet

Router

A device that transports packets between IBA subnets



InfiniBand speed (physical layer)

- InfiniBand uses serial stream of bits for data transfer
- Linkwidth
  - 1x – One differential pair per Tx/Rx
  - 4x – Four differential pairs per Tx/Rx
  - 12x - Twelve differential pairs per Tx and per Rx
- LinkSpeed
  - Single Data Rate (SDR) - 2.5Gb/s per lane (10Gb/s for 4x)
  - Double Data Rate (DDR) - 5Gb/s per lane (20Gb/s for 4x)
  - Quad Data Rate (QDR) - 10Gb/s per lane (40Gb/s for 4x)
  - Fourteen Data Rate (FDR) - 14Gb/s per lane (56Gb/s for 4x)
  - Enhanced Data rate (EDR) - 25Gb/s per lane (100Gb/s for 4x)
- • Linkrate
  - Multiplication of the link width and link speed
  - Most common shipping today is 4x ports QDR

Infiniband speed for data transfer..

- For SDR, DDR and QDR, links use 8b/10b encoding:
  - every 10 bits sent carry 8bits of data
- Thus single, double, and quad data rates carry 2, 4, or 8 Gbit/s useful data, respectively.
- For FDR and EDR, links use 64b/66b encoding
  - every 66 bits sent carry 64 bits of data.

Infiniband performance

Infiniband type	LInk Speed	Data Speed	Max Bandwidth (at application level)
SDR 4x	10 Gigabit	8 Gigabit	1 GB/sec
DDR 4X	20 Gigabit	16 Gigabit	2GB/sec
QDR 4X	40 Gigabit	32 Gigabit	4GB/sec

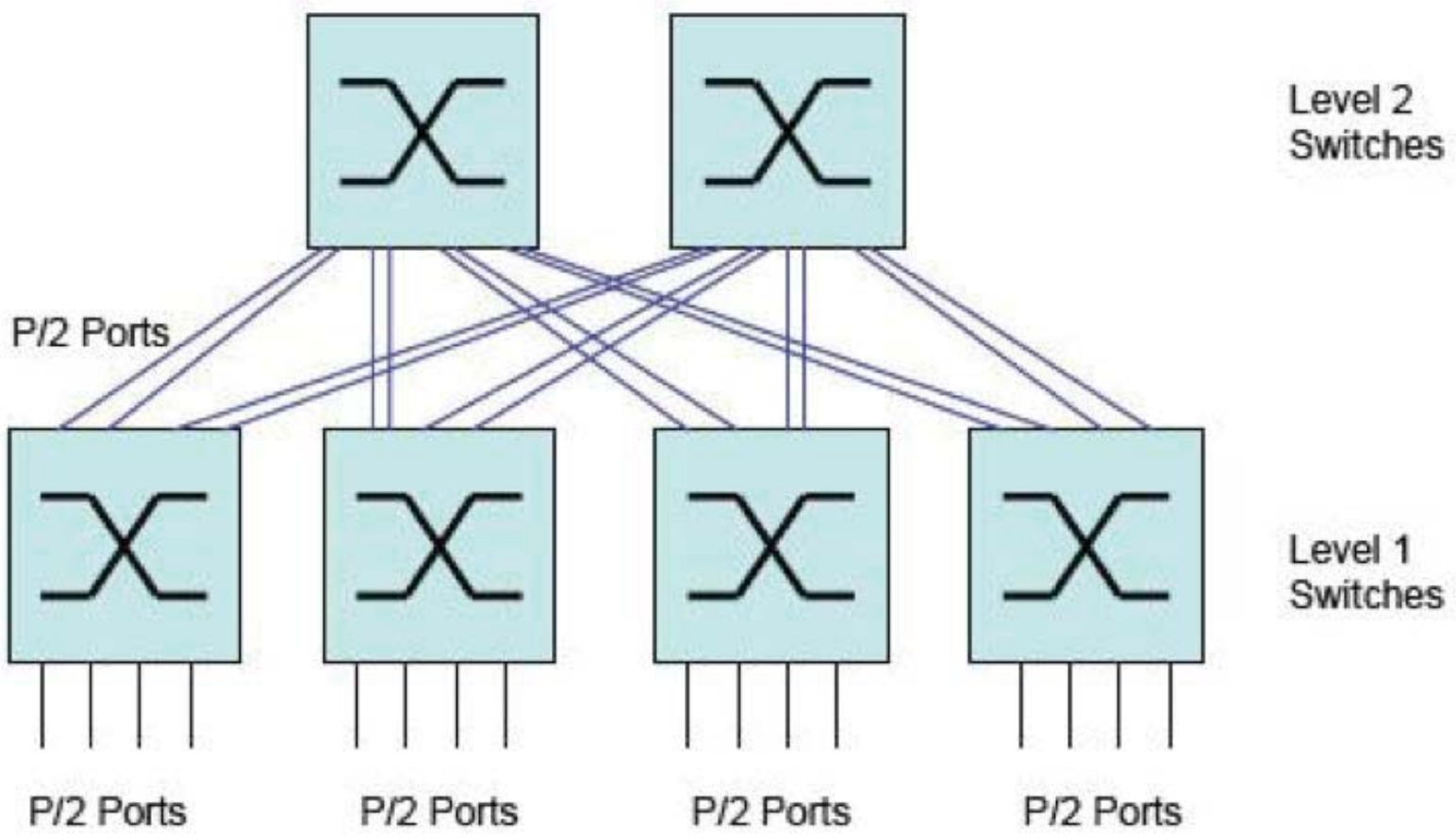
We do not take into account the additional physical layer overhead requirements for common characters or software protocol requirements..

Topologies

Common network topologies

- Fat tree
- Mesh
- 3D torus
- CBB (Constant Bi-sectional Bandwidth)
  - type of Fat tree can be oversubscribed 2:1 to 8:1
  - oversubscription can reduce bandwidth but most applications do not fully utilize it anyway

Two level CBB



Vendors

- Two IB vendors: Mellanox and Qlogic (now Intel)
  - Aligned with many server vendors: IBM, HP, Dell
  - And many integrators: Appro, Advanced Clustering, Microway
- Broadly two kinds of adapters
  - Offloading (Mellanox) and Onloading (Qlogic/INTEL)
- Adapters with different interfaces:
  - Dual port 4X with PCI-X (64 bit/133 MHz),
  - PCIe x8,
  - PCIe 2.0 and HT
- MemFree Adapter
  - No memory on HCA Uses System memory (through PCIe)

Connections

- SDR and DDR: copper CX4
- QDR and FDR: copper or fiber
  - QSFP: Quad Small Form Factor Pluggable
  - also called mini-GBIC (Gigabit Interface Converter)

4x CX4 Fiber

4X CX4

4X QSFP

4x QSFP Fiber

IB software

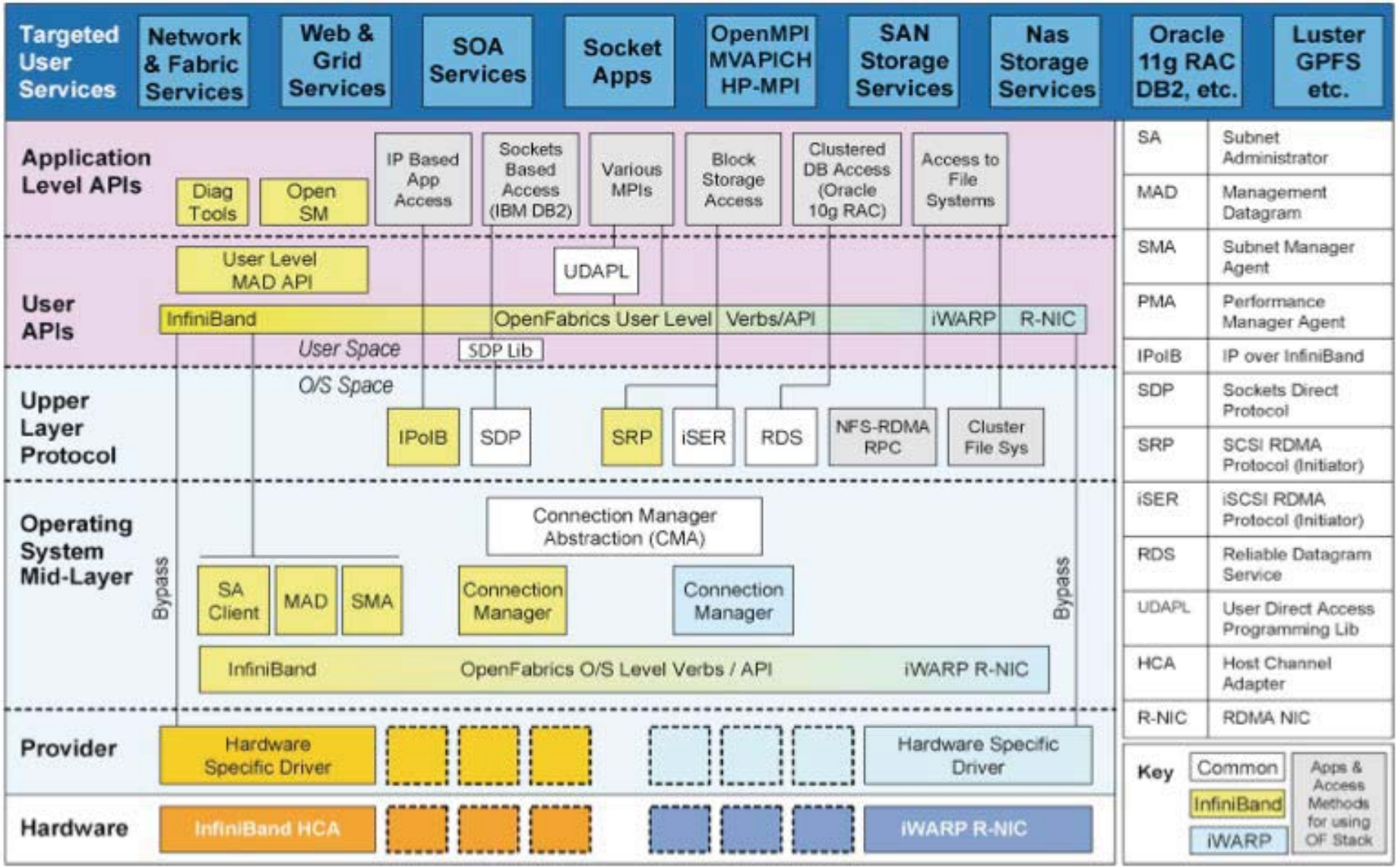
- Provided by OpenFabric (www.openfabrics.org)
- Open source organization (formerly OpenIB)
- Support for Linux and Windows Design of complete stack with `best of breed' components
- Linux Distribution is now including it (check out carefully which version)
- Users can download the entire stack and run
  - Latest release is OFED 3.5.x
- QLogic/Intel and Mellanox could have special add-ons

OFED…

OFED stack includes:

- device drivers
- performance utilities
- diagnostic utilities
- protocols (IPoIB, SDP, SRP,…)
- MPI implementations (OpenMPI, MVAPICH)
- libraries
- subnet manager

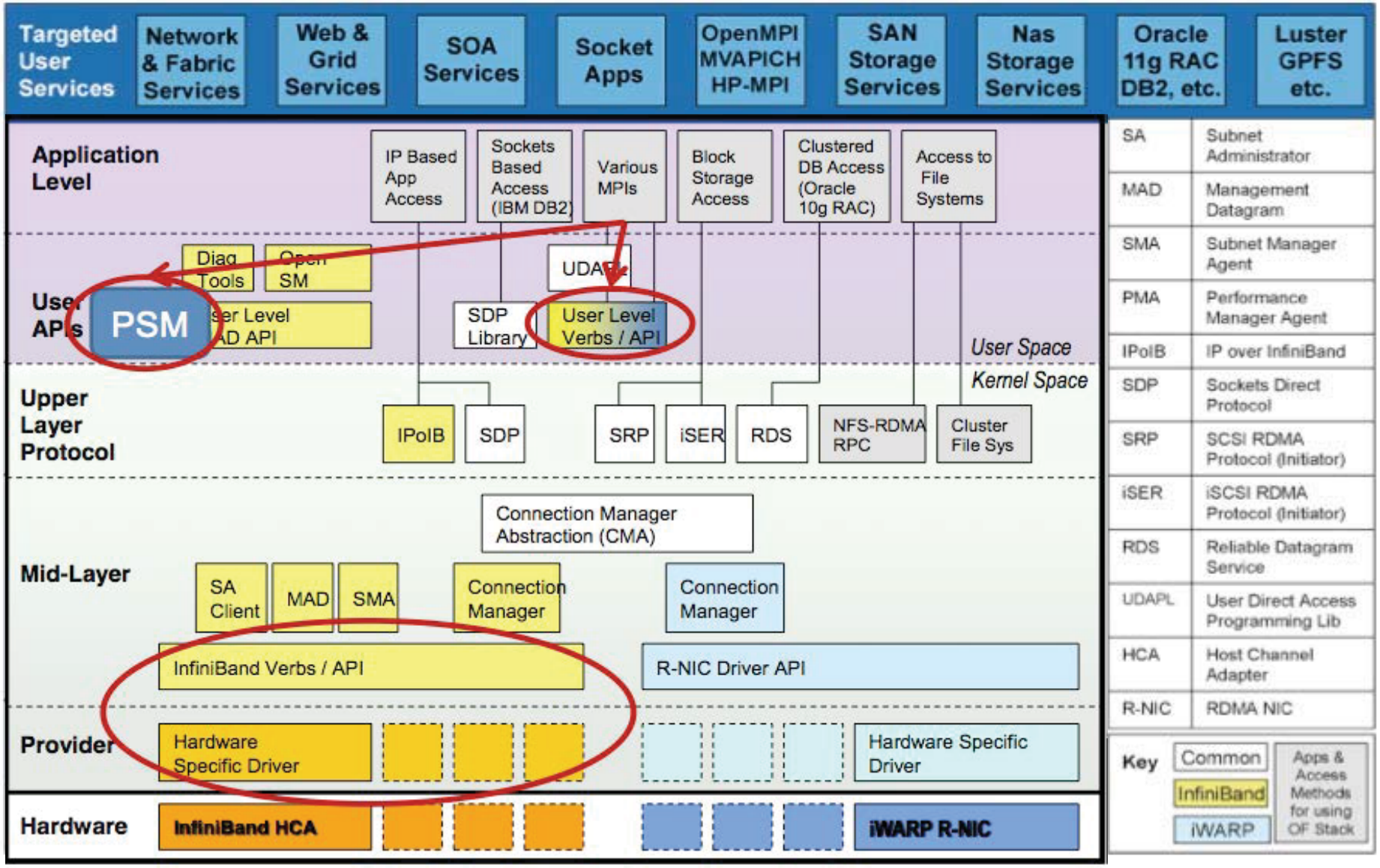
IB software stack..



Subnet Manager

- The Subnet Manager (SM) is mandatory for setting up port ID, links and routes
- OpenSM is an Infiniband compliant subnet manger included with OFED
- Ability to run several instance of osm on the cluster in a Master/Slave(s)configuration for redundancy.
- Routing is typically static:The subnet manager tries to balance the routes on the switches.
  - A sweep is done every 10 seconds to look for new ports or ports that are no longer present.
  - Established routes will typically remain in effect if possible.
  - Enhanced routing algorithms:
    - Min-hop, up-down, fat-tree, LASH, DOR, Torus2QOS

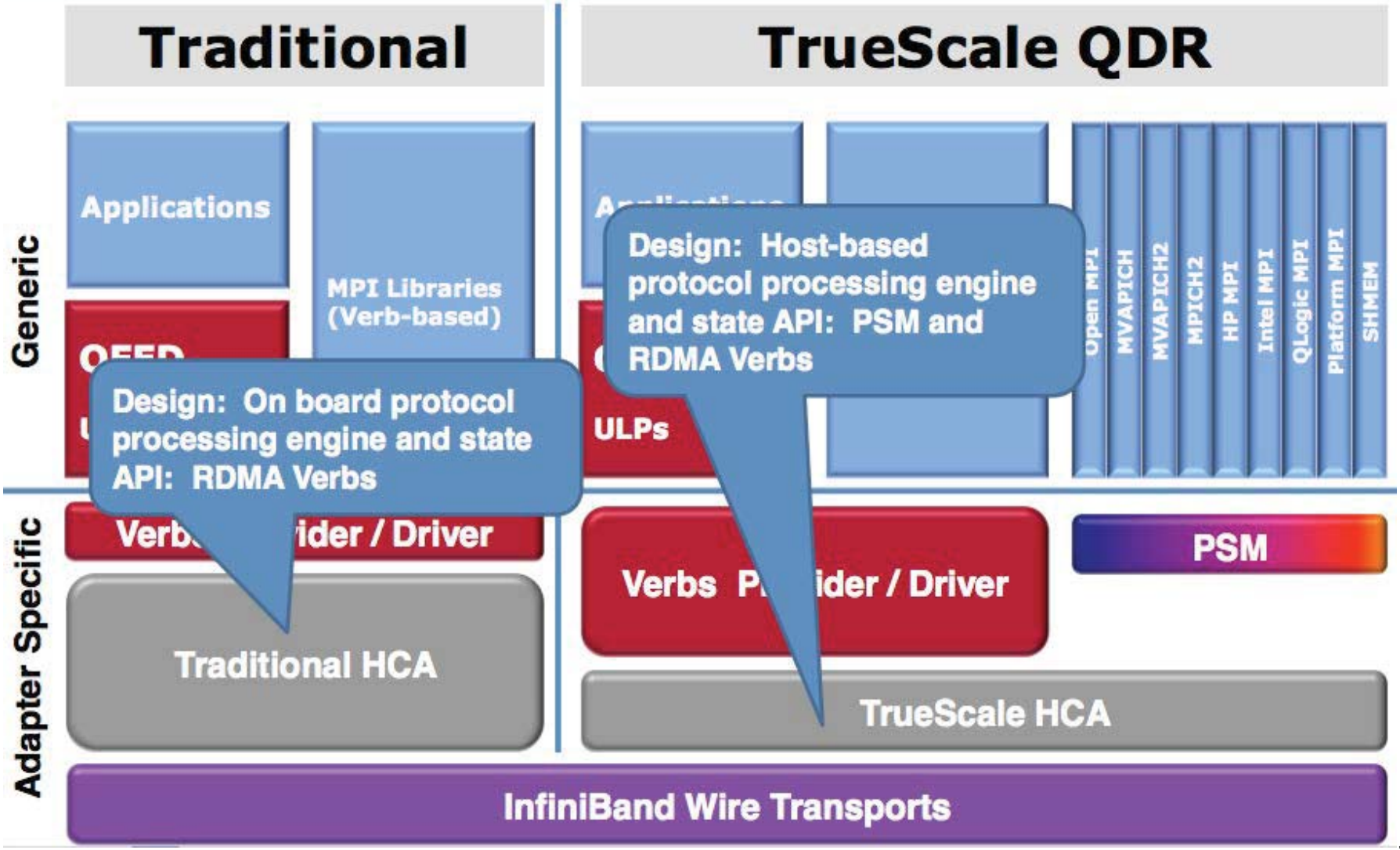
IB software stack for Intel/Qlogic



Performance Scaled Messaging

- PSM added as part of OFED by Intel/Qlogic
  - InfiniBand Library and API
  - Designed to maximize the performance and scalability of MPI applications
  - Integrated with all major MPI's & supports all Collectives
- PSM API is a good impedance match for MPI semantics
  - Allows efficient layering of higher level MPI libraries
  - Optimized communications for MPI, while minimizing CPU utilization

PSM vs... ibverbs



Infiniband prices..

- HCA
  - Mellanox HCA ~ 536 euro
  - Truescale HCA ~ 240 euro
- Switches
  - QLogic 12300 18Port 40Gb/s IB Switch ~ 4000 Euro

    Mellanox MIS5023Q-1BFR - 18Port 40Gb/s IB Edge Switch ~ 4000

Infiniband performance

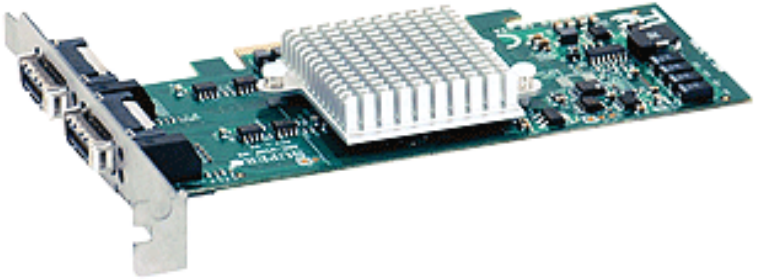
- Mellanox:
  - 3.2Gb/second MPI bandwidth
  - 1.4 ~ 1.5 microsecond latency
- Qlogic/Intel:
  - 2.6Gb/second MPI bandwidth
  - 1.9 ~ 2.0 microsecond latency
- Gigabit card
  - 120 Mb/sec MPI bandwidth
  - 50/60 microseconds latency

Which network for your cluster ?

Difficult choice

- A few questions to help
  - Which kind of cluster (HTC or HPC) ?
  - Which kind of application ?
    - Serial/Parallel
    - Parallel loosely coupled / tightly coupled ?
    - Latency or bandwidth dominated ?
  - I/O considerations
    - Only MPI or Storage as well ?
  - Budget considerations

Gigabit vs InfiniBand



COMMODITY

easy to manage

Complex to manage

Expensive

High latency

Low BW

Low latency

High BW

Why is (low) latency so important?

According to **Amdahl's law:**

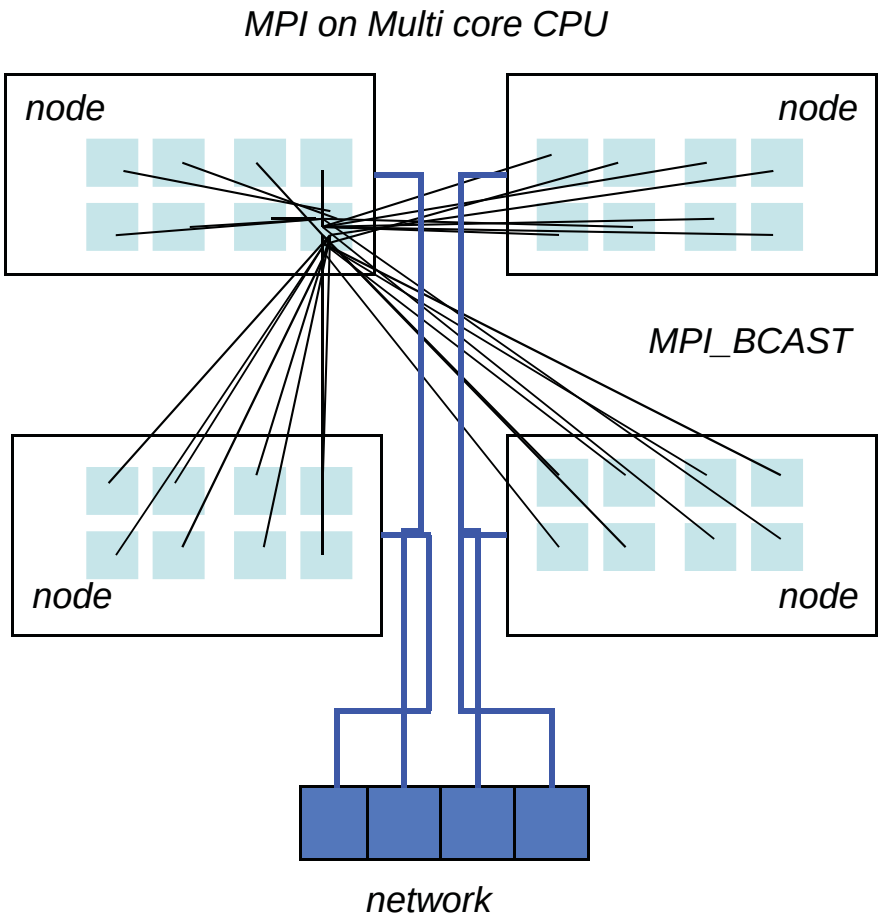
- **a high-performance parallel system tends to be bottlenecked by its slowest sequential process**

- in all but the most embarrassingly parallel supercomputer workloads, **the slowest sequential process is often the latency of message transmission across the network**

A few more considerations

- In general the compute/communication ratio in a parallel program remains fairly constant.
- So as the computational power increases the network speed must also be increased.
- As multi-core processors proliferate, it is increasingly common to have 8, 10 or even 16 MPI processes sharing the same network device.
- Contention for the interconnect device can have a significant impact on performance.

Multicore & MPI issues



1 MPI process / core:
- Stress network
- Stress OS

Final remarks

- High speed networks should have:
  - high bandwidth
  - high throughput
  - low latency
- Things to keep in mind:
  - choose the right topology for your needs (both physical and logical)
  - figure out what will be your typical data patterns (small/large chunks, frequent access, ...)
  - bet on reliable hardware
  - consider the cost