High Performance Networking

Infiniband or Ethernet?

Martin Hilgeman

EMEA product technologist HPC

Agenda

Agenda

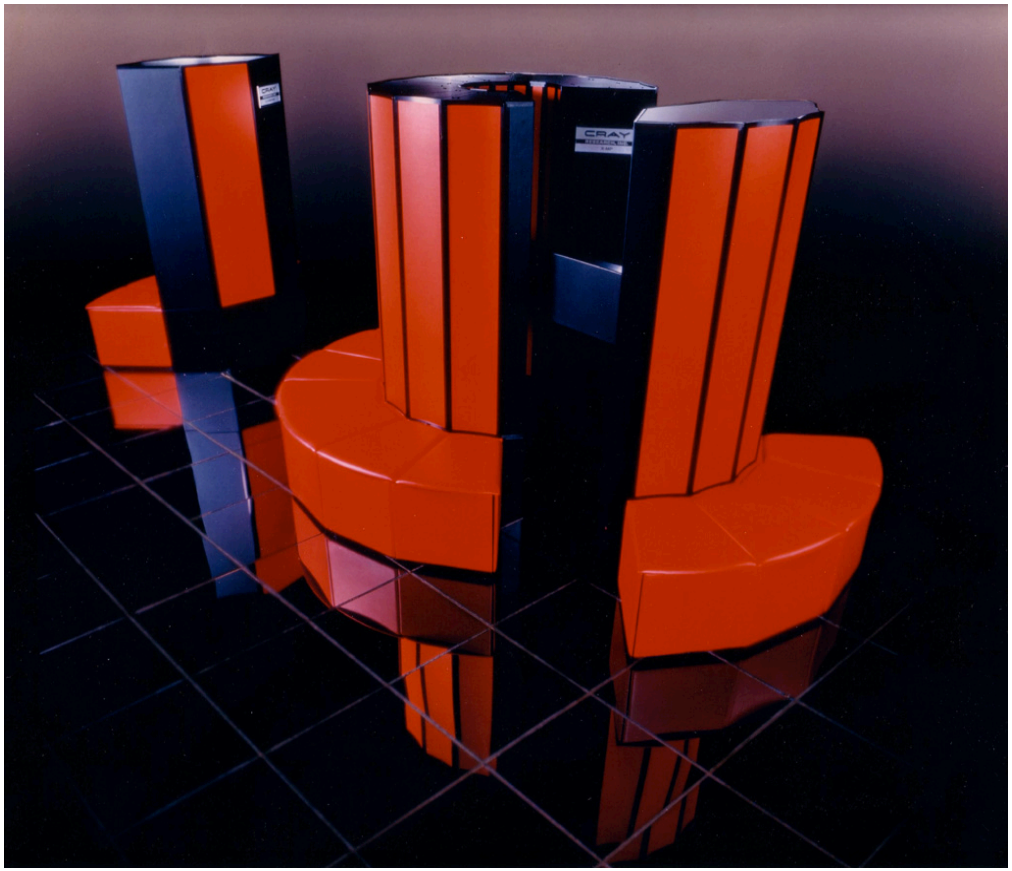
- What is High Performance Computing?
- Why HPC needs a network
- Networking 101
- MPI
- Benchmark environment
- Results
- Summary

High Performance Computing (HPC)

What is HPC?

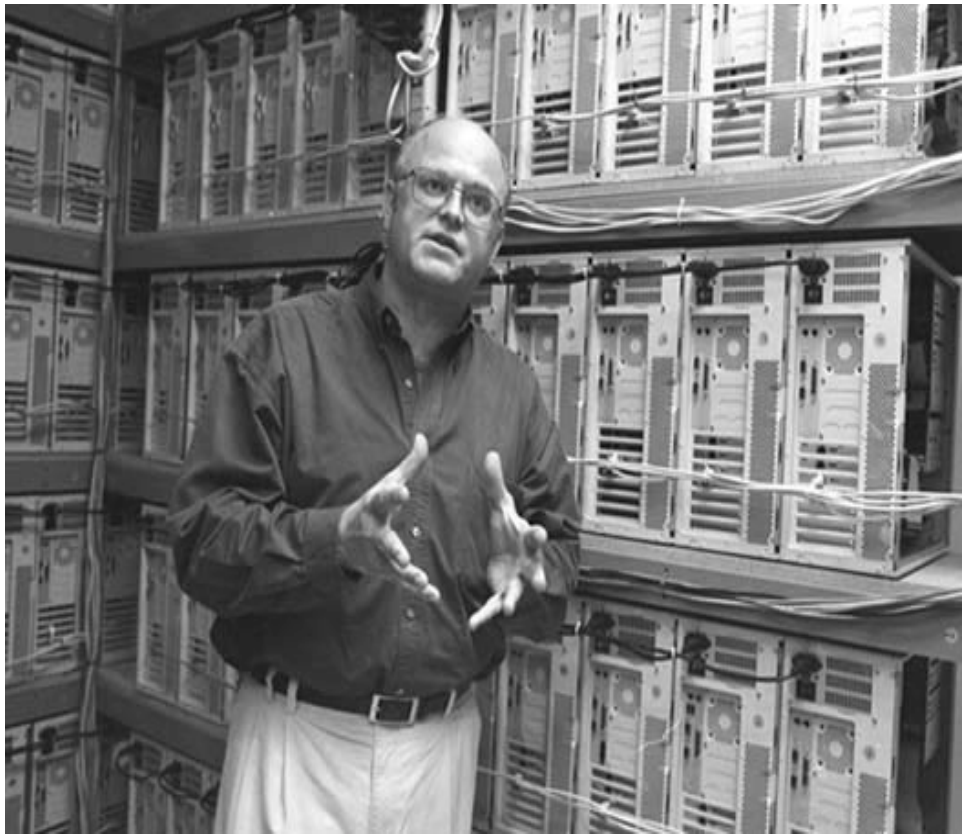
- HPC stands for High Performance Computing
- Performance is critical, both in terms of throughput and capability
- HPC systems typically try to solve problems that too big to fit onto a single computer or workstation
- Important sectors are
- Oil and gas exploration
- Weather forecasting
- Materials design
- Academic sciences

1980s – specialized machines for vector processing



- Note: money left over after building the computer for nice red paint and seats
- Specialist HPC machines
- Proprietary software stack

1990s – The Beowulf project (NASA)



- Beowulf cluster put together using standard off-the-shelf computers
- Specially tuned Ethernet drivers developed on Linux
- Complete open-source software stack

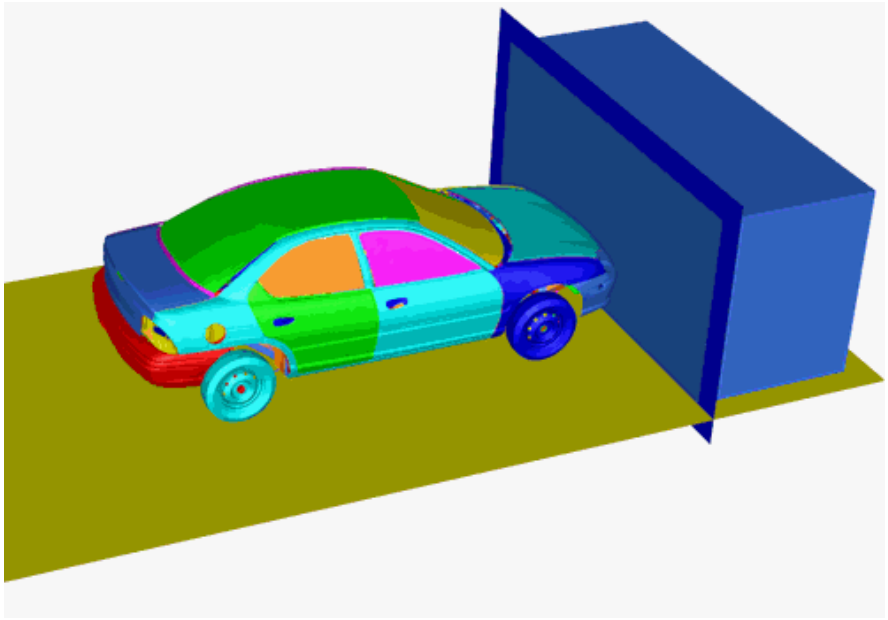
What we are doing now

Workloads – Real life problems (1)

Think of workloads like:

- A calculation for a car manufacturer simulating an NCAP crash test

LSTC
Livermore Software Technology Corp.

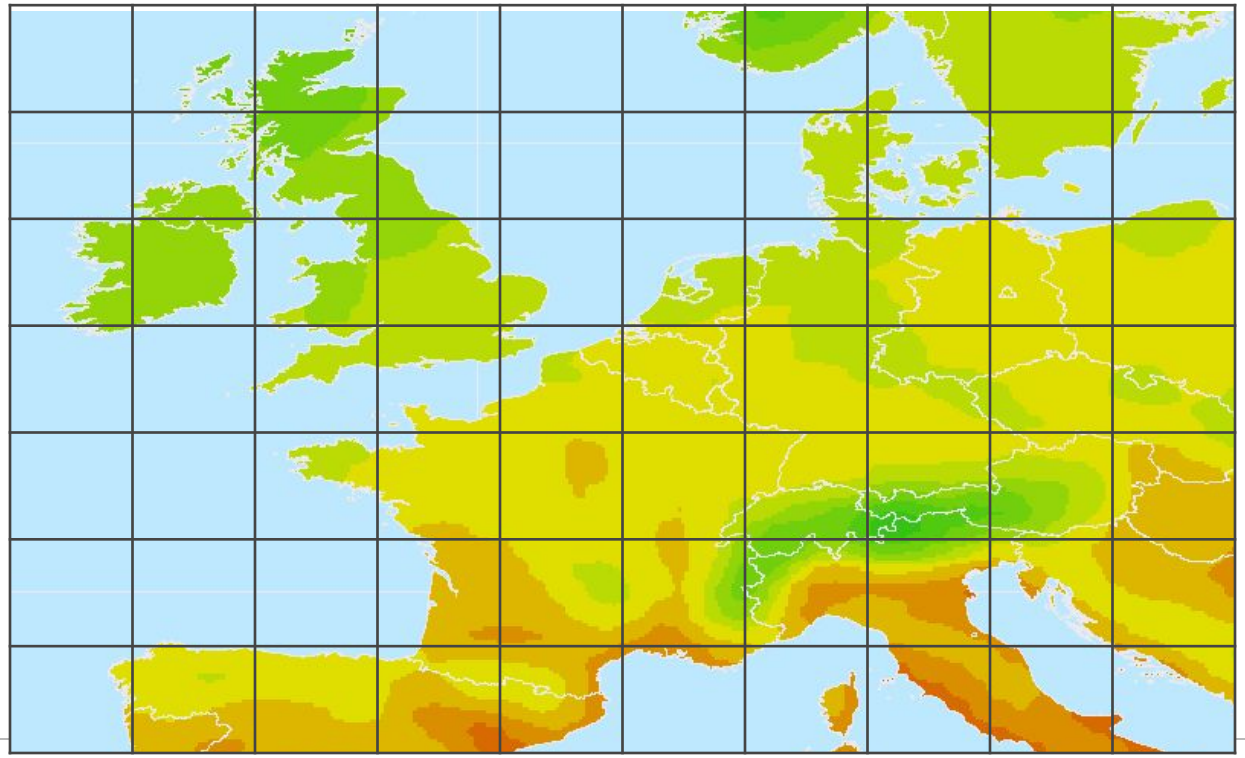


Workloads – Real life problems (2)

Think of workloads like:

- A run for Met Office that calculates the weather for Northern Europe in a 50x50 km resolution up to 1500 meters above sea level

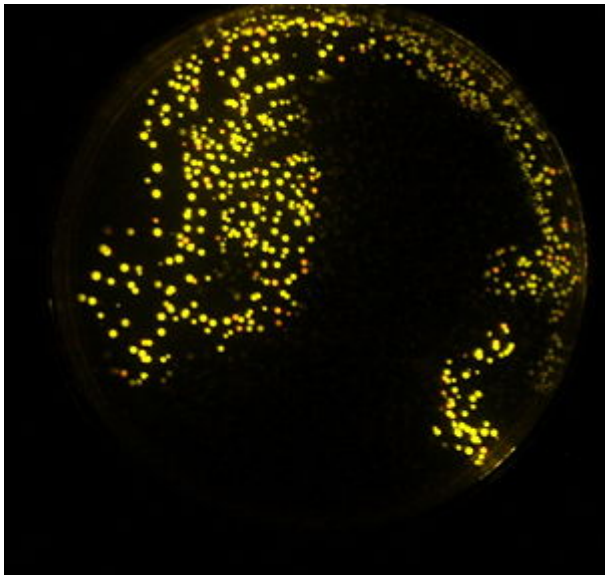
Met Office

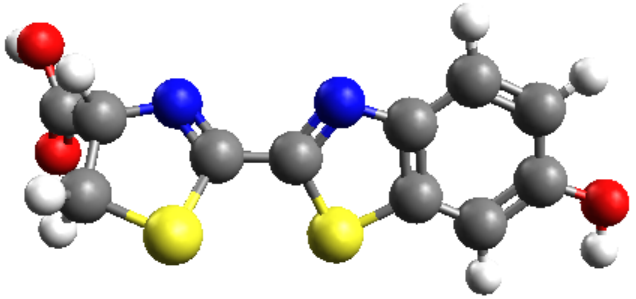


Workloads – Real life problems (3)

Think of workloads like:

- A run for the chemistry department calculating the molecular properties of the *luciferin* molecule





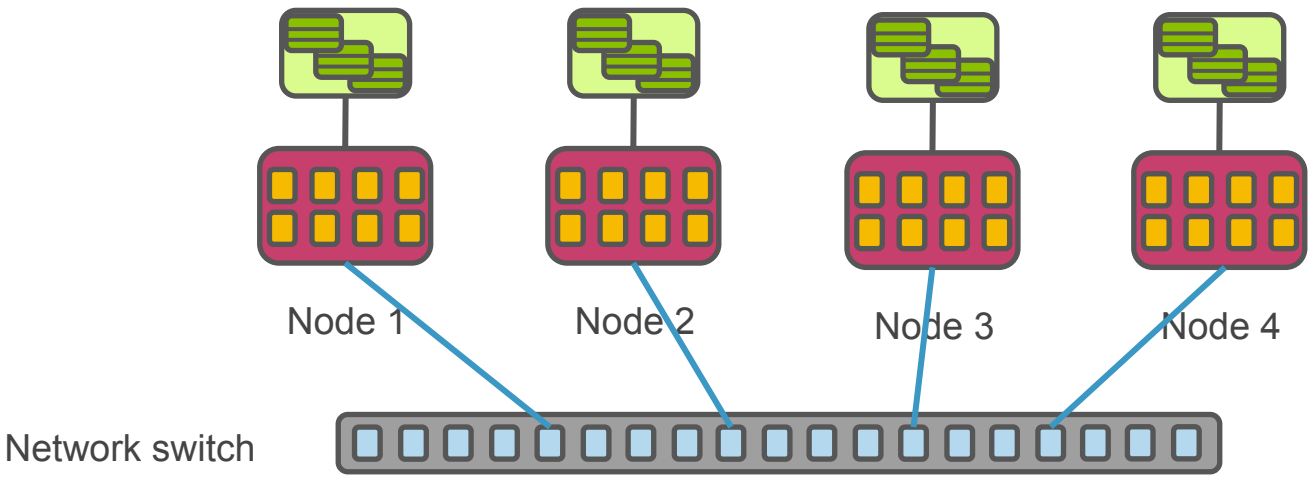
GÅMESS

$$H\Psi(\mathbf{R},\mathrm{r}) = E\Psi(\mathbf{R},\mathrm{r})$$

HPC networking

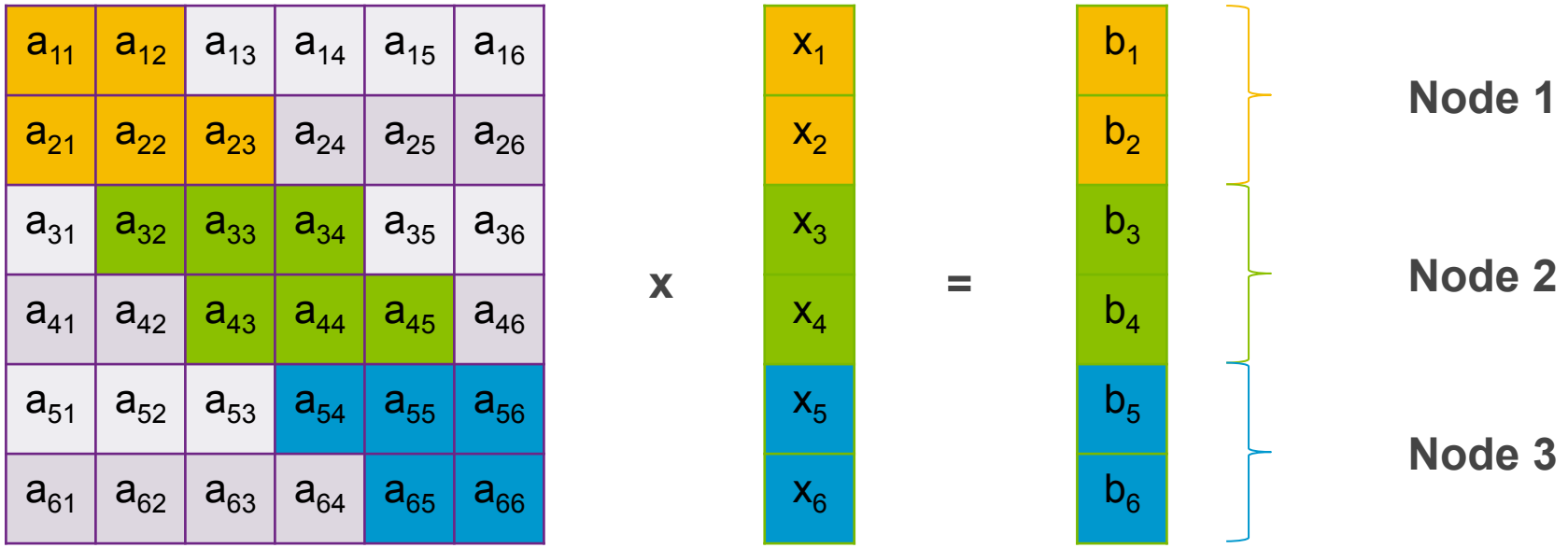
How are all these cores connected?

- As mentioned before, a cluster consists of nodes that are connected with each other through some kind of network



Why are nodes connected with each other (1)?

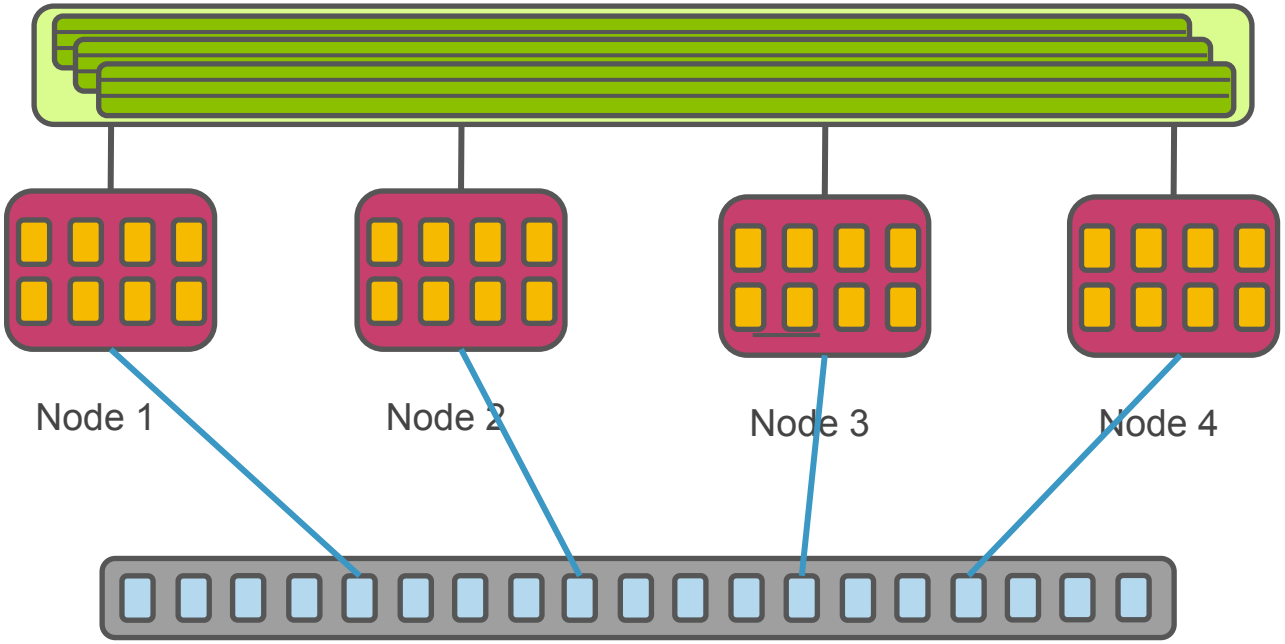
- To divide a computational problem across nodes



- Nodes need to communicate to distribute data and to merge partial results into a final result

Why are nodes connected with each other (2)?

- *"Big data"* : To work on a problem that is too big to fit on one node



- Nodes need to communicate to get data from each other's memory

Make a cluster look like a single system

Obviously we need a network between the components (aka nodes) that is

- Reliable
- Fast
- Expandable
- Cost effective

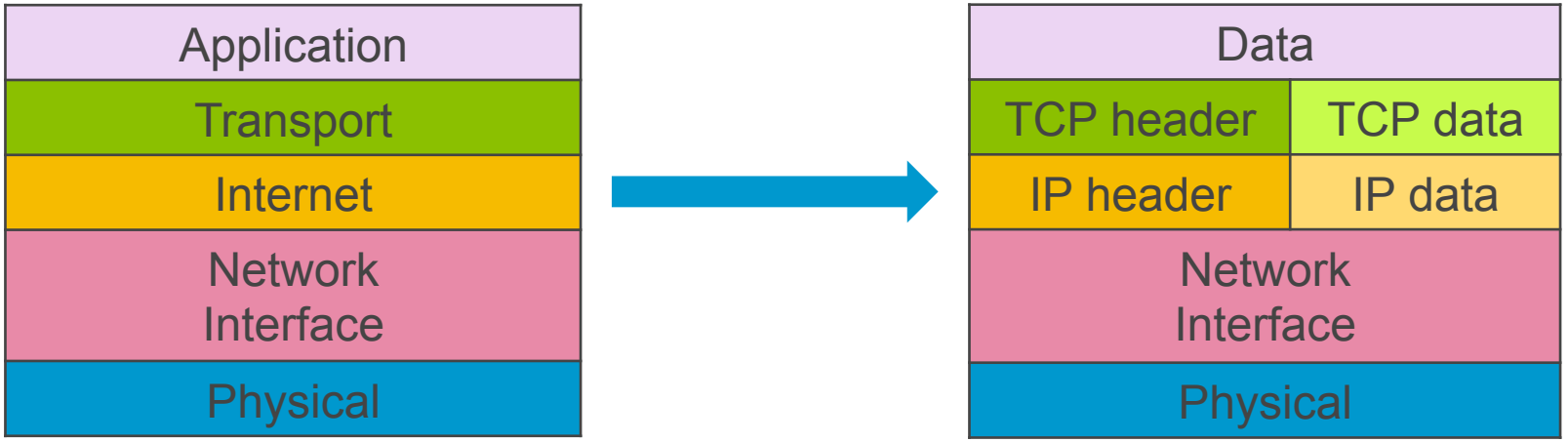
Being HPC, the speed is what matters most. This can be expressed as

- Bandwidth (how much data can be transferred)
- Latency (how much time does it take to travel)

Networking 101

How about protocols?

- TCP/IP has had it's 30th Birthday
- RFC 1122 in practice

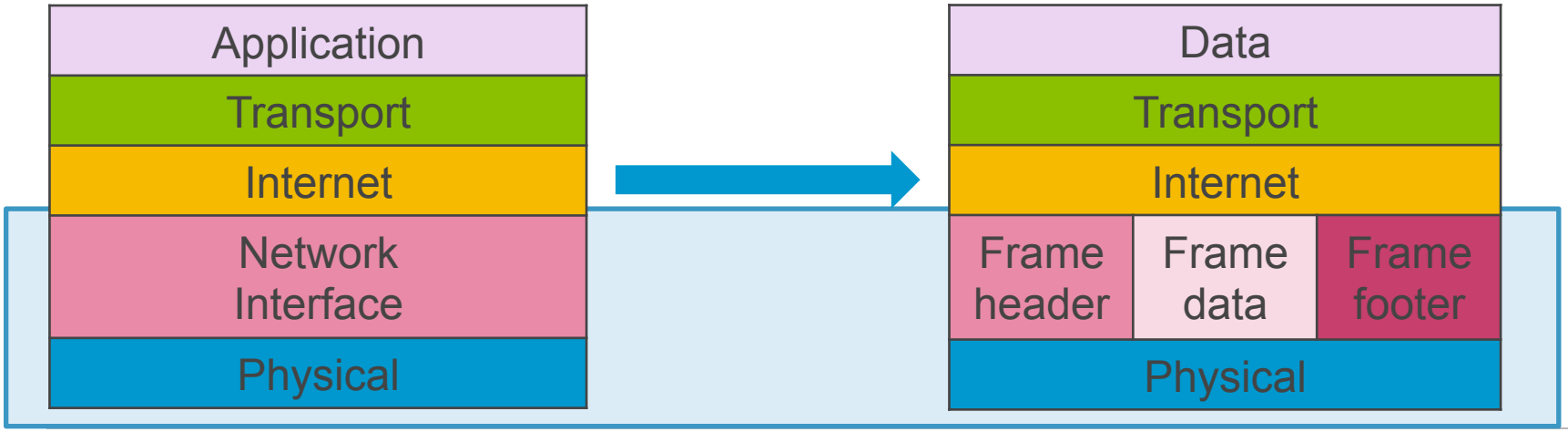


- 4 layers to transfer data from one device to the next

Ethernet

In the old days there was Ethernet

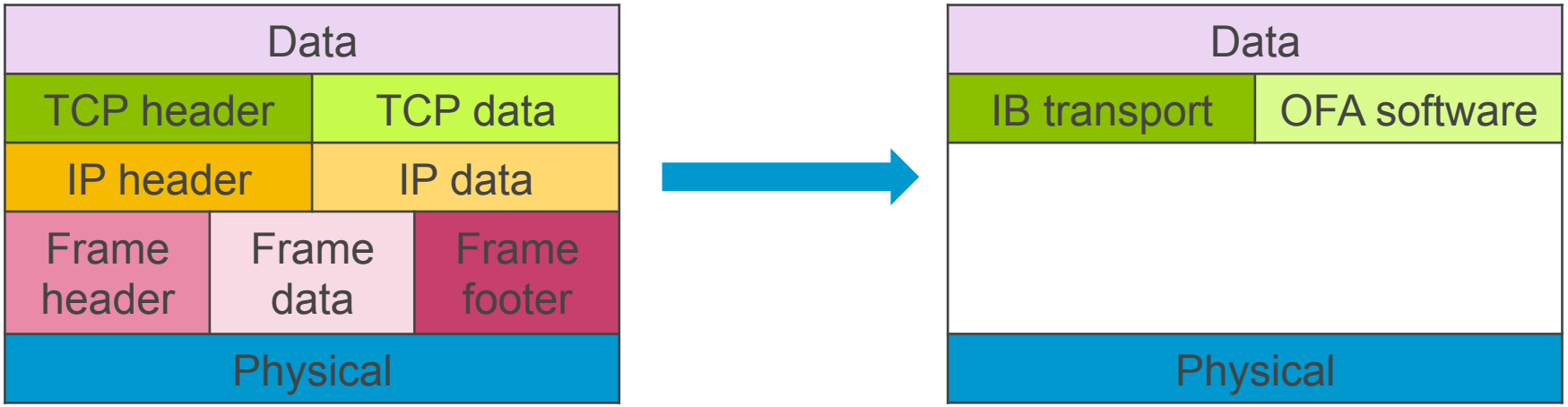
- 1 Gbps bandwidth
- 30 µs latency
- Data is being divided into frames
  - Contains source and destination address
  - Collisions occur then multiple sources do transmit at the same time



Infiniband

De-facto standard in HPC

- Up to 56 Gbps bandwidth
- <1 µs latency
- Uses multiple point-to-point serial links
  - Switched fabric
  - Data is sent in packets of 4kB to form a message



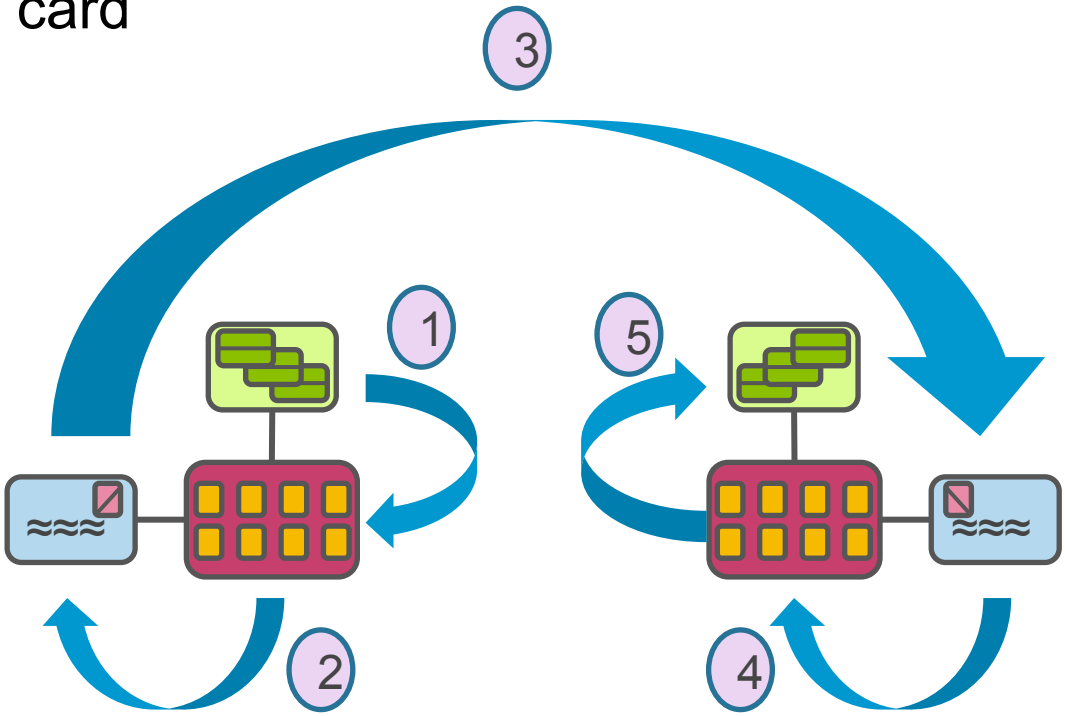
Infiniband message types

Data is transferred in packets of 4 kB that form a message. Types can be:

- Direct memory access read/write from/to a remote node (RDMA)
- Channel send or receive
- Multicast transmission.
- Atomic operation

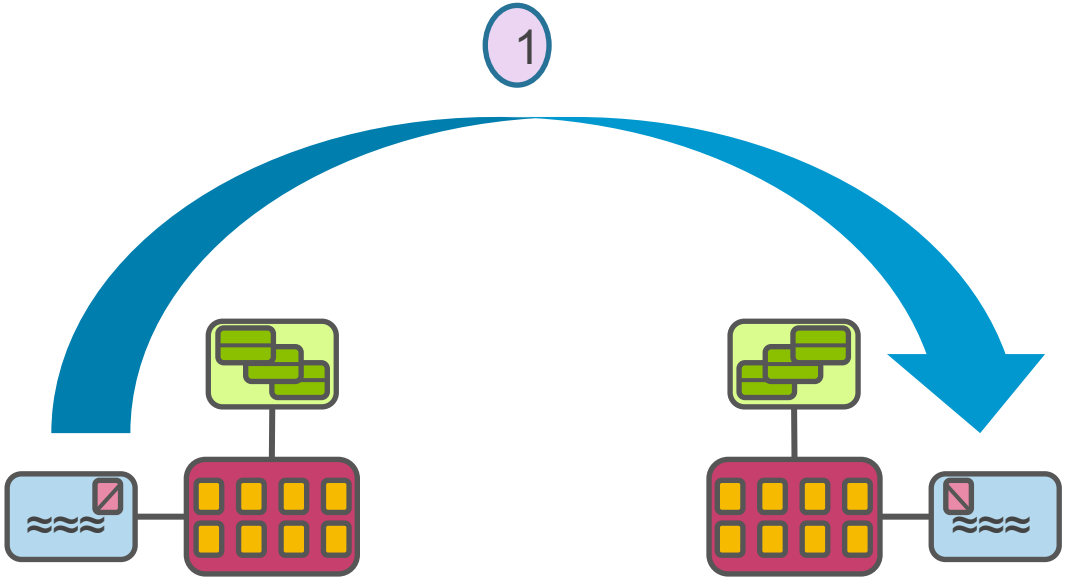
Data transfer between 2 nodes

- Data is pre-staged in local memory and copied to a buffer on the network card



Data transfer between 2 nodes with RDMA

- Data is directly copied to the memory of the remote node



RDMA over Ethernet

RDMA can also be done over Ethernet by

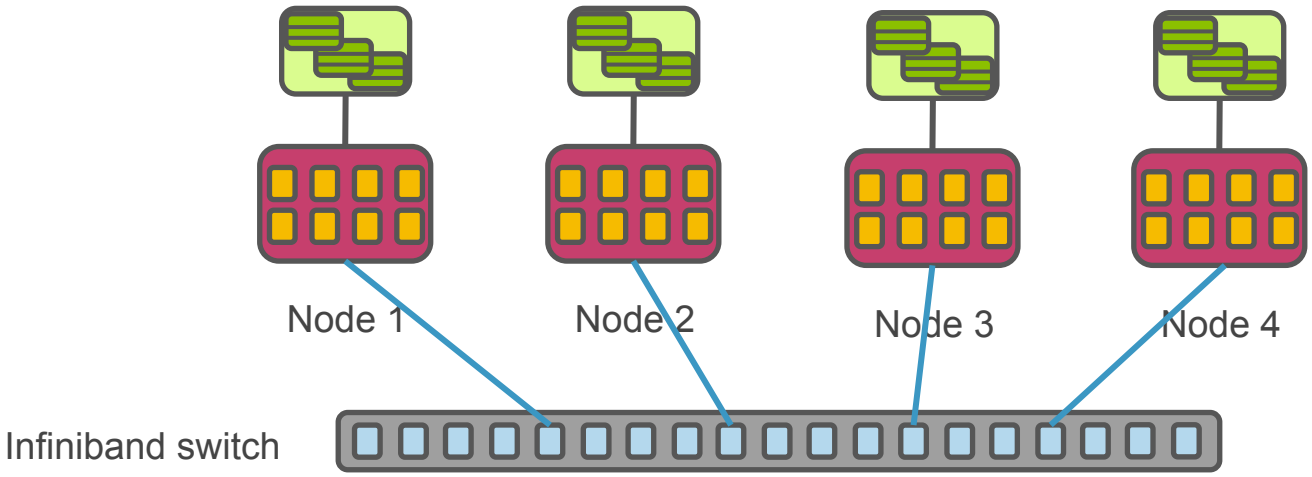
- Stripping the IB GUIDs out of the data header and replace them with MAC addresses (RoCE, Mellanox-only)
- Alternative, zero-copy implementation for TCP that works on top of IP (iWARP)

- Bandwidth is what is offered by Ethernet, but what about latency?

MPI

MPI 101

- MPI is a programming interface that allows one to use multiple systems (or nodes) to work on a single problem in parallel
- Each cluster node has its own memory space and is interconnected through a **fast, low latency network**

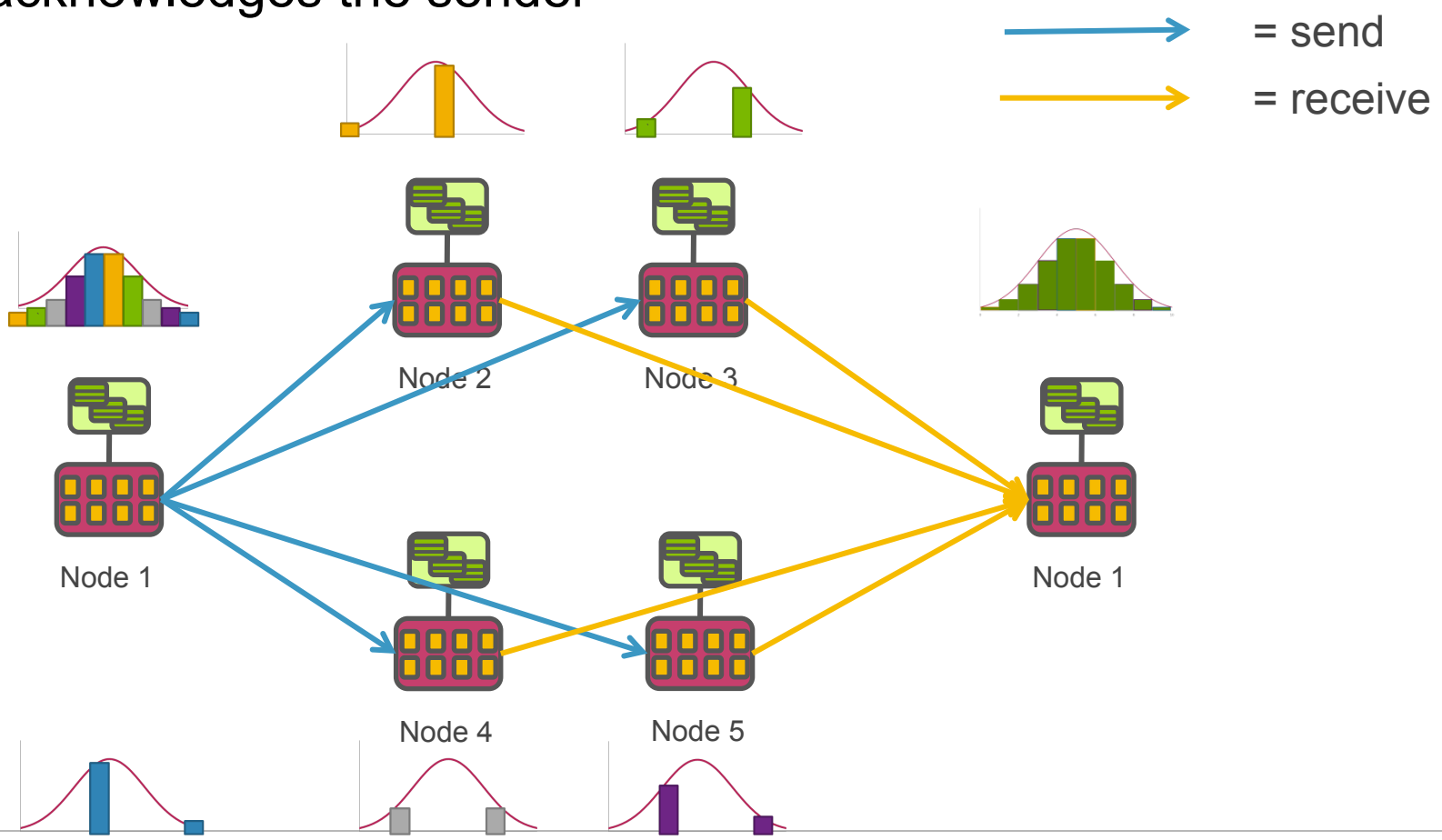


Think of an express service

Data is transferred between nodes as *messages* (with an envelope containing *sender*, *destination* and *reference* tag)

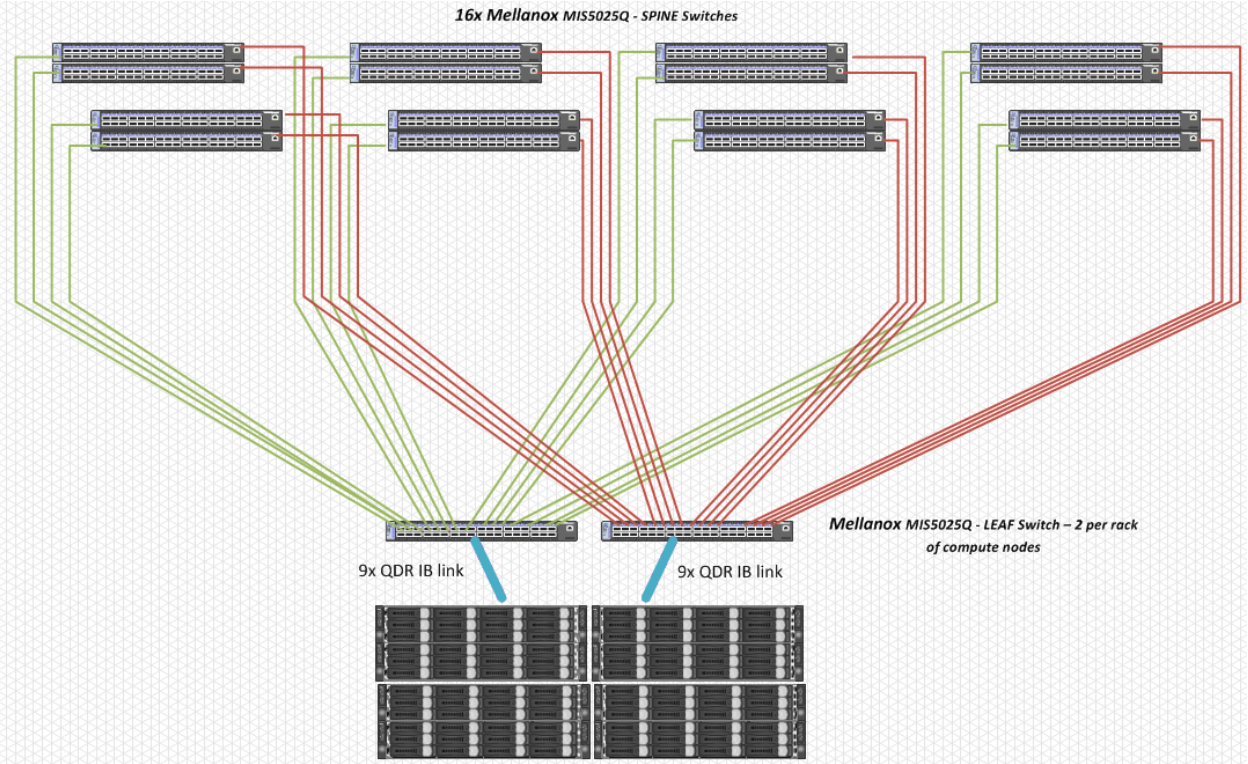
MPI 101

- MPI communication is normally **2-way** (the receiver acknowledges the sender



MPI 101

- 90% of all the HPC applications that run in parallel use MPI
- Performance is measured in terms of bandwidth and latency



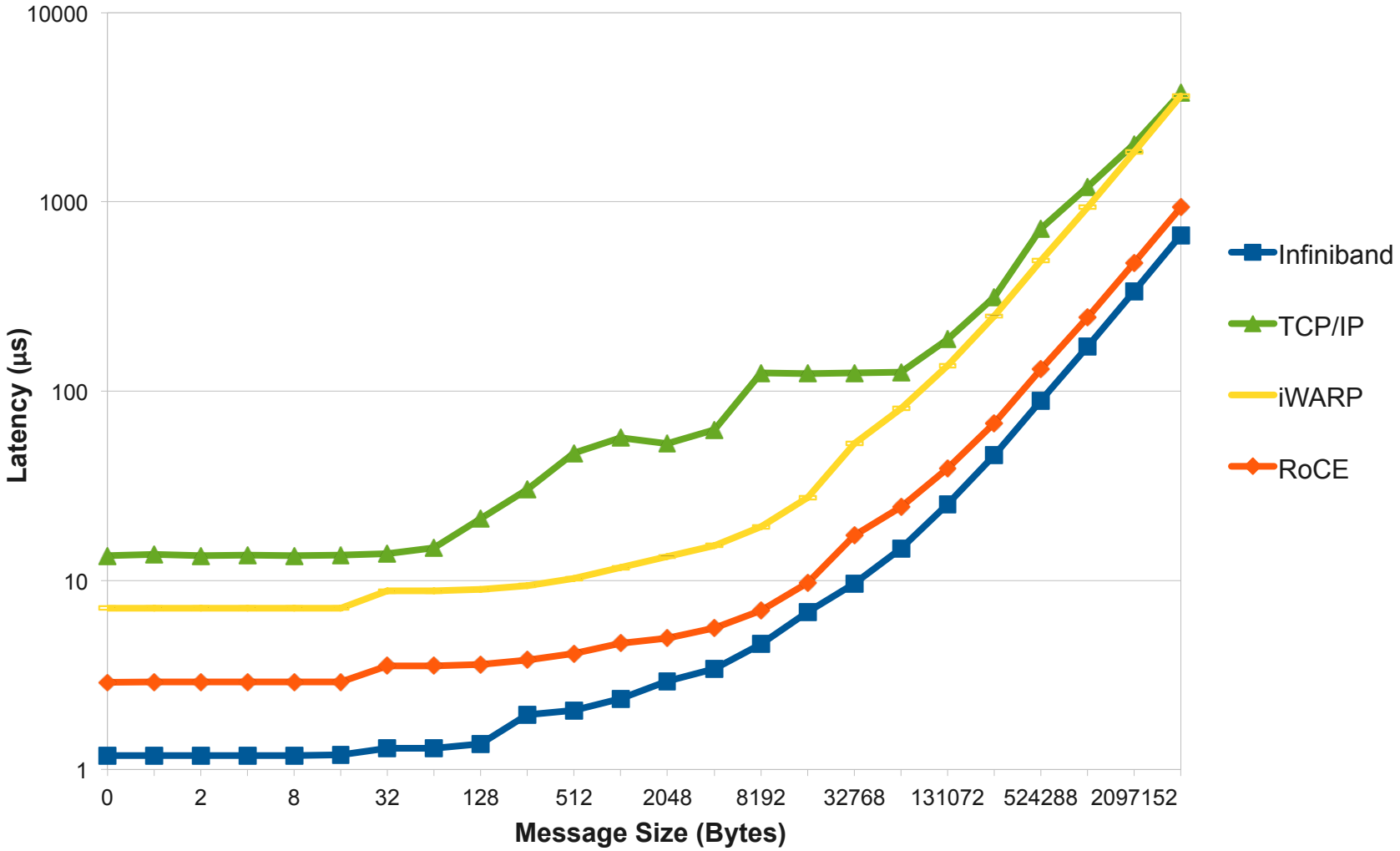
Benchmark setup

4 different test beds

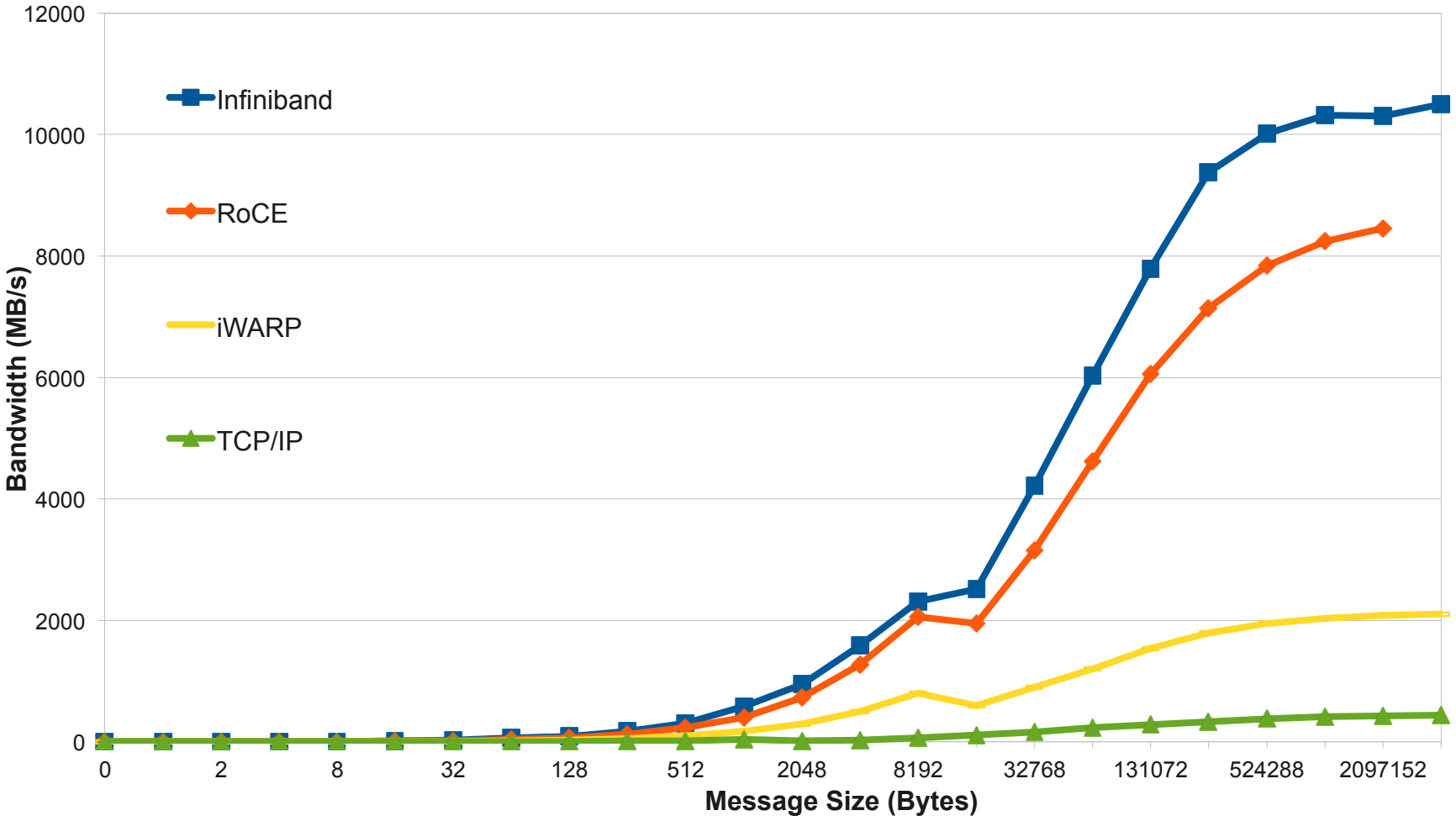
Network Type	Speed (Gbps)	Nodes	Network card	Network switch
Infiniband	56	4	Mellanox ConnectX-3	Mellanox SX6536
RoCE	40	4	Mellanox ConnectX-3	Dell Force10 Z9000
iWARP	10	4	Intel X520	Dell PowerConnect 6348
TCP/IP	10	4	Intel X520	Dell PowerConnect 6348

Performance

Point-to-Point latency



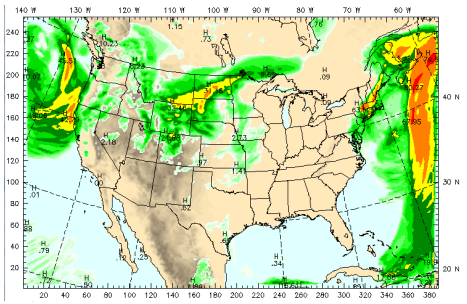
Point-to-Point bandwidth



What about applications?

Application parallel load imbalance and their varying communication pattern can be far bigger performance factors

- Three different applications have been tested:
  - Numerical weather forecasting (WRF)



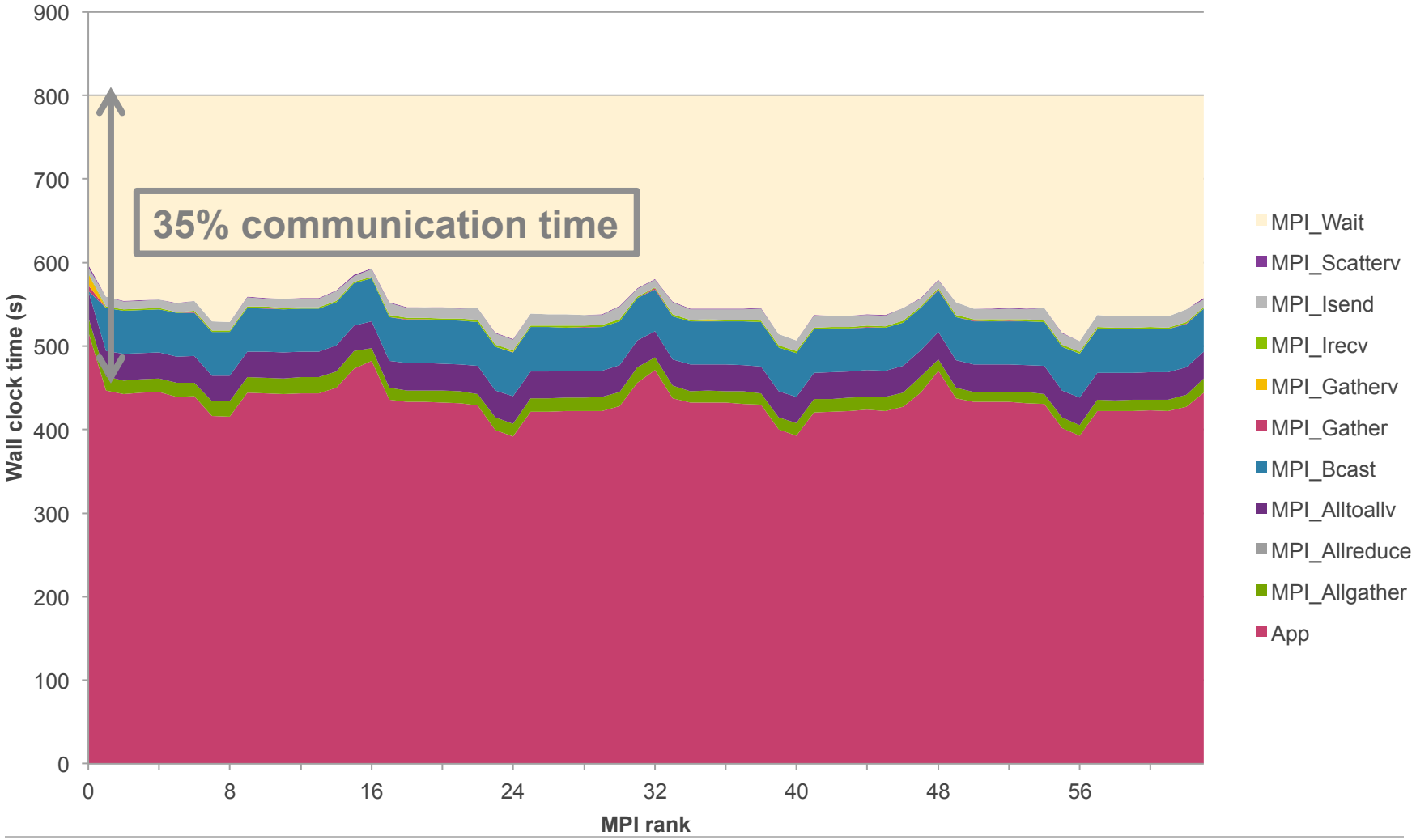
  - Molecular dynamics (GROMACS)

  - Materials science (VASP)

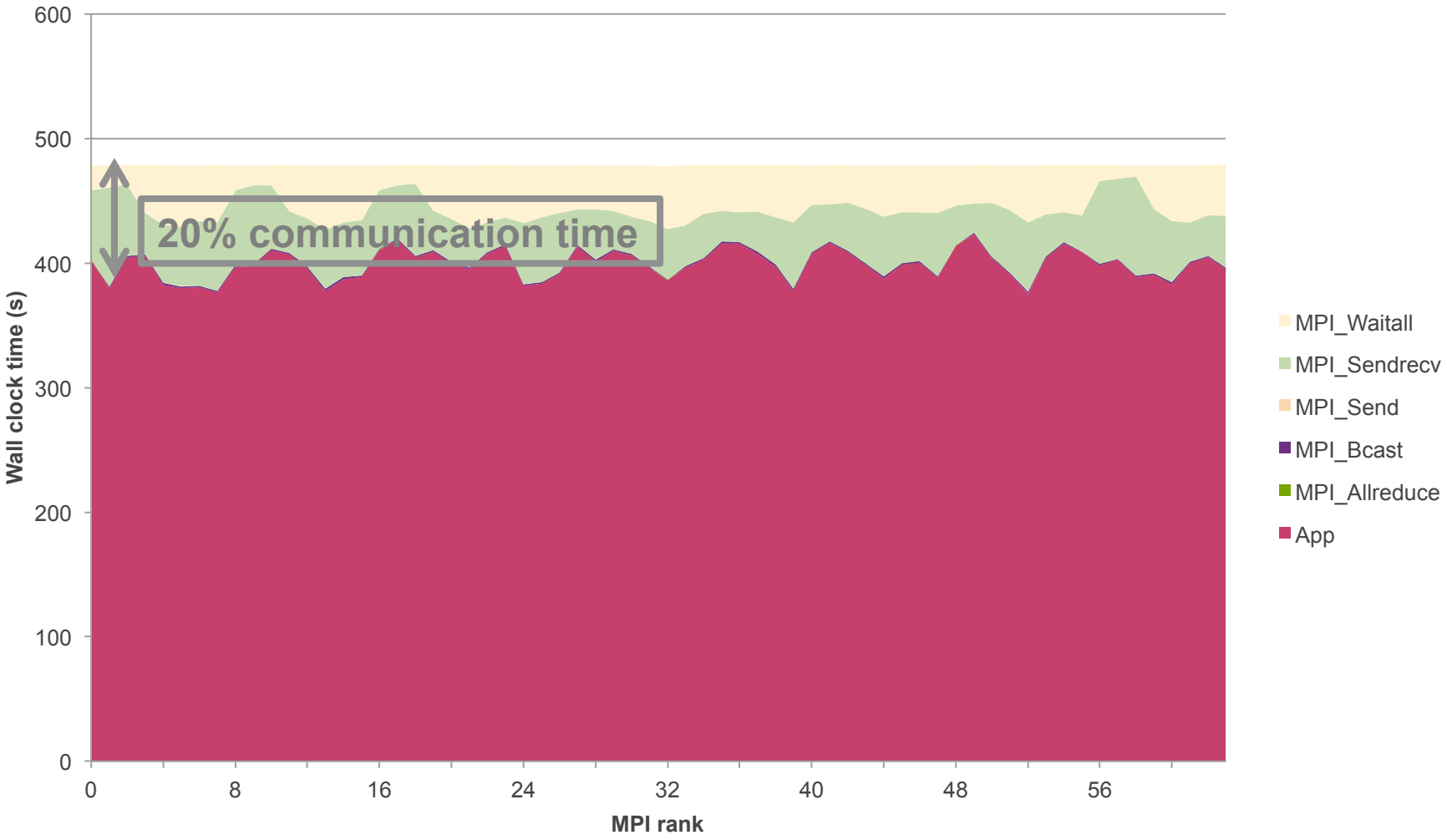
Application characteristics

Application	Segment	Communication Pattern	Notes
WRF	Weather forecasting	Nearest neighbor	Lot of data movement, but all “local”
GROMACS	Molecular dynamics	Point-to-point	Latency sensitive
VASP	Materials science	All-to-all	Lot of data movement and very bandwidth intensive

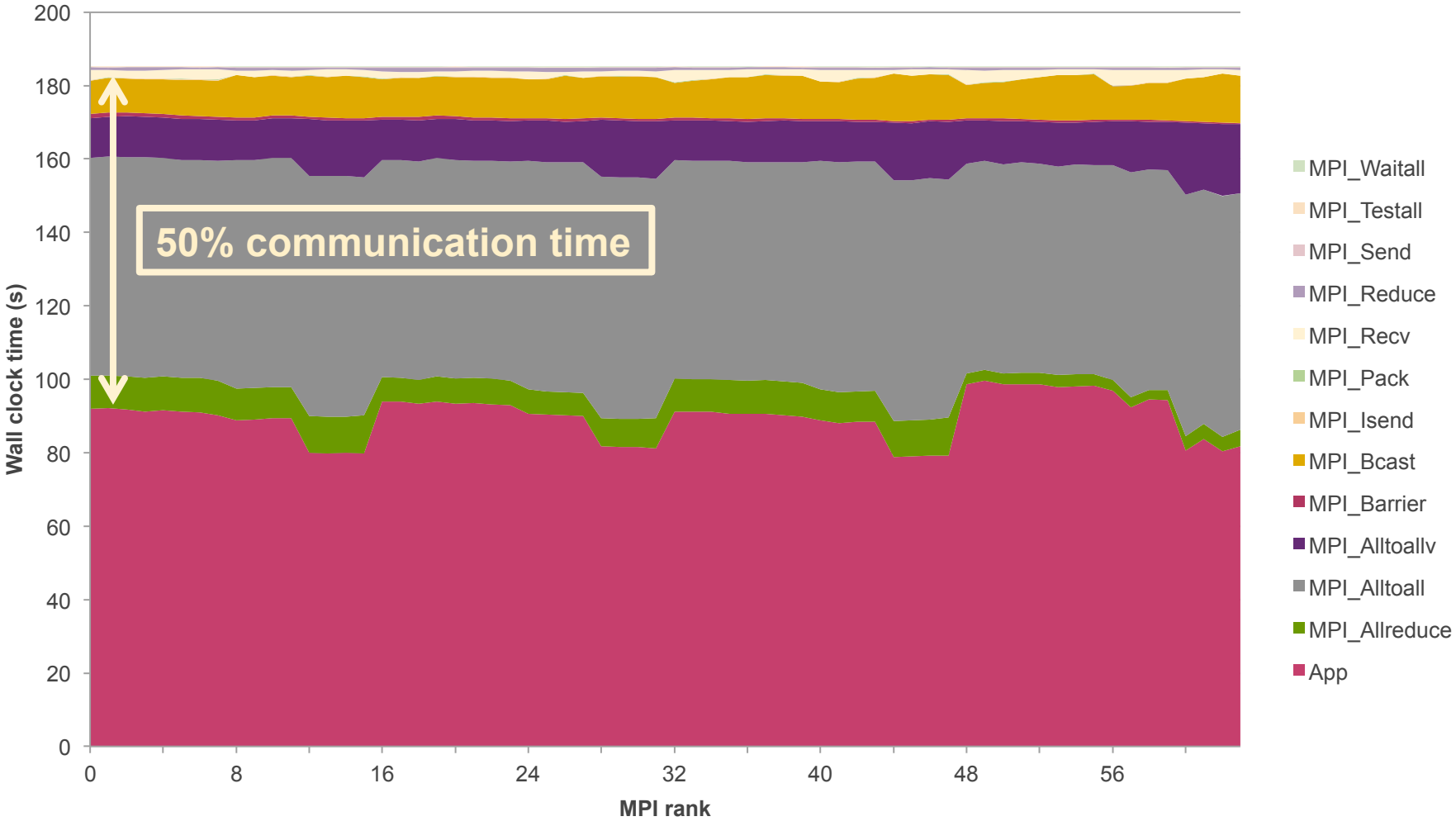
WRF on 64 cores, Infiniband



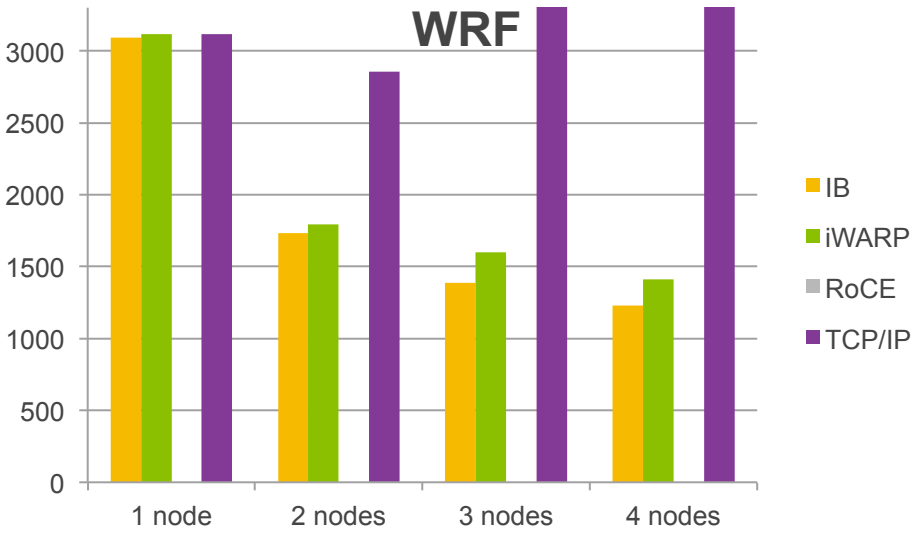
GROMACS on 64 cores, Infiniband

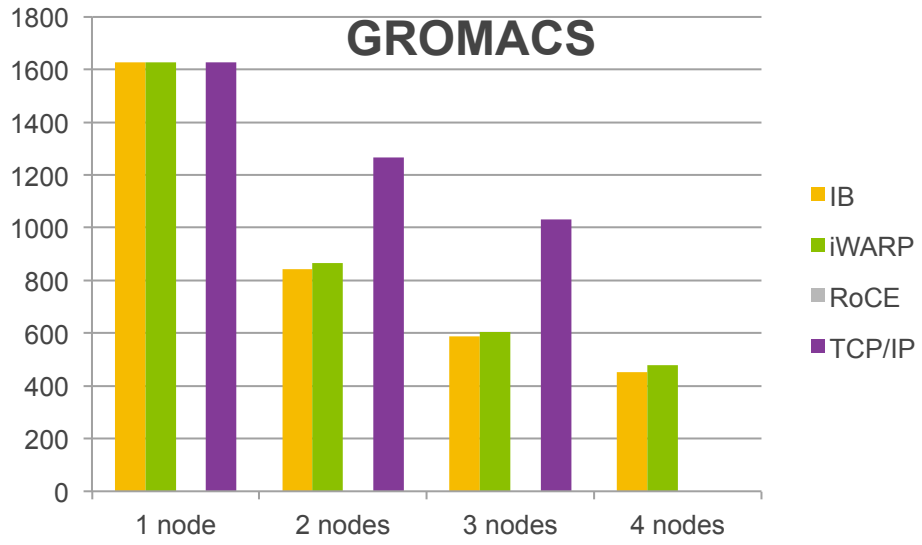


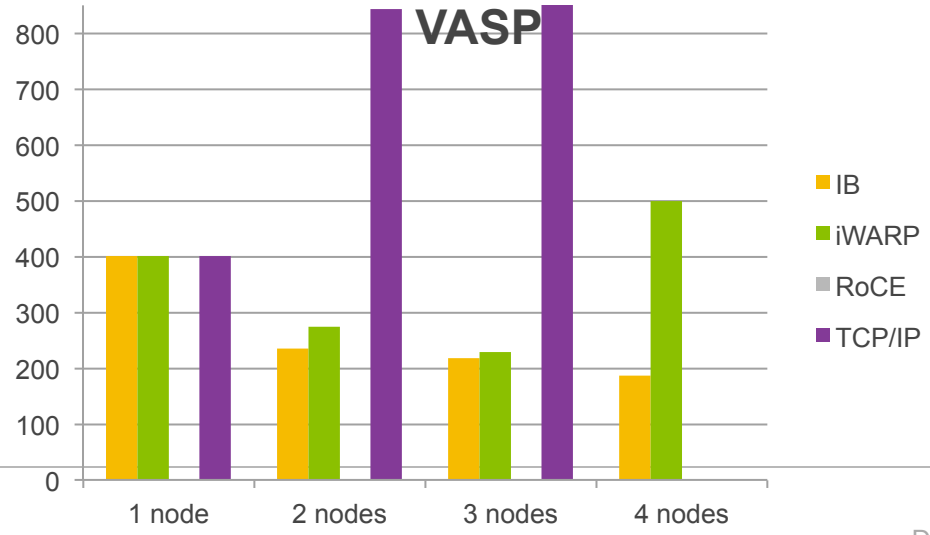
VASP on 64 cores, Infiniband



Results

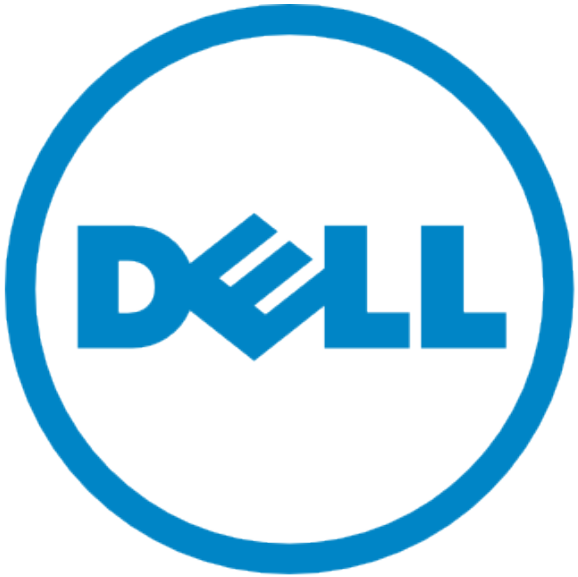

WRF
3000
2500
2000
1500
1000
500
0
1 node
2 nodes
3 nodes
4 nodes
IB
iWARP
RoCE
TCP/IP


GROMACS
1800
1600
1400
1200
1000
800
600
400
200
0
1 node
2 nodes
3 nodes
4 nodes
IB
iWARP
RoCE
TCP/IP


VASP
800
700
600
500
400
300
200
100
0
1 node
2 nodes
3 nodes
4 nodes
IB
iWARP
RoCE
TCP/IP

Summary

- Infiniband is still the best performing and most versatile networking fabric for HPC workloads
- For some workloads, iWARP can be a viable and cost effective alternative
- Classical Ethernet is not useful for parallel applications
- RDMA protocols is still not of production quality, so your mileage will vary



The power to do more