Current status of Infiniband for HPC

Peter Kjellström – 20041019

Agenda

- Overview

  What is Infiniband, photos...

- Hardware

  Current hardware, future plans, vendors

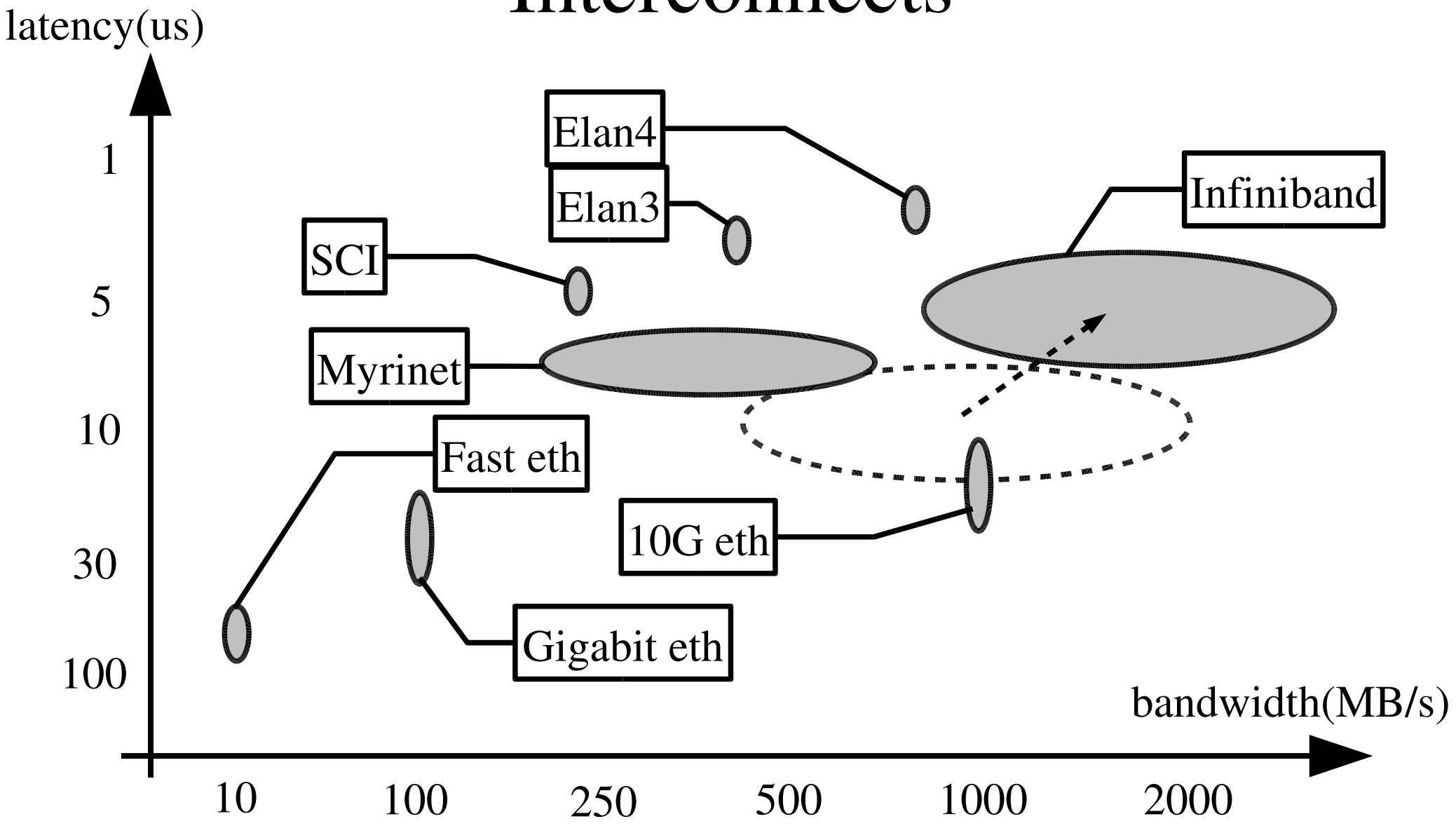
- Software

  Drivers, MPI, etc.

- Our experiences

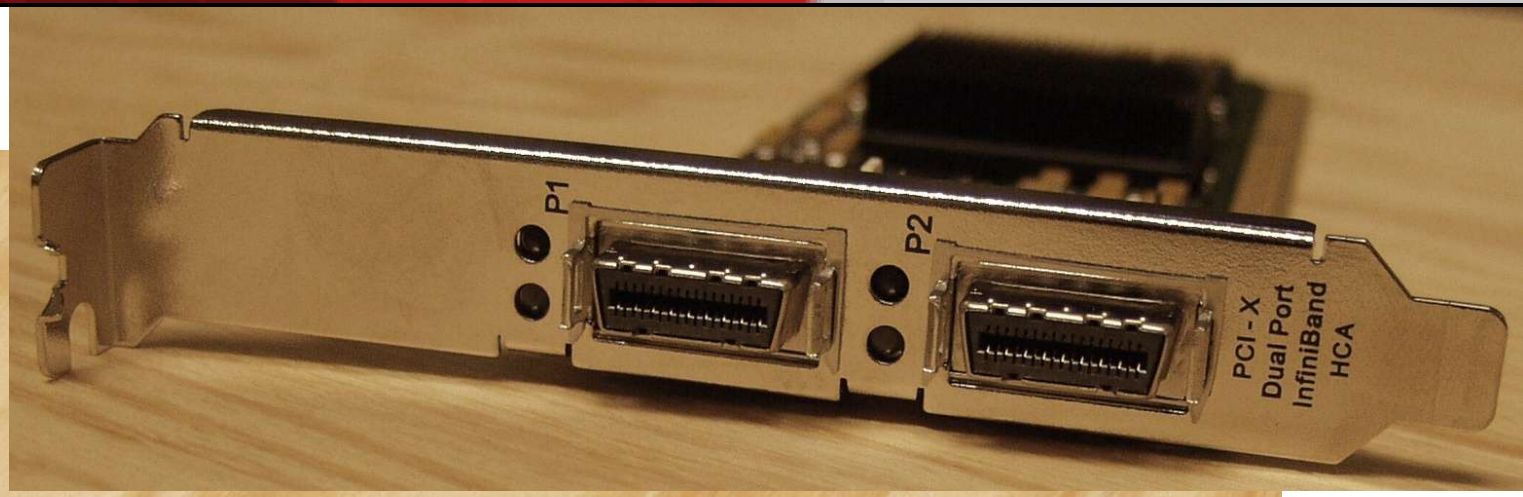
- Live demo stuff

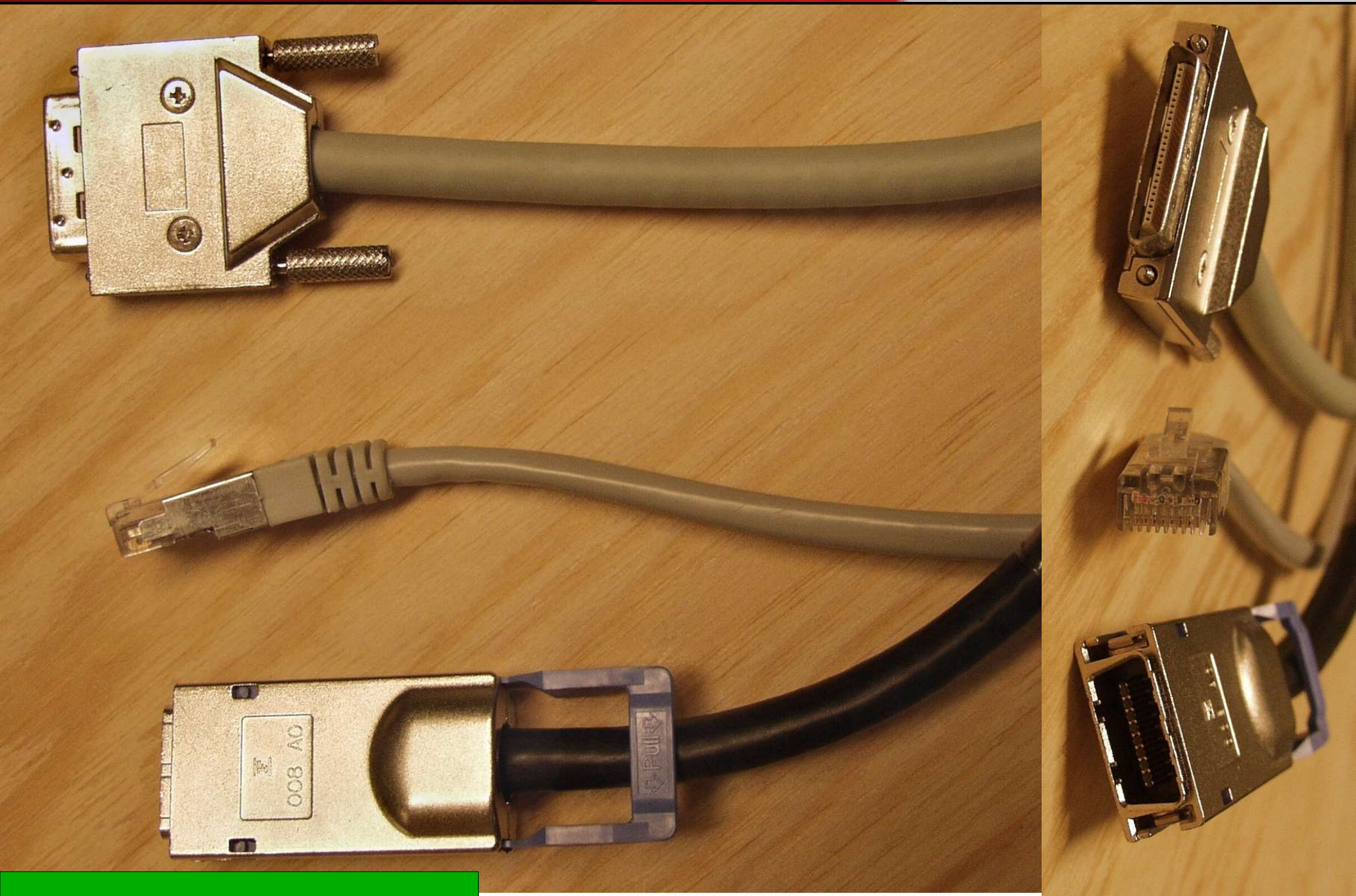
Interconnects

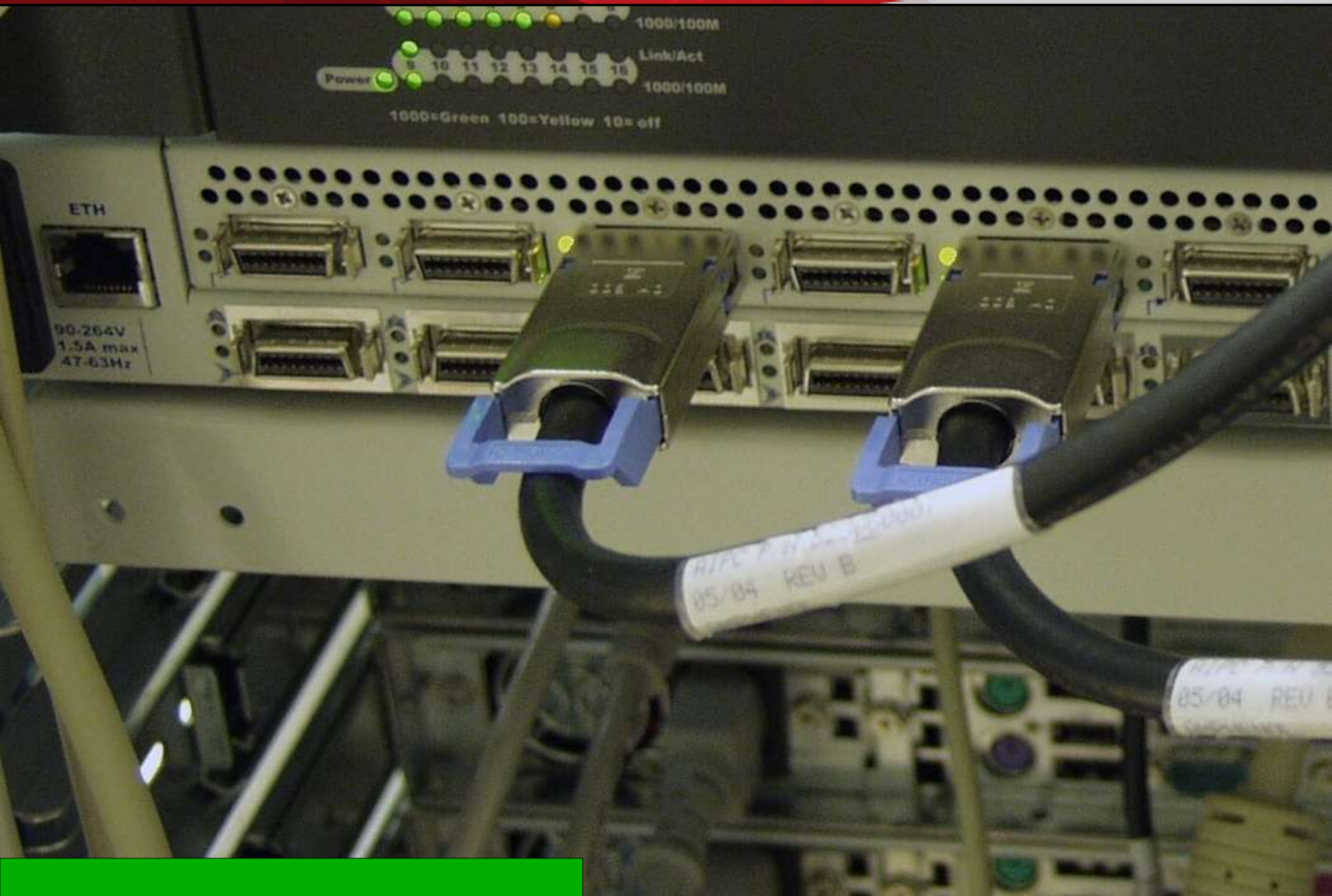


Infiniband overview

- Standardized switch based multi gigabit interconnect
- Like PCI-express it's serial, 1x is today 2.5 Gbps
- The vast majority of IB ports are 4x, 10 Gbps
- DDR is almost here and QDR is on the horizon
- Infiniband targets not only HPC but also storage and enterprise systems

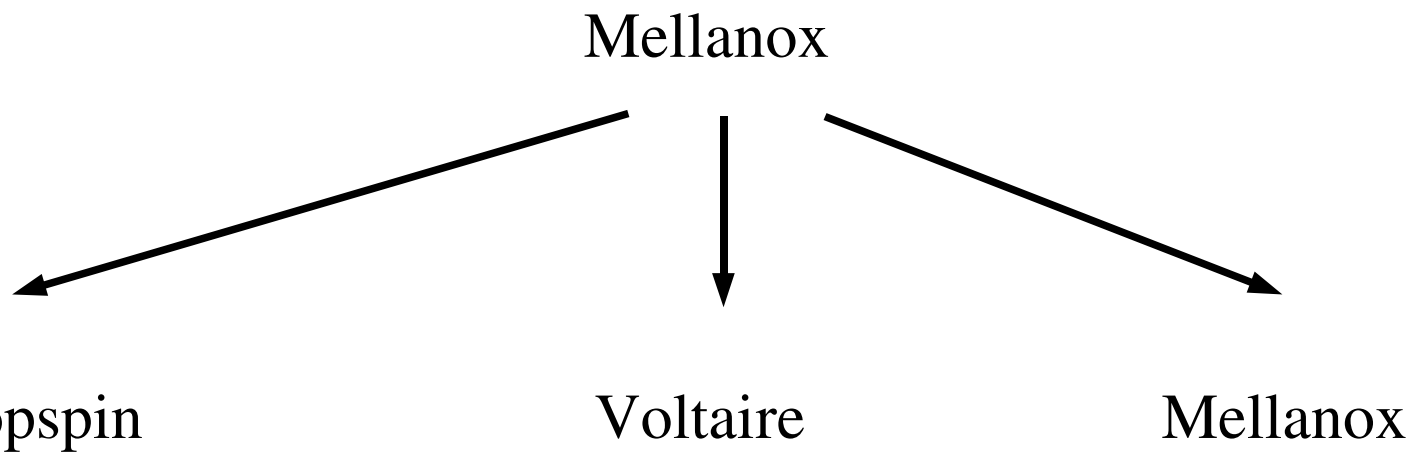

P1
P2
PCI - X
Dual Port
InfiniBand
HCA


008 AO
PULL


1000/100M
Link/Act
1000/100M
9 10 11 12 13 14 15 16
Power
1000=Green 100=Yellow 10= off
ETH
90-264V
1.5A max
47-63Hz

Who makes it, who sells it?

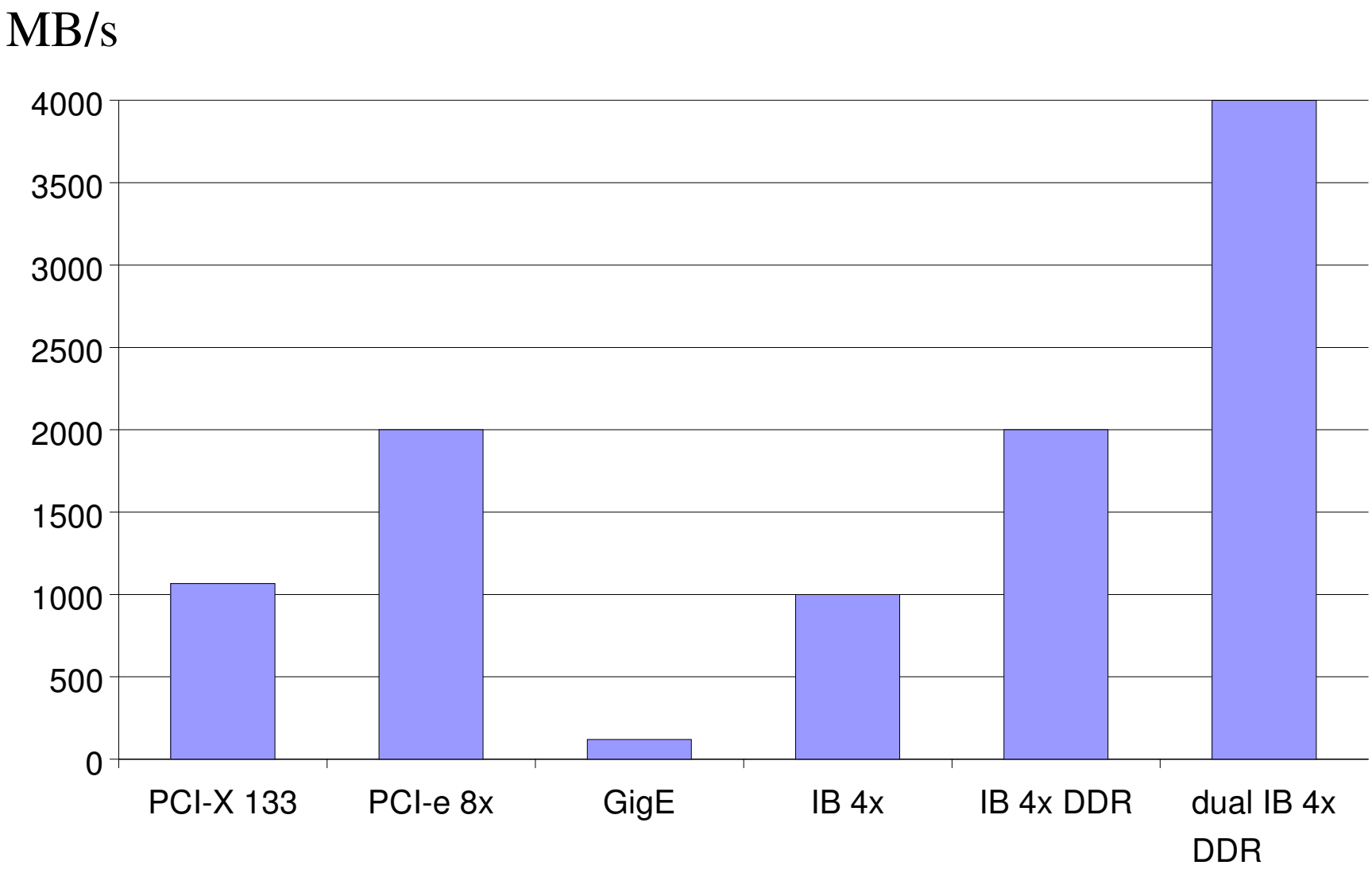
- Currently pretty much only one chip maker, Mellanox
- Several makers of HCAs and switches, Topspin, Voltair, Mellanox...



Hardware products

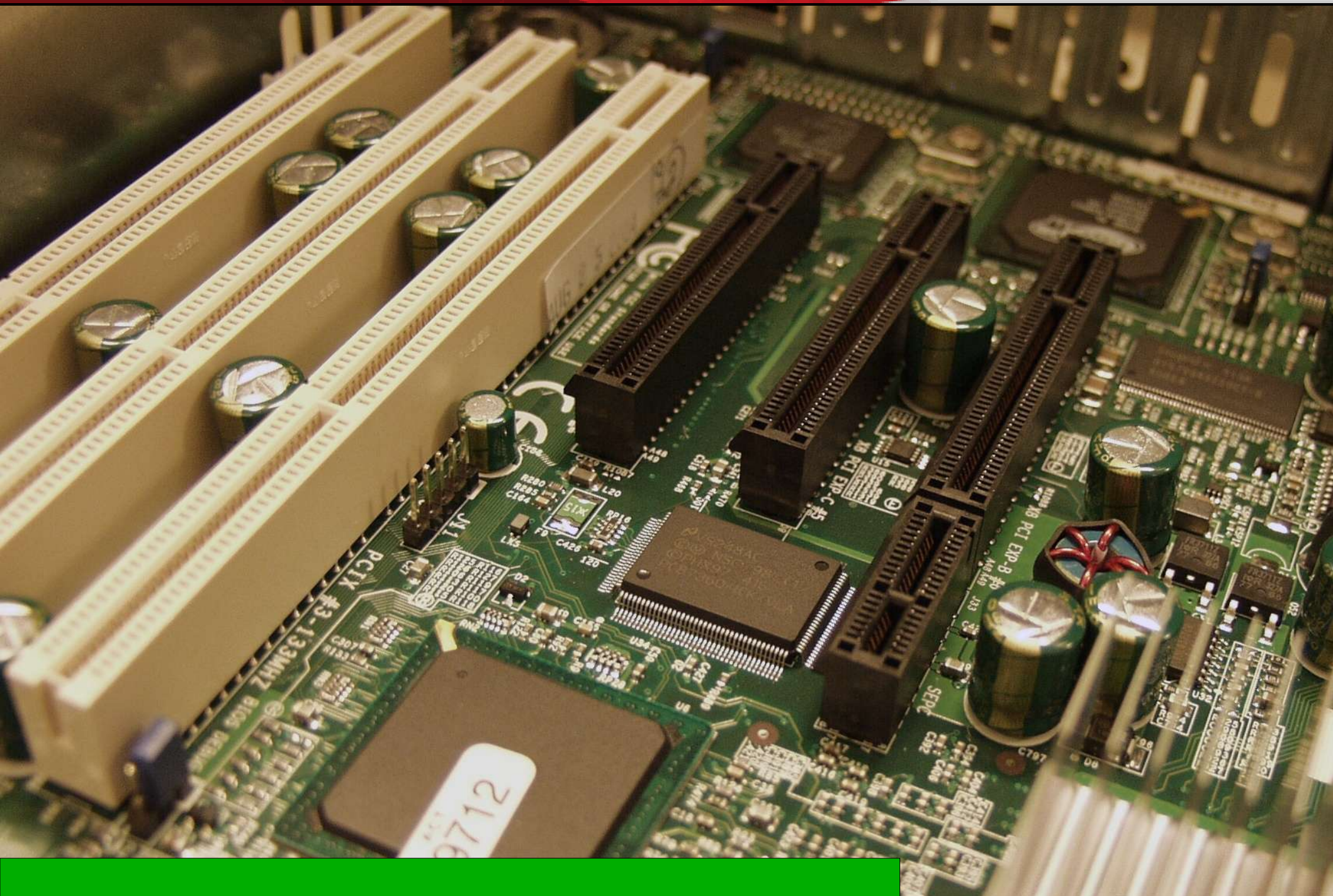
- Infiniscale III based
- Switches from 24 ports to 288 ports
- HCAs for PCI-X and PCI-express (low prof, onboard...)
- Switchports for both copper and fiber
- Some other stuff, FC to IB modules, etc.

Interconnect vs system bus



PCI-express

- Latency drops from good PCI-X @5us to 3.5us
- 8x PCI-e gives 2 GB/s both directions


PCIX #3-133MHZ BUS
J11
SEP2
C707

Software

- Driver stack from OpenIB.org

  Lots of ongoing development, everybody uses it in one form or another

- Filesystems

  PVFS2 has support for native IB, Luster will have support, NFS over rdma

- MPI

  MVAPICH, MPICH-VMI, LAM, Scali MPI-connect, others...

- Other supported protocols

  Native IP support, SDP (Socket Direct Protocol), SRP (SCSI Rdma Protocol)

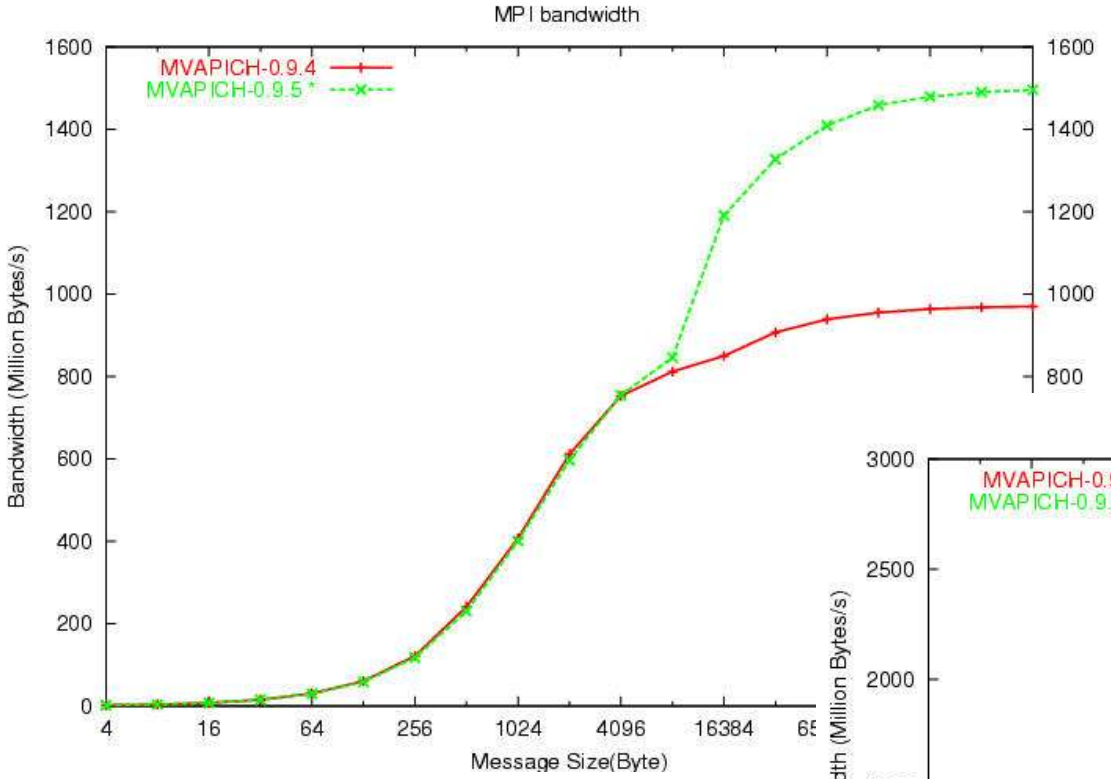
Current IB testing at NSC

- Infiniband testbed A
  - 4+1 3.2 GHz “prescott” P4, PCI-X, 2G RAM/cpu
  - E7210 chipset with 533 MB/s PCI-X (266 MB/s HA
- Infiniband testbed B
  - 2 dual 3.2 GHz “Nocona” Xeon, PCI-Express, 2G RAM/cpu
- New HCA cards, PCI-express
- More OpenIB testing, more LAM
- Some storage testing on the way

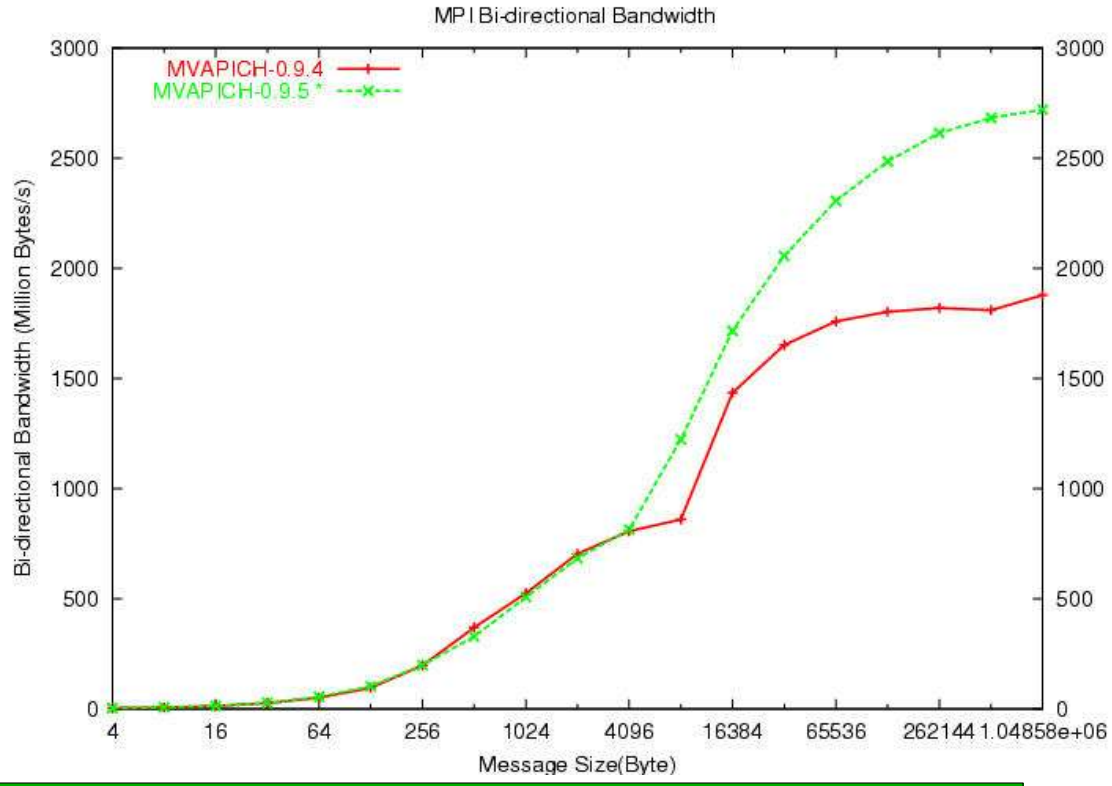
Our experiences

- Lots of MPI implementations to choose from
- LAM runs ok (7.1, 7.1.2b), Scali MPI-connect fine
- Lower latency than I expected from the beginning (~5.5 us on PCI-X)
- Both software and hardware has been very stable for us
- Nocona PCI-X was not fast without tuning...

The very latest and greatest IB performance numbers



MVAPICH dual vs single port



Future directions

- Proven in HPC environments
- DDR, QDR
- Storage products, moving on from HPC
- In-band functionality, boot, IPMI, console, management
- IB on motherboards
- Linux kernel adoption (hopefully)
- More vendors, more volume

Live demo

- Scali MPI-connect, LAM functionality
- VAPI level performance on Nocona PCI-X