InfiniBand – Experiences at Forschungszentrum Karlsruhe

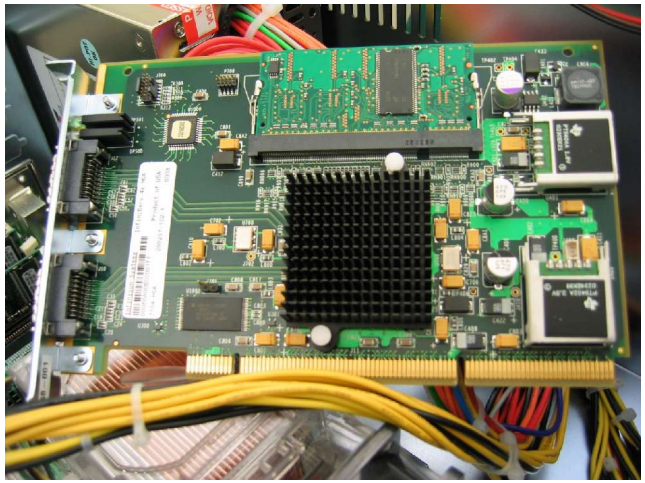
A. Heiss, U. Schwickerath

Credits: Inge Bischoff-Gauss
Marc García Martí
Bruno Hoeft
Carsten Urbach

- InfiniBand-Overview
- Hardware setup at IWR
- HPC applications:
  - MPI performance
  - lattice QCD
  - LM
- HTC applications
  - rfio
  - xrootd

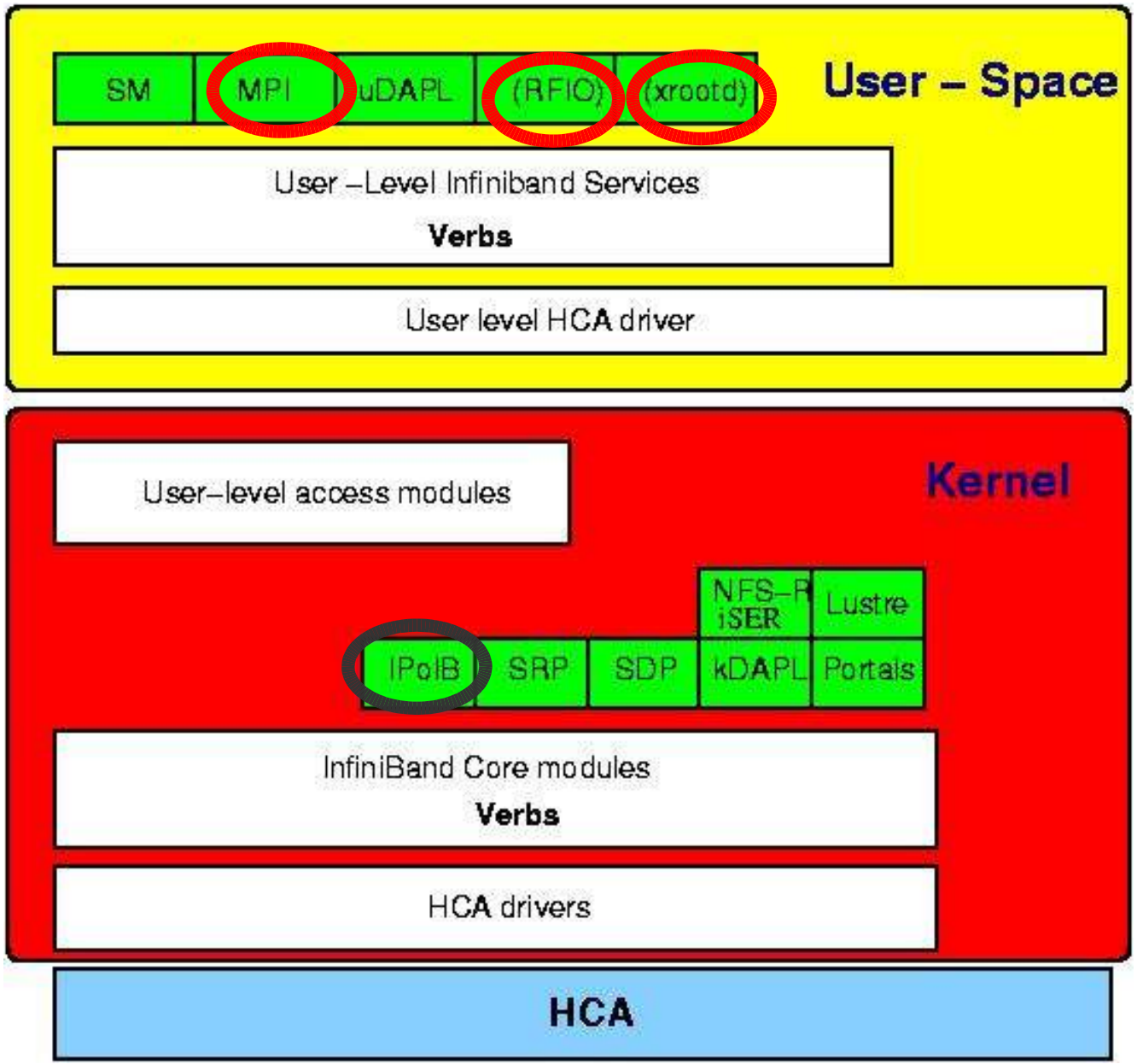
InfiniBand – Overview

- Channel-based, serial, switched fabric providing 2.5, 10 or 30 Gb/s bidirectional bandwidth. 1, 4 or 12 wire pairs carrying voltage differential signals per direction (1X, 4X, 12X).
- Usable bandwidth is 80% of signal rate: 250 MB/s, 1 GB/s, 3 GB/s. (soon: DDR)
- Copper cables (up to 15m) or fibre optics.
- Host Channel Adapters (HCAs) provide up to two ports each: redundant connections possible.



- HCAs for PCI-X (64bit, 133MHz) and PCI-Express.
- Onboard chips expected soon

Software overview



Notes:

- http://openib.org
- kernel space drivers now ship with 2.6 kernel (since 2.6.11)
- Verbs API implementation can be vendor specific
- RFIO and xrootd prototypes by IWR

InfiniBand Test equipment at IWR

Opteron Cluster

- 16 V20z Dual Opteron, 4GB RAM, InfiniCon IBA drivers, SL303/304, Kernel 2.4.21, PBS, 2.2 GHz (for production purpose)
- 13 V20z Dual Opteron, 4GB RAM Mellanox GOLD, SL303/304, Kernel 2.4.21, LoadL+PBS, AFS, 2.2 GHz
- InfiniCon InfinIO 9100 4x- InfiniBand switch
- Mounted into fully water cooled rack
- Installed and management with the QUATTOR toolkit

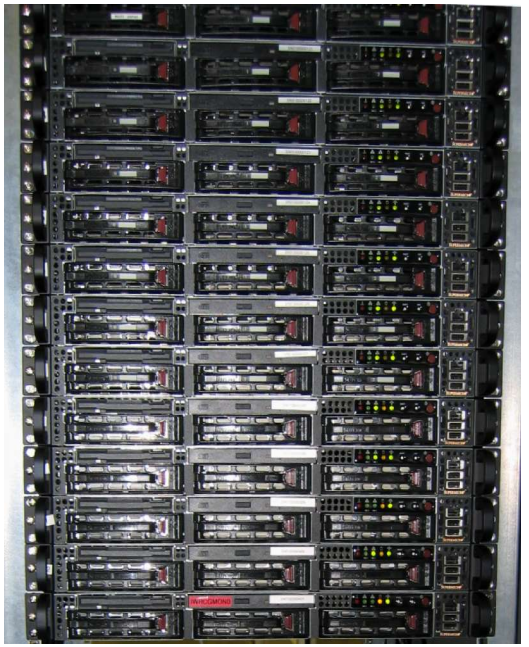
HPL results: 171.4GFlops (26 nodes, 52CPU's)
(75% of theoretical peak performance)

InfiniBand Test equipment at IWR

Xeon Cluster and Blade Center

- 12 Dual Xeon, 2.4 Ghz, 4x-InfiniBand, RH7.3

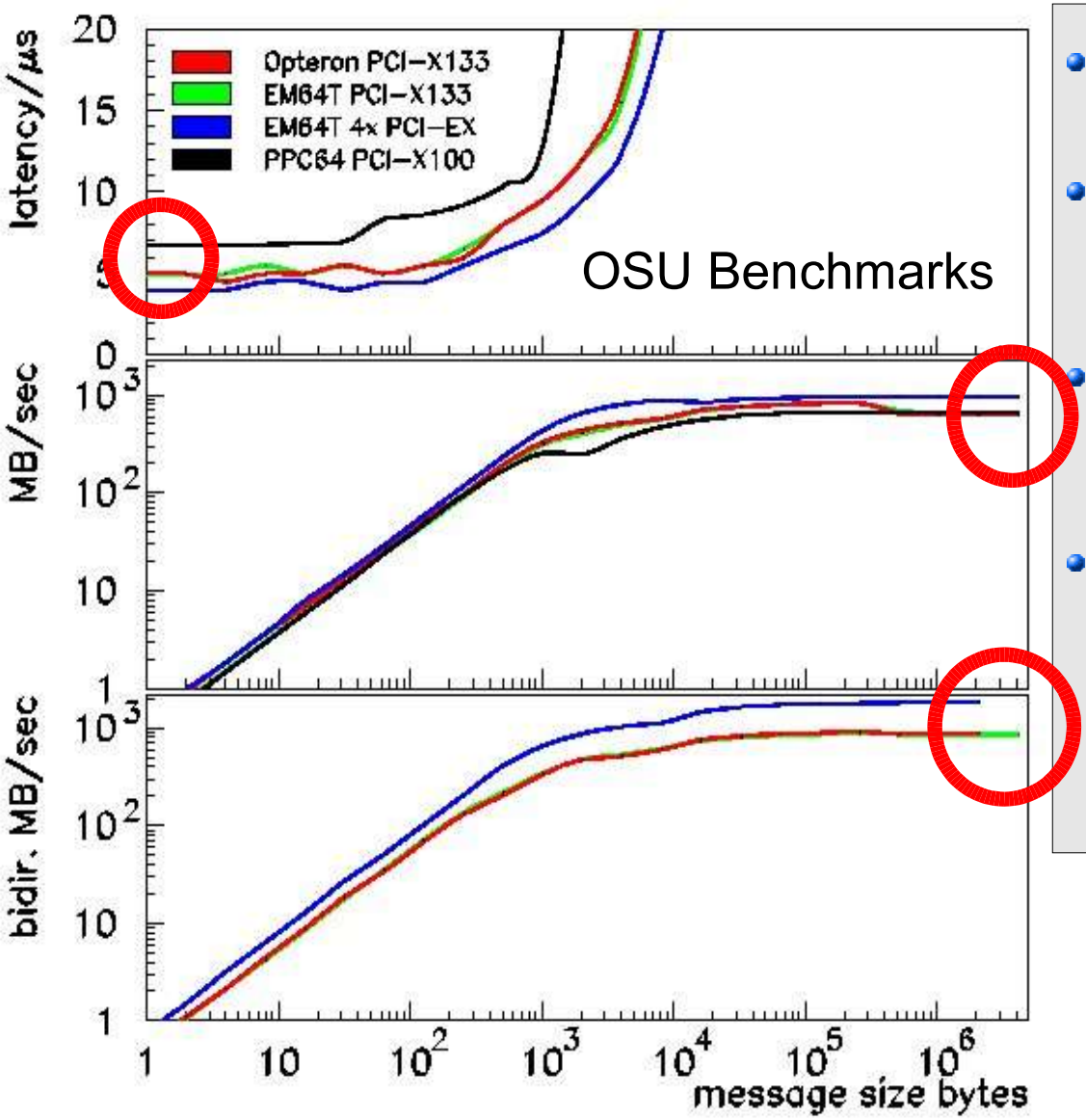
  Kernel 2.4.26, Mellanox drivers suite
- 16 Port 4x Mellanox switch (reference design)
- Rack mounted, air cooled



Temporary equipment used for tests

- HP Xeon64 with 4x PCI-Express and 133MHz PCI-X, 3.4GHz, Dual-CPU, 4GB RAM
- NEC Quad-Opteron, 16GB RAM, 133MHz PCI-X
- IBM JS20 PPC64 blades with 4x-InfiniBand daughter card at 100MHz speed. Not an official IBM product but technology prototype, kindly provided by IBM/Böblingen
- 2 IBM Xeon (2.6GHz) nodes with Intel 10GE ethernet cards

MPI Raw-Performance (64Bit)



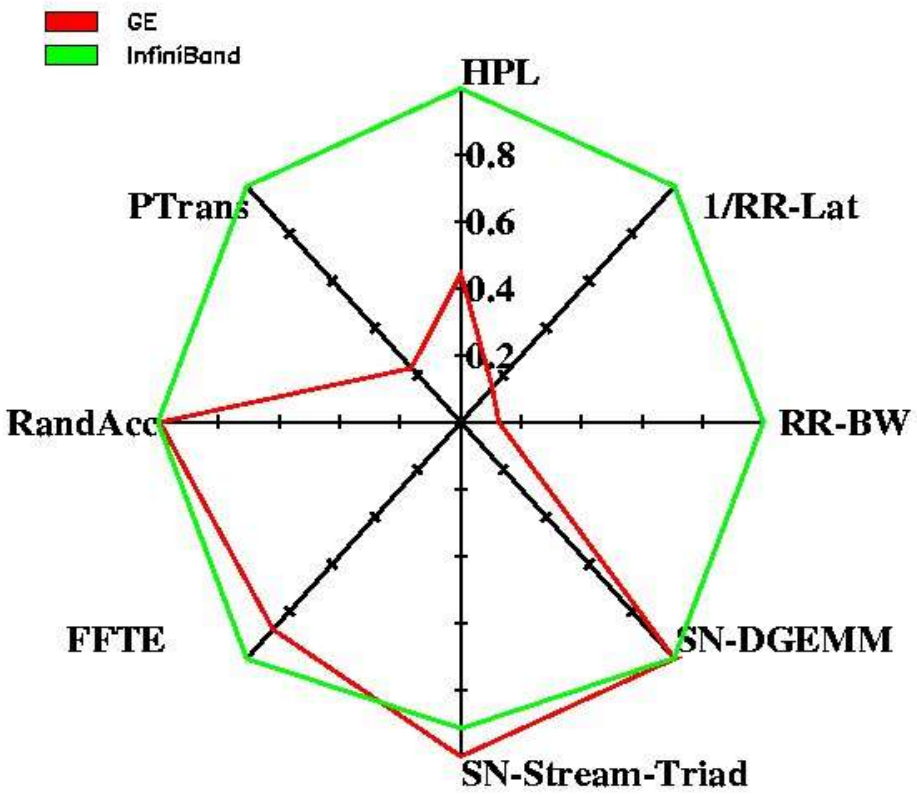
Notes:

- best latency with PCI-Ex (4µs)
- best throughput with PCI-Ex (968MB/s)
- bidirectional BW with PCI-Ex up to 1850MB/s
- JS20 throughput matches experiences with Xeon nodes at 100MHz PCI-X speed, but note the better floating performance of the PPC970FX CPU.

Disclaimer on PPC64:
Not an official IBM Product.
Technology Prototype.
(see also slide 5)

HPCC benchmark suite (0.8beta)

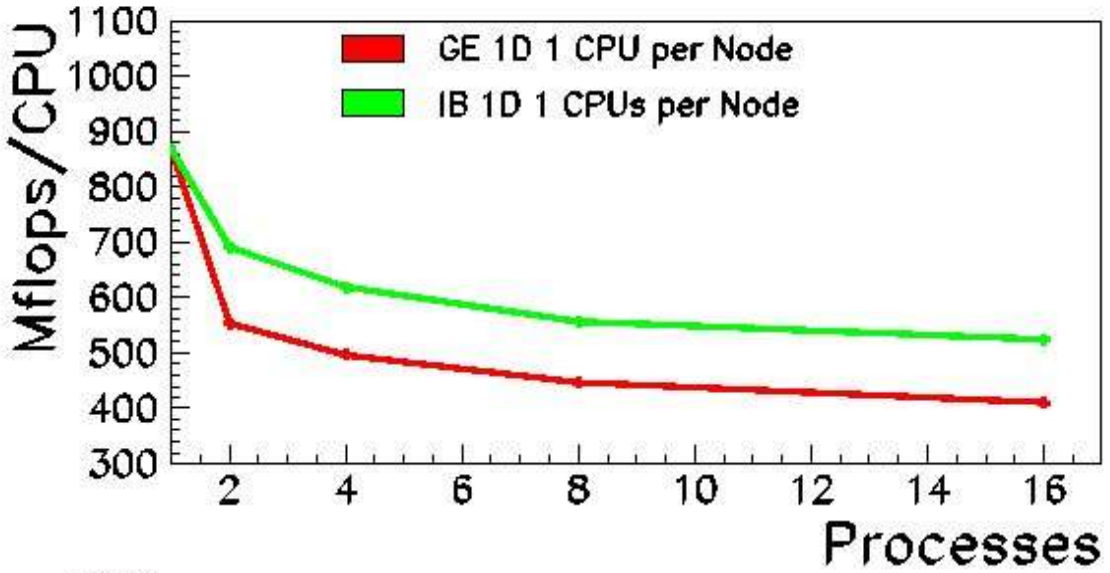
- Comparison GE wrt/ IBA
- GE not tuned, on-board
- Same benchmark parameters
- Same nodes
- 8 nodes, 16 CPUs
- HPL p x q = 4 x 4, N=31208
- NB=40,64,80,96
- HPL 56.46 GFlops (79.5% of peak)

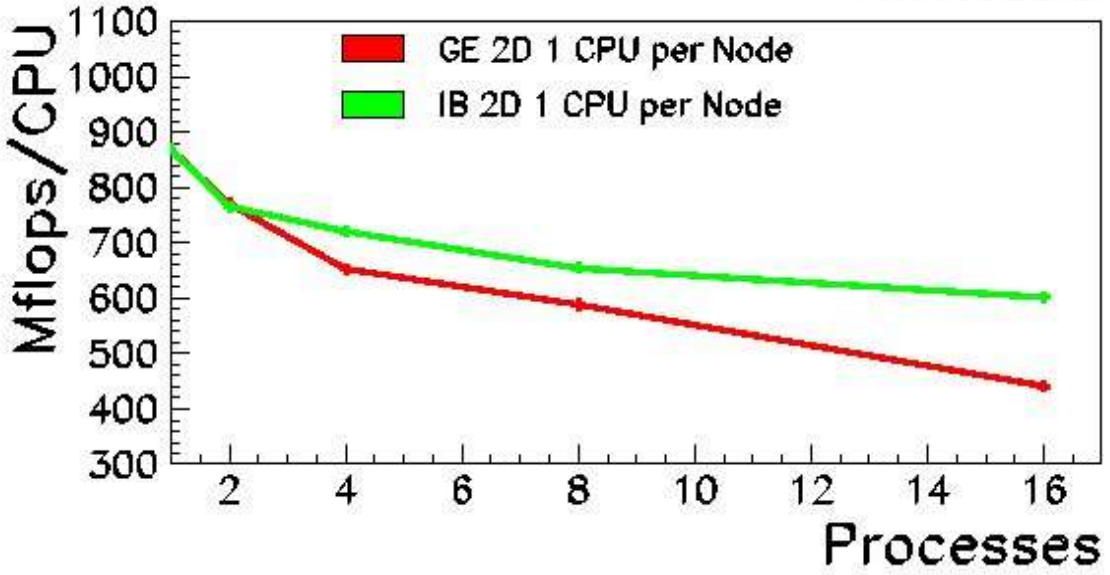


http://icl.cs.utk.edu/hpcc/

Lattice QCD Benchmark GE wrt/ InfiniBand

- Memory and communi-cation intensive application
- Benchmark by C. Urbach
- See also CHEP04 talk given by A. Heiss





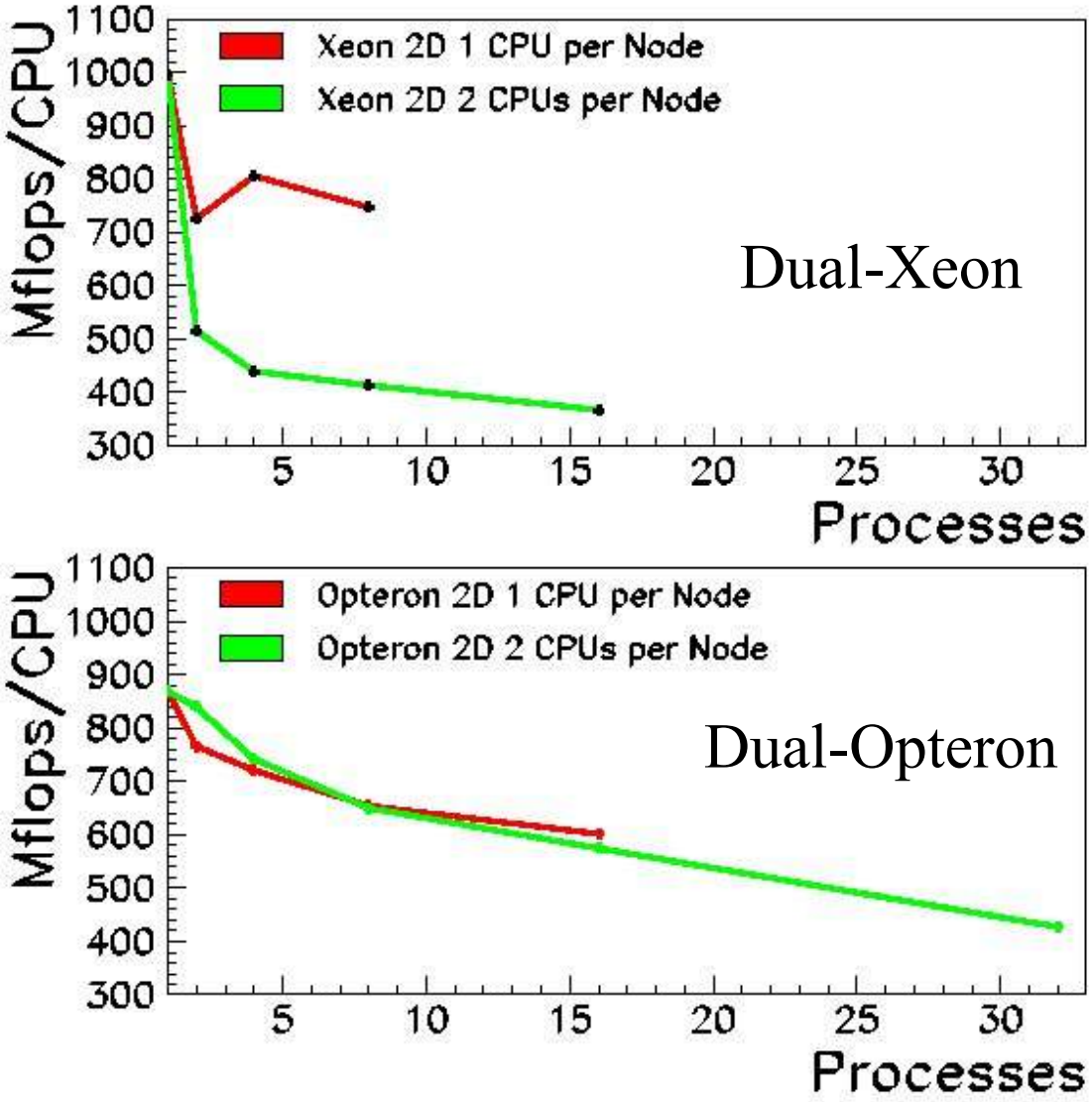
Significant speedup by using InfiniBand

Thanks to Carsten Urbach FU Berlin and DESY Zeuthen

Lattice QCD Benchmark Xeon wrt/ Opteron

CampusGrid

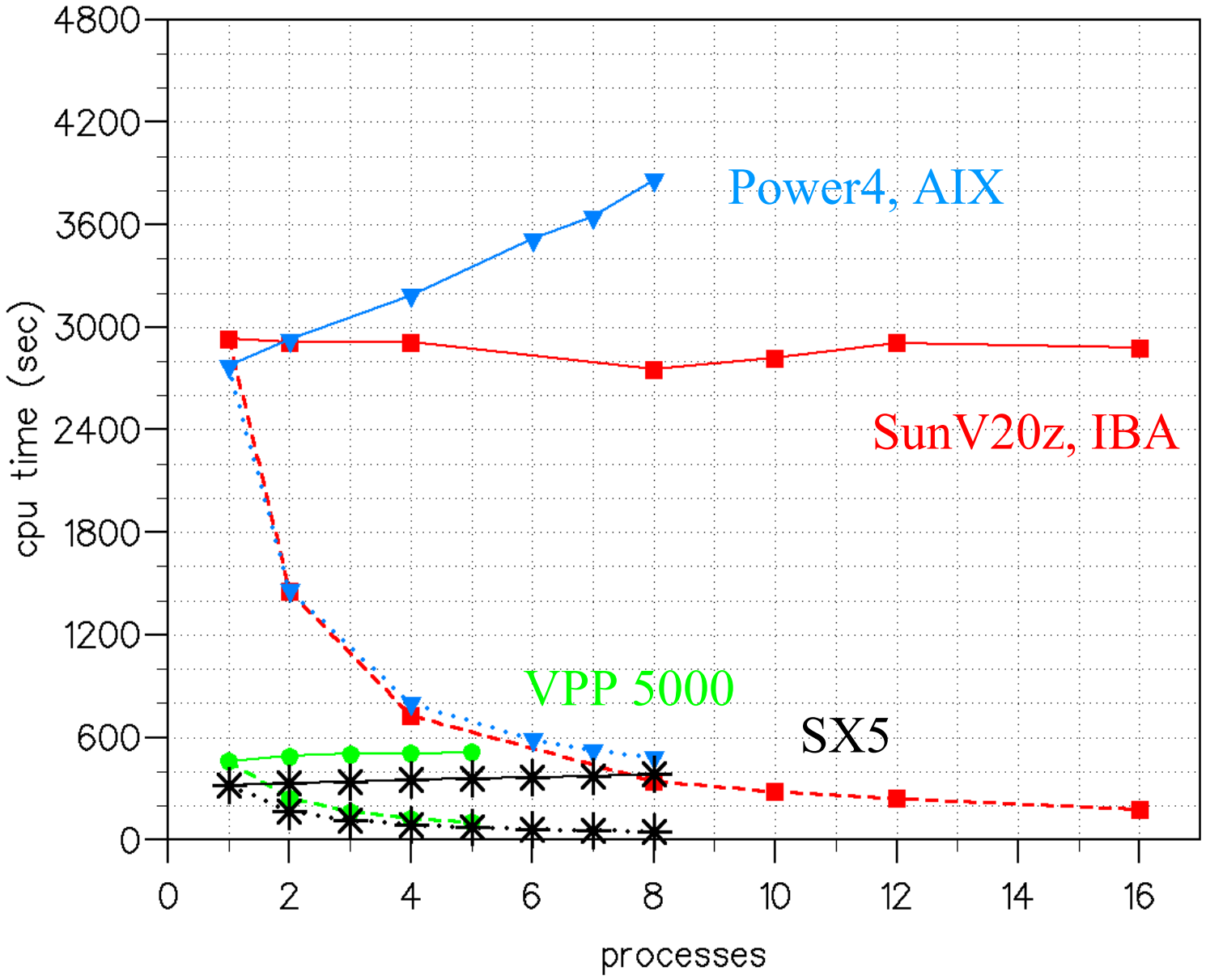
Comparison Xeon with Opteron using **one or two** CPU's



- Opteron: network as good as SMP
- Speed up drop on Xeons when using both CPU's
- Effect not visible on Opterons
- Possible reason:
  Memory bottle neck
  by Northbridge on Xeon

All measurements done at IWR
Thanks to Carsten Urbach
FU Berlin and DESY Zeuthen

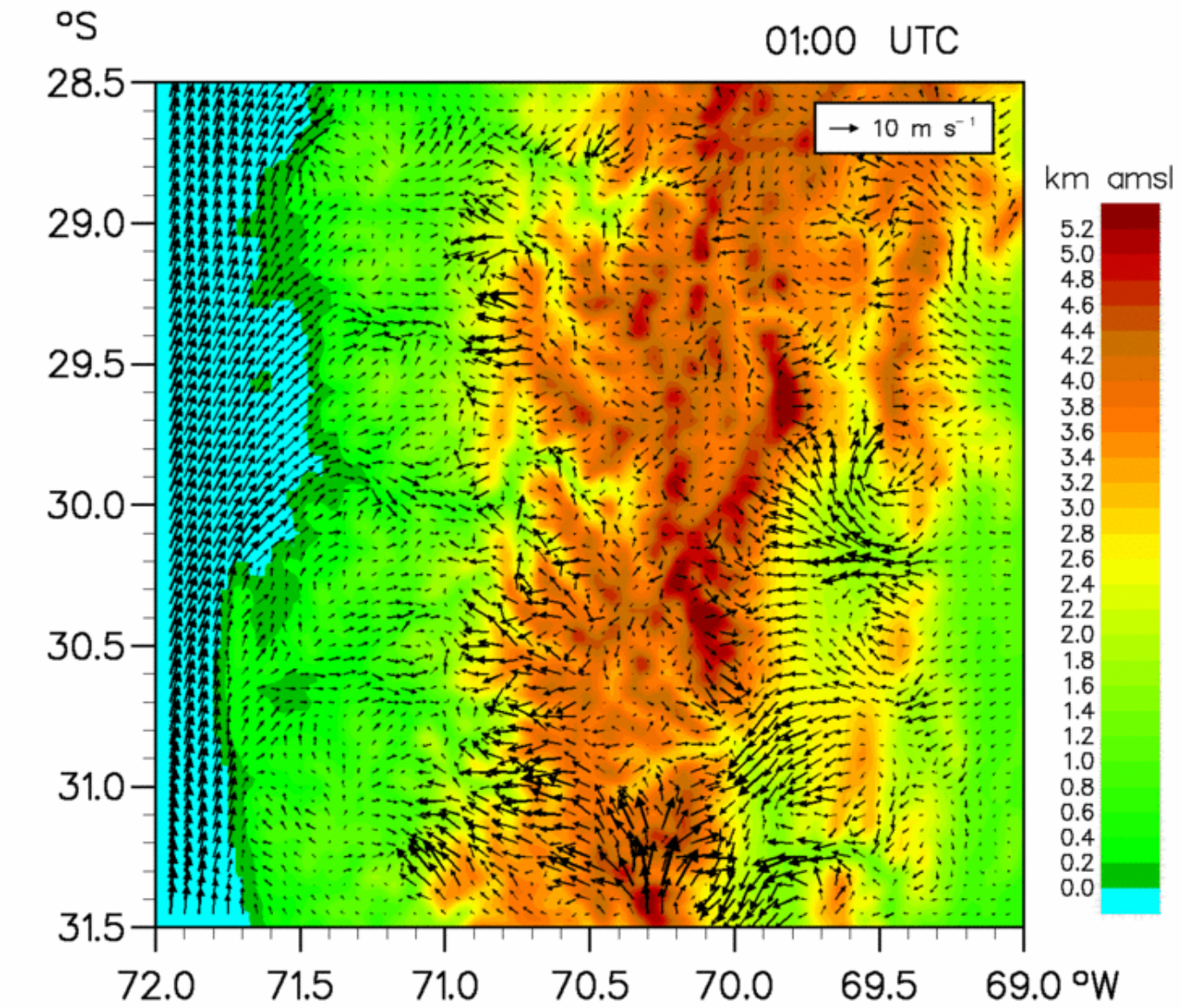
The Local Model (LM) of Deutscher Wetterdienst



- surface wind simulation
- grid size 2.8km
- Chile, 28.03.2000
- 241 x 261 grid points
- 1h simulation time
- dashed: real time used
- solid: total CPU time
- InfiniBand: V20z
  - NCSA MPI
  - Mellanox Gold

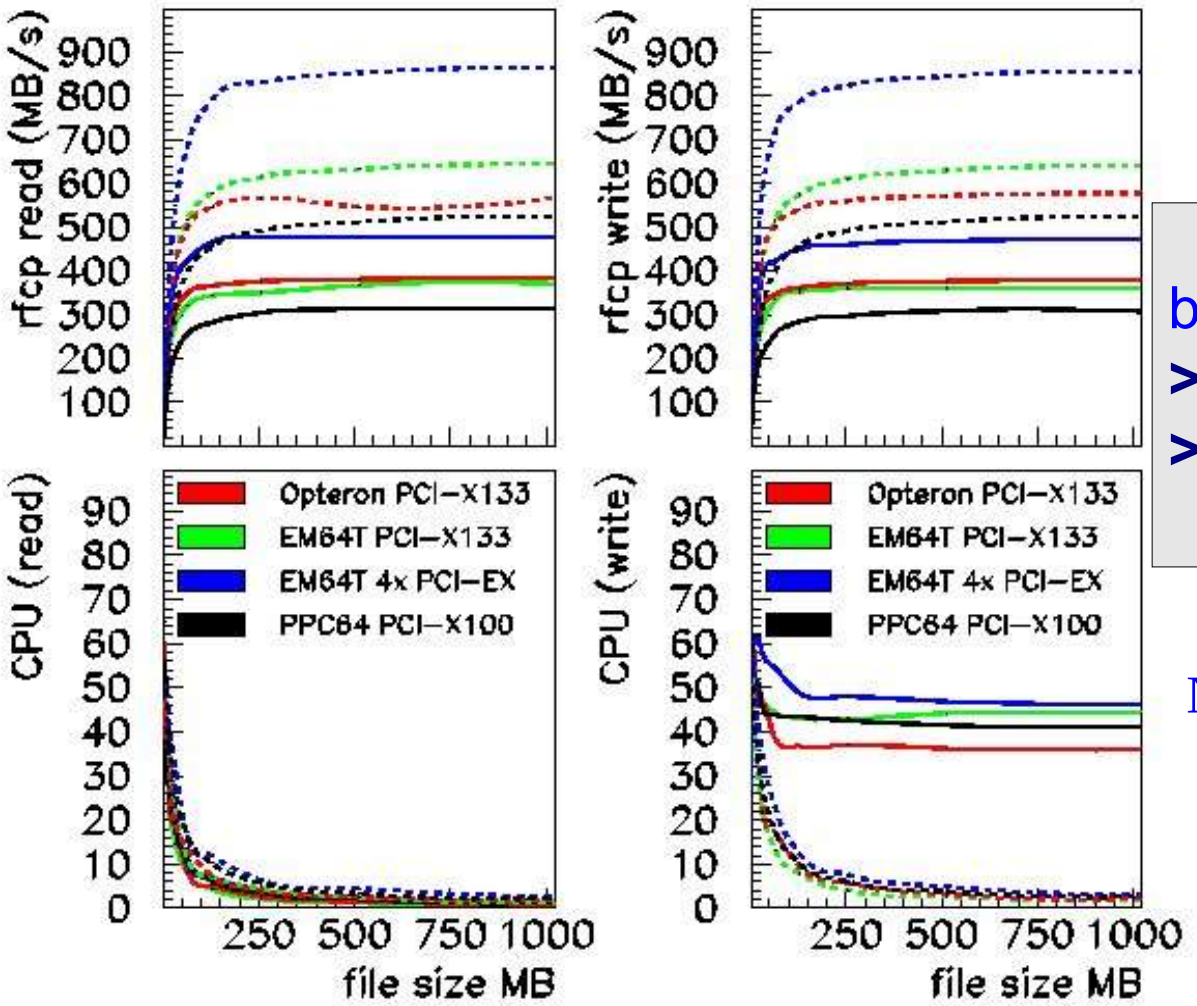
Measurement done by Dr. I. Bischoff-Gauss

The Local Model (LM): 1 day simulation result



RFIO/IB Point-to-Point file transfers (64bit)

PCI-X and PCI-Express throughput



solid: file transfers cache->/dev/null
dashed: network+protocol only

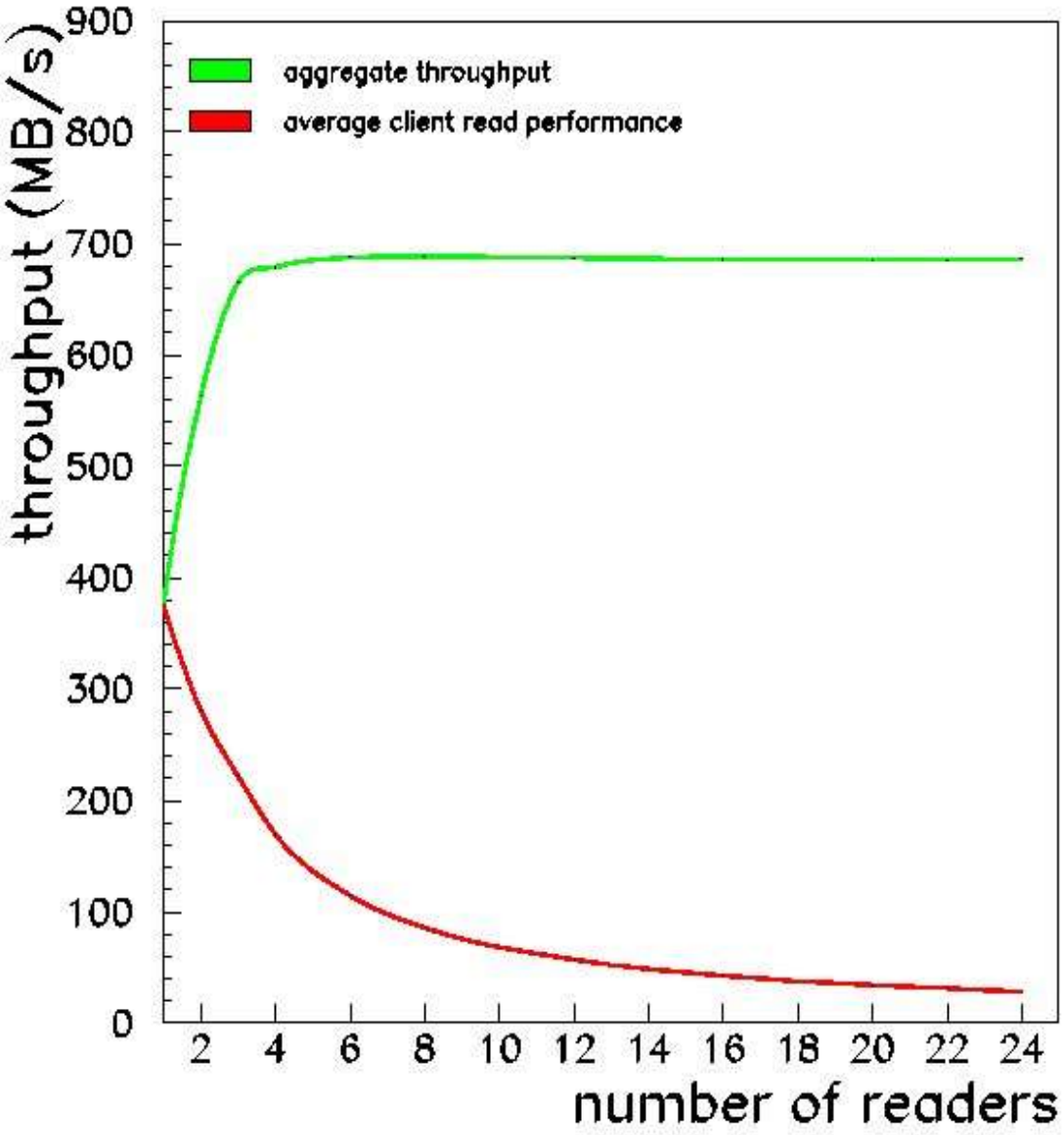
Notes

best results with PCI-Express:
> 800MB/s raw transfer speed
> 400MB/s file transfer speed

RFIO/IB see ACAT03
NIM A 534(2004) 130-134

Disclaimer on PPC64:
Not an official IBM Product.
Technology Prototype.
(see also slide 5 and 6)

RFIO/IB throughput (mixed setup)



Notes:

- NEC Quad-Opteron Server SuSE SLES9, Kernel 2.4.21, 16GB RAM, 2.2GHz
- Testfile: 1024MB random data
- Readers: 12 Dual Xeon 2.4GHz RH7.3 based, Kernel 2.4.16
- All readers read the same file at the same time (to /dev/null)
- See also CHEP04 talk by A. Heiss

The Extended Root Daemon

CampusGrid

What is the Xrootd package?

- Toolkit developped by SLAC and INFN (Padova) for easy data access for the BaBar experiment
- File based data access
- Simple, fault-tolerant, flexible security
- Standalone suite with clients and server packages
- Fully (and heavily) multithreaded
- Release version now distributed with the ROOT package

Here:

focus on raw data throughput, using a simple file copy method (xrdcp)

Xrootd and InfiniBand

Xrootd on **native InfiniBand**

Challenges to be addressed:

- Queue Pairs instead of sockets
- Memory management challenges
  - Use of RDMA requires the buffers to be known to the sender in advance
  - Send method requires preposted receive requests
- Xrdcp does not destroy it's physical connections before exit

Features and status of prototype:

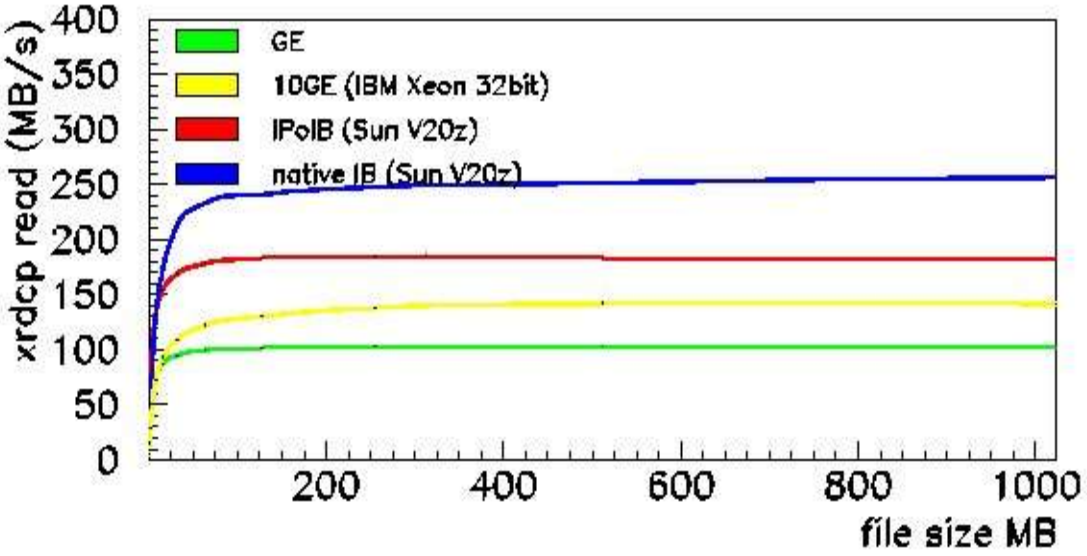
- Makes use of IB_SEND method instead of RDMA
- Allocate private send and receive buffers associated with each QP
  - Last connection times out at end
  - ROOT interface not yet implemented

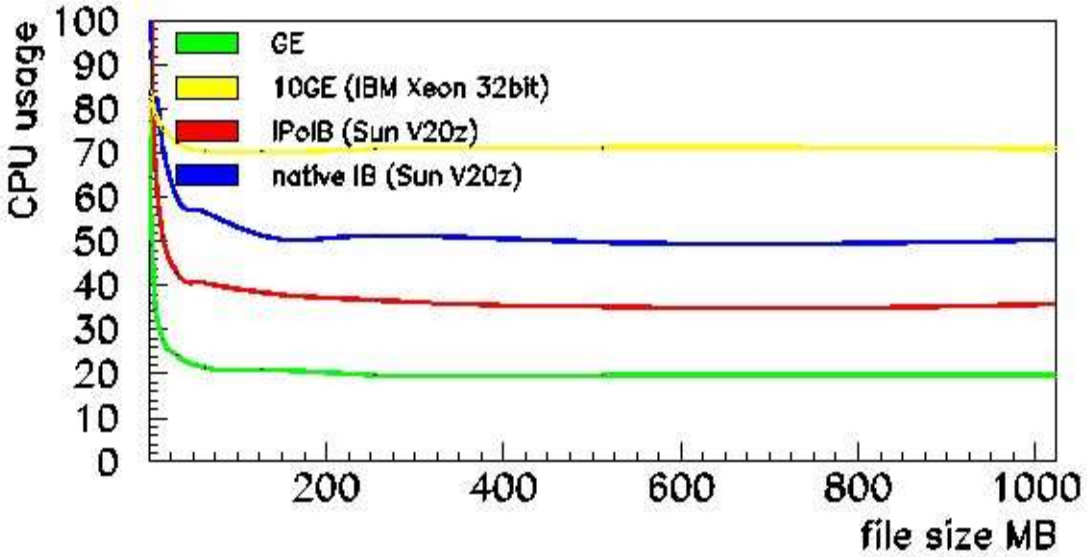
Xrootd and InfiniBand

Notes:

- **IPoIB notes:**
  - Dual Opteron V20z
  - Mellanox Gold drivers
  - SM on InfiniCon 9100
  - same nodes as for GE
- **Native IB notes:**
  - proof of concept version
  - based on Mellanox VAPI
  - using IB_SEND
  - dedicated send/recv buffers
  - same nodes as above
- **10GE notes:**
  - IBM xseries 345 nodes
  - Xeon 32bit, single CPU
  - 1 and 2 GB RAM
  - 2.66GHz clock speed
  - Intel PRO/10GbE LR cards
  - used for long distance tests

First preliminary results





Xrootd and InfiniBand

Outlook/next steps:

- fix known problems
  - memory management
  - client/xrcdp resource cleanup
  - fast connection ending
- implement missing parts
  - integration into ROOT toolkit
- performance enhancements
  - get rid of local buffers
  - maybe implement buffer recycle mechanism
  - allow use of RDMA based transfers
  - requires discussion/interaction with developpers

Summary & Outlook

- InfiniBand offers nice performance for small prices
- usable for both HPC and high throughput applications at the same time
- technology is developping and prices keep falling
- software and drivers are freely available
- see also:

http://www.fzk.de/infiniband