InfiniBand for High Energy Physics

Dr. Andreas Heiss,
Dr. Ulrich Schwickerath
Institute for Scientific Computing (IWR), Forschungszentrum Karlsruhe

Gerardo Ganis
CERN

- Motivation
- InfiniBand – Technology Overview
- Hardware Setup at FZK
- MPI Performance
- RFIO over InfiniBand
- Work in progress: ROOTD/IB

Motivation

The Problem:

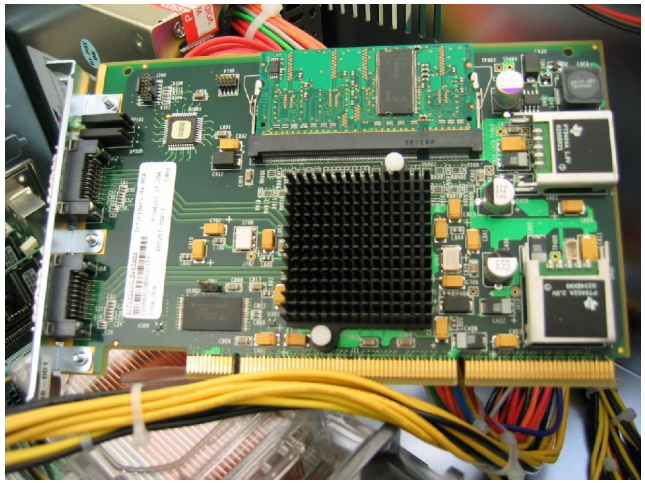
- Computing power has increased faster than network bandwidth.
- Amount of data to be moved and processed for HEP experiments will increase drastically.
- ”Classical” TCP/IP communication uses large portion of CPU power.
- Parallel applications require high bandwidth and low latency.
- Big clusters require a scalable networking technolgy.

The Solution ?

InfiniBand – A fast interconnect technology
with open specifications released end of 2000 by the
InfiniBand Trade Association (IBTA)
http://www.infinibandta.org

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 two ports each: redundant connections possible.

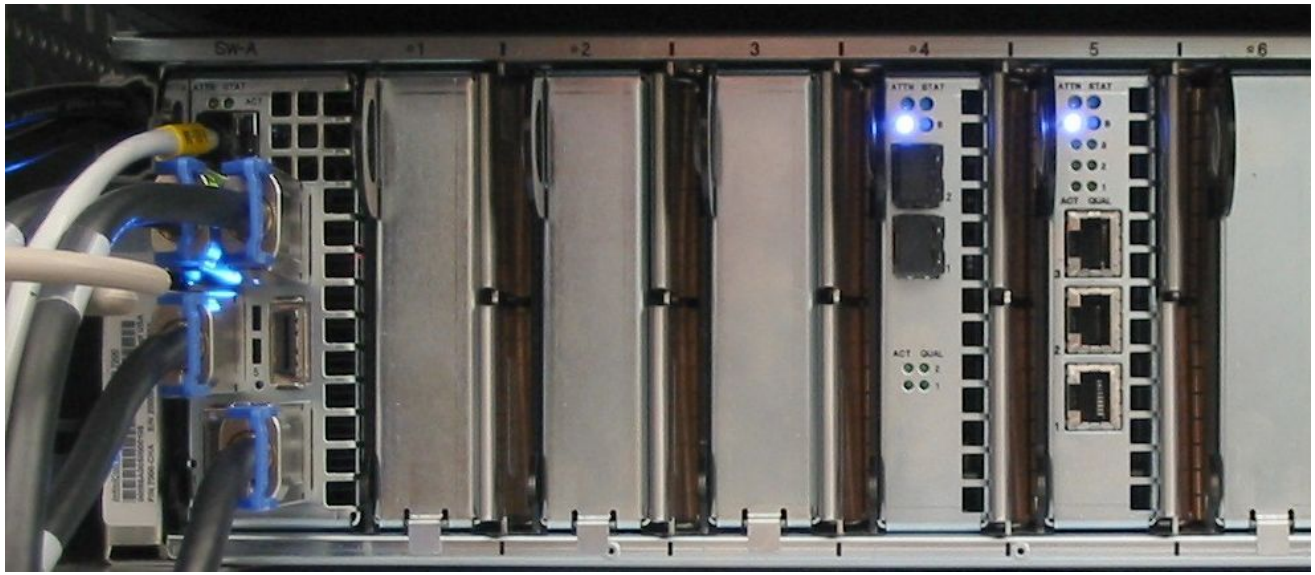


- HCAs for PCI-X (64bit, 133MHz) and PCI-Express.
- Onboard chips available soon.

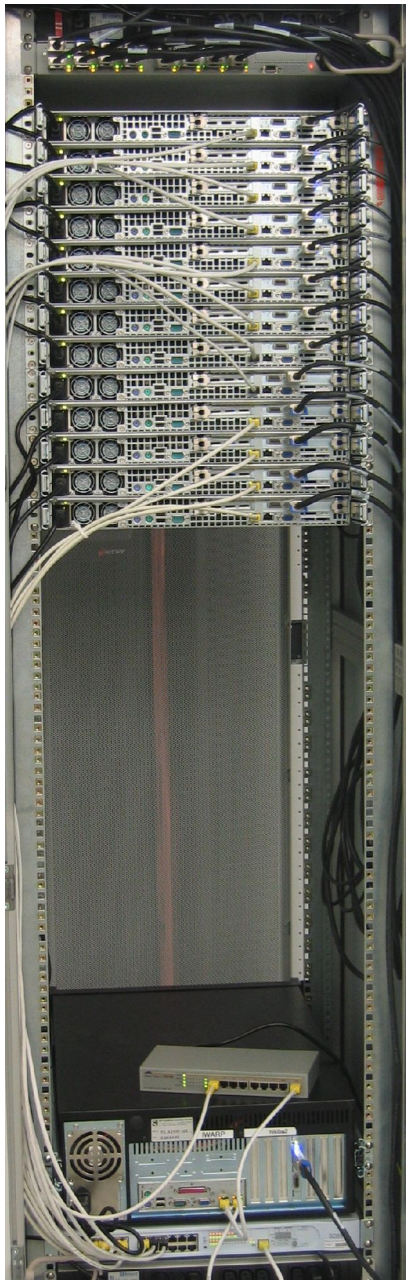
InfiniBand – Overview

- ”Reliable” data transfers: hardware takes care of data integrity.
- Remote Direct Memory Access (RDMA) capabilities.
- Unified fabric for Inter Process Communication (MPI), network (IPoIB, SDP) and storage (SRP).
- Gateway modules for Fibre Channel and Gigabit Ethernet.
- Drivers for IA32, IA64, X86_64, PowerPC, ... ; Linux and Windows. (open source: see http://www.openib.org)

Infinicon InfinIO7000 shared I/O system with switch module, fibre channel and GE modules.



Current Hardware Setup at IWR



Mellanox 16-port switch

13 Dual-Xeon 2.4 GHz worker nodes:
- Infinicon InfiniServ7000 HCAs (4X)
- Supermicro X5DPR-iG2+ board
  (Intel E7501 chipset)
- 2 GB RAM
- RH 7.3, Kernel 2.4.26
- soon: Scientific Linux
  CERN

Interactive node
Dual Xeon 2.4 GHz
Tyan board

Coming up:
24+ node Opteron cluster
Infinicon InfinIO 9100 switch
Sun V20z compute nodes
- Dual Opteron 248 (2.2 GHz)
- 4 GB RAM
- Scientific Linux
  CERN

MPI Performance

MPICH 1.2.5.2 / OSU 0.9.2

D.K. Panda, Ohio State University
http://nowlab.cis.ohio-state.edu/projects/mpi-iba

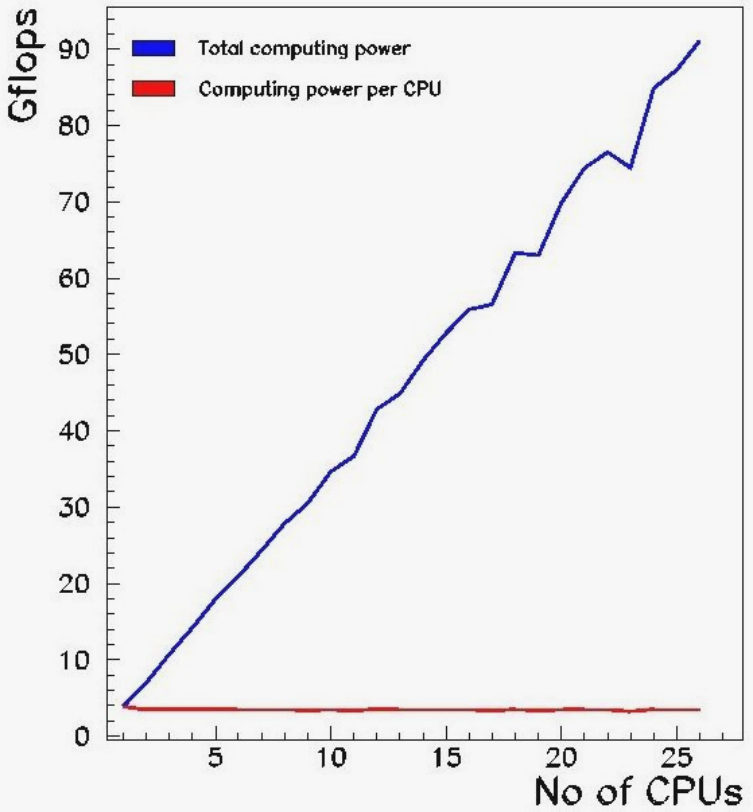
Peak bandwidth ≈ **800 MB/s**
Latency ≈ **6.5 µs**

Up to 880 MB/s between two processes on same node for 16 kB messages but only ≈ 400 MB/s for messages > 16 kB
SMP latency ≈ 1.3 µs

Better: OSU 0.9.4

AMD Opteron:
Bandwidth up to **825 MB/s**
Latency **5.5 µs**

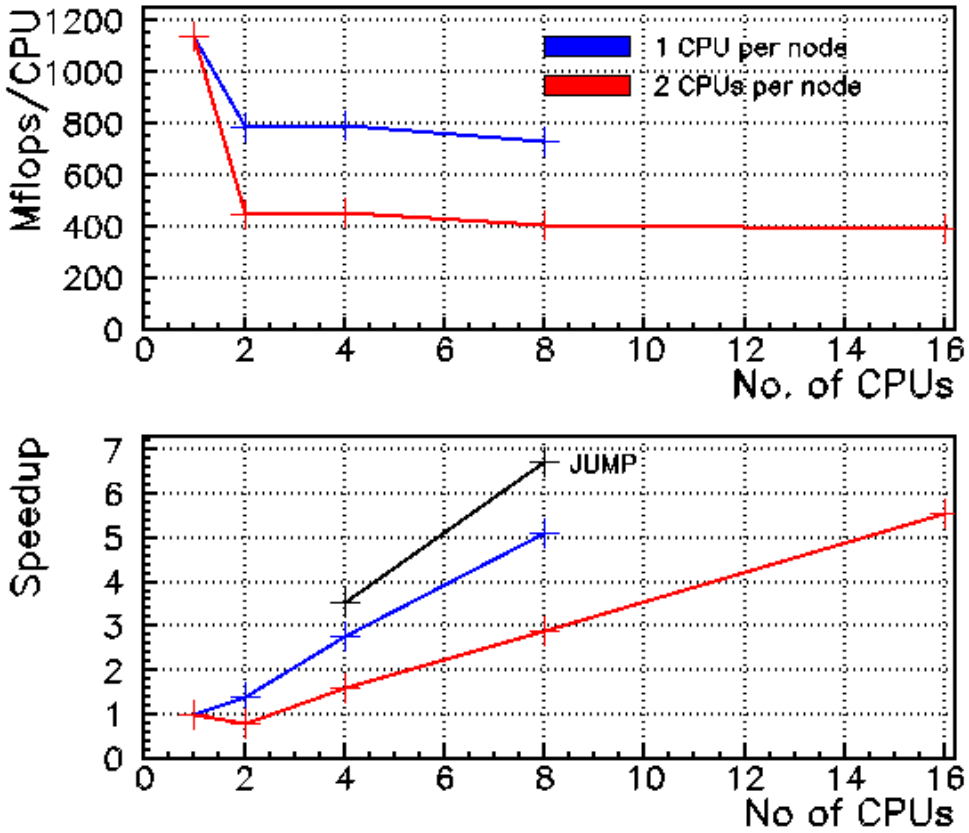
HP Linpack benchmark results (1 to 26 CPUs)



Max. performance **92 Gflops**
R_{max}/R_{peak} = **74%**

MPI Performance

A real world application on the Xeon cluster: Lattice QCD



Memory and communication intensive application.

Possible reasons for low speedup: performance suffers from

- Memory bottleneck due to the single memory controller (Northbridge) on dual Xeon machines.
- Low MPI bandwidth for large messages between two processes on same node.

Thanks to Carsten Urbach
FU Berlin and DESY Zeuthen

Data Intensive Computing w/ InfiniBand

- Data intensive applications suffer from limited bandwidth and CPU overhead for (TCP/IP) network communication.
- InfiniBand IPoIB bandwidth (< 2 Gbit/s) is not satisfying, CPU utilization high due to TCP overhead.
- So far no file transfer / streaming protocol available for native InfiniBand.
- Large amounts of data will need to be moved for LHC data taking and processing.

Port of ***RFIO*** to InfiniBand by FZK.
Make use of RDMA capabilities for low CPU utilization.

RFIO over InfiniBand

RFIO (Remote File Input/Output):
Efficient protocol for remote file access

- Under development at CERN since 1990 (SHIFT project)
- Now part of the CASTOR Storage Manager Project
- Comes with
  - Posix like API library (rfio_open(), rfio_read(), ...)
  - Daemon (rfiod)
  - Set of standard tools (e.g. rfcp, rfdir, rfrm, rfstat, ...)
- RFIO interfaces in ROOT, CERNLIB (PAW), ...

More information on RFIO / CASTOR:

- http://castor.web.cern.ch
- The following presentation by Dr. Durand

RFIO over InfiniBand

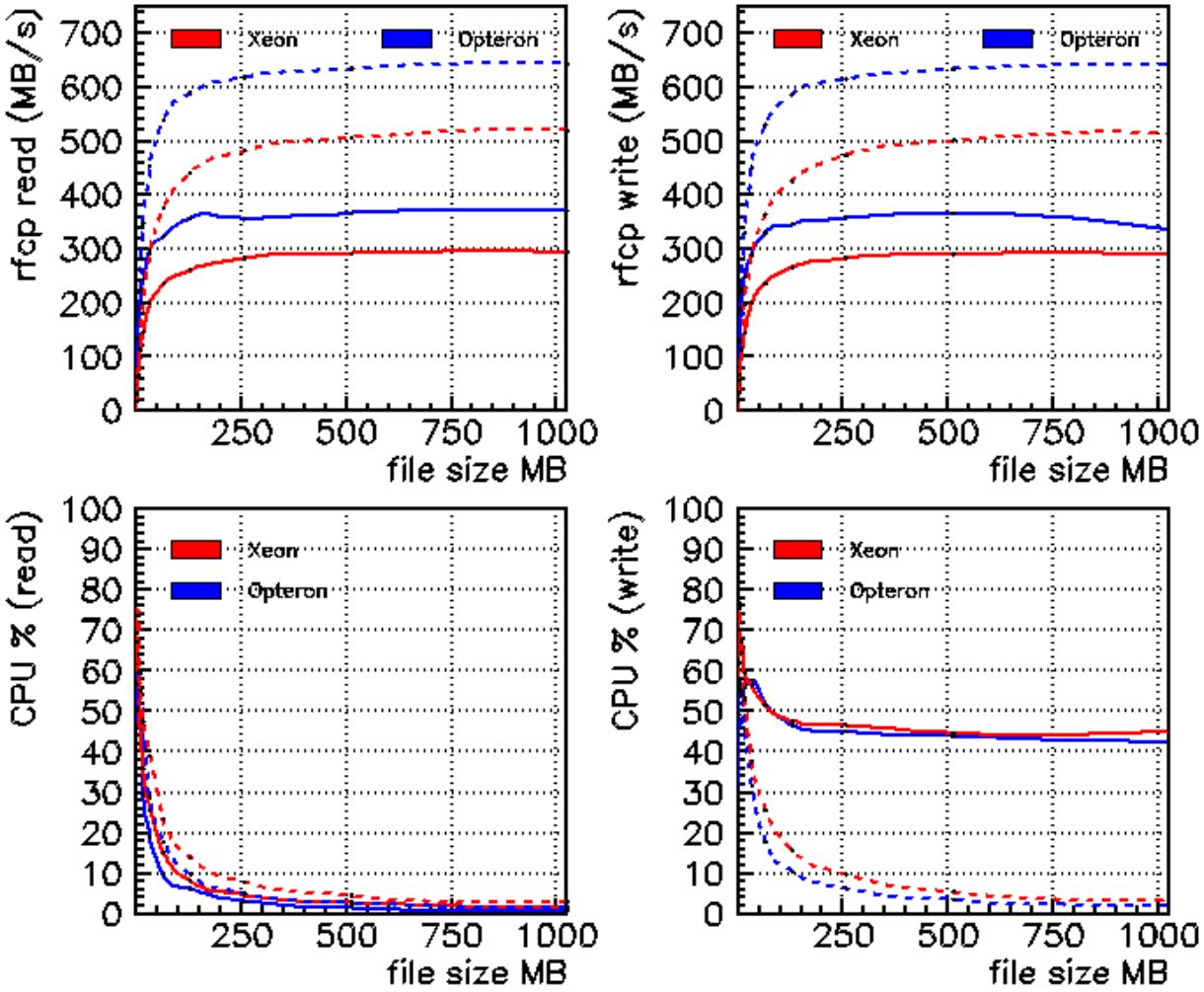
- InfiniBand version does addressing of remote hosts via TCP/IP for compatibility reasons.
  - Nothing changes from the users point of view.
  - No need to modify existing applications (if dynamically linked against libshift.so)

- InfiniBand rfiod has to be multi-threaded due to InfiniBand HCA driver issue.
  → Porting to InfiniBand was more complicated than expected in the beginning.
  'Normal' rfiod does fork().

Implementation based on Mellanox 'Verbs' API (VAPI):
http://www.mellanox.com
http://www.openib.org

RFIO over InfiniBand

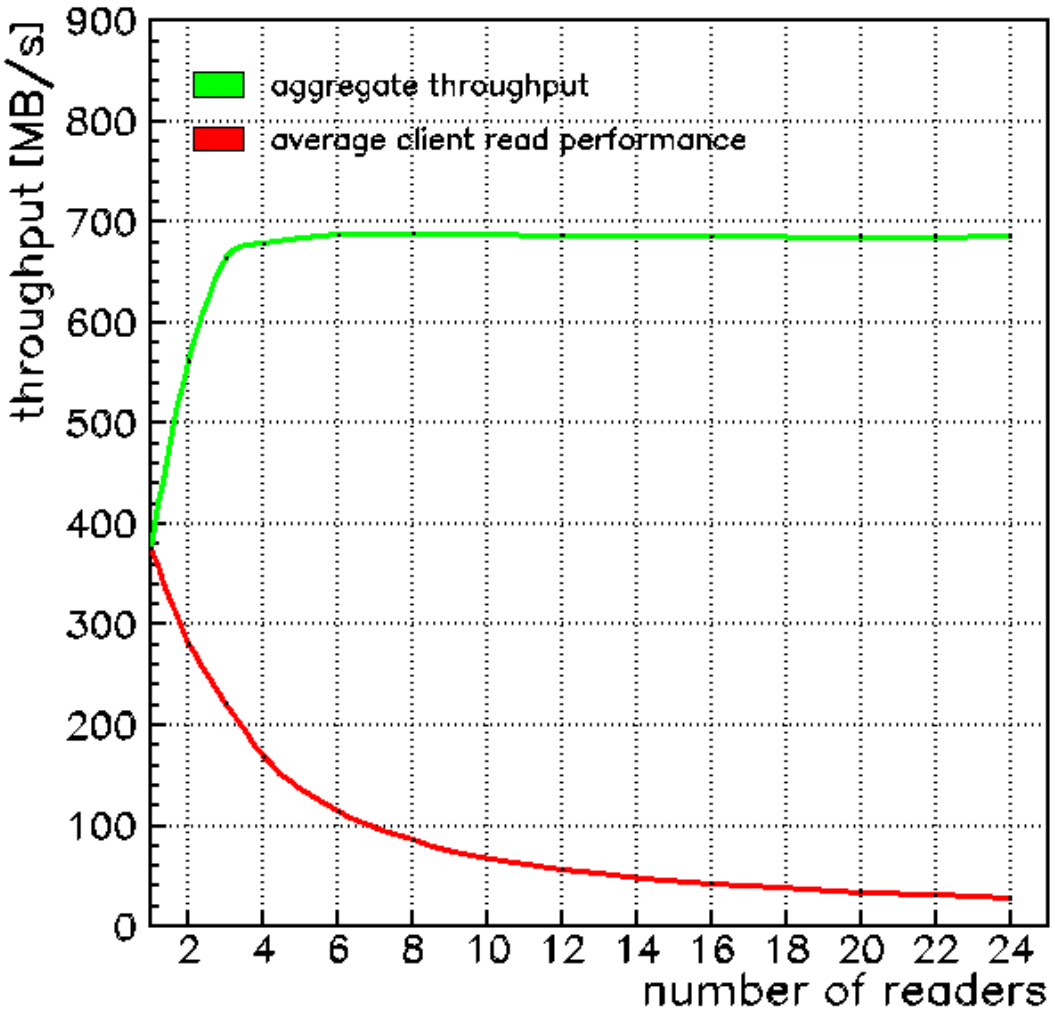
RFIO performance: single client (RDMA streaming protocol)



- Client and server connected via one 4X InfiniBand port to Mellanox switch (dual-Xeon nodes) Infinicon switch card (dual-Opteron nodes).
- Each measurement repeated 100 times.
- Measurements with unfilled buffers show pure memory + interconnect performance.

RFIO over InfiniBand

Multiple clients



Clients on Dual-Xeon nodes, rfiod on quad-Opteron server connected with one 4X InfiniBand port to the Mellanox switch.

- 1 GB sized random data test file
- Two “warm up” transfers to cache the file.
- Each measurement repeated 250 times.
- Total amount of data transferred for this plot is almost 30 TB.

Aggregate throughput ≈ 700 MB/s

ROOTD via native InfiniBand

Porting (X)ROOTD to native InfiniBand is a promising project:

- Need fully multi-threaded (X)ROOTD version.
- Supports different network interfaces.
- Should be easy to write plugins for various interconnects, e.g. InfiniBand, Myrinet, Quadrics or Dolphin.
- InfiniBand plugin: clients would talk to daemon using a special protocol (VAPI://)

Work is in progress, every volunteer willing to participate is highly welcome!

Summary / Outlook

- InfiniBand interconnect is a promising technology based on open standards.
- Interesting for HPC and data intensive computing (HTC).
- Opteron cluster will be built this year at FZK.
- RFIO/IB transfer rate > 300 MB/s, low CPU usage.
- (X)ROOTD/IB project started.
- For updates of results and project status look at

http://www.fzk.de/infiniband