First experiences with the InfiniBand (TM) Interconnect

Andreas Heiss, Ulrich Schwickerath, Institute for Scientific Computing (IWR)
Forschungszentrum Karlsruhe
76021 Karlsruhe, Germany

- InfiniBand: What is this? Key features of a new technology
- Test equipment at the IWR: Hardware used for measurements
- Performance tests and scalability
- High-Throughput-Computing (HTC): RFIO over InfiniBand
- Summary and Outlook

All numbers are preliminary
Motivation:

The problem(s):

- computing power has increased much faster than the interconnects
- increasing need of applications for high bandwidth within computer centers
- parallel computing applications need low latency
- scalability to thousands of computing nodes

The solution (?): InfiniBand (TM)

- merged best of Next Generation I/O (NGIO) and Future I/O (FIO) projects
- specifications released in autumn 2000 by InfiniBand Trade Association (IBTA)
What is InfiniBand?

A fast interconnect technology with open specifications

Key-features:

- low latency channel oriented switched fabric
- runs over copper or fibre cables
- speed: 2.5, 10 or 30 GBit/s (1x,4x,12x)
- (un)reliable and (un)connected data transfers
- RDMA capable
- redundant connections possible
- only one fabric for HTC and HPC applications

Notes:

- **reliable connections**: hardware takes care of the integrity of your data
- **RDMA**: one machine can directly put data into a registered memory of another node without going through the processor
- TeraScale System/Virginia (No. 3 of top 500 list) uses InfiniBand (TM)
Available software

- **low level drivers** available for different architectures and operating systems
  
  (IA32, IA64, X86_64, PowerPC OS: Linux and Windows)

Available high level protocols:

- **IPoIB**: creates virtual ethernet devices, transparent for all applications
- **SRP**: using block storage devices over InfiniBand(TM) fabric
- **MPI**: several implementations, commercial and free, including MPI2
- **DAFS, DAPL, SDP** and more ....

See also [http://infiniband.sourceforge.net](http://infiniband.sourceforge.net)
Hardware setup at the IWR:

**iWarp Cluster**: evaluation of a production environment:
- 8 worker nodes: 2.4GHz Dual Xeon 2GB RAM, GE on board
- 1 interactive node: 2.4GHz Dual Xeon, 1GB RAM, GE on board
- Interconnect: 4x InfiniBand, 16Port 'fat tree' 4x InfiniBand switch

**Test equipment:**
- for software development and hardware evaluation
  - 3 2.4GHz Dual Xeon, GE on board
  - 4x Infiniband between nodes
  - InfinIO7000 Chassis, FC and GE card to 4x InfiniBand backplane
  - 2 IDE Raid boxes with 7x120GB disks each
  - 2GB/s FC connection between RAID and InfiniBand Fabric
MPI performance and scalability

- 1-byte latency: \(~7 \mu s\)
- peak bandwidth 780MB/s
- good scalability (up to our 24 CPU's)
- peak performance \(~70\) GFlops
Status of iWARP cluster:

- available for tests of MPI application
- several test accounts to different persons to test real life applications.
- up and running since late summer, without major problems

Work in progress:

- optimisation of file I/O using the two IDE Raids systems, connected via SRP to fabric
- test of different file systems on the raid systems (cluster file systems, RAID0 ...)
- tests with real world MPI applications (Lattice QCD, climate predictions and others)
- job forwarding queue from cross-grid test bed, first tests are in progress
- software development targeting at High Throughput Computing (HTC) applications
High Throughput Computing (HTC) application: RFIO over InfiniBand

About RFIO:
- efficient protocol for large data transfers
- under development at CERN since 1990
- now part of the CASTOR software suite
- interfaces to RFIO exist in ROOT, CERNLIB, PARROT etc

Basic idea: implementation of an alternative fast streaming protocol (rfcp)

- profit from high transfer rates well above GE capabilities
- keep CPU usage at a low level

Solution (?): combine RDMA and reliable connection (RC) features of InfiniBand(TM)
High Throughput Computing (HTC): RFIO over InfiniBand(TM)

Status of the project:
- code is in early beta development state
- in contact with CERN group A. Horvath/A. v. Praag
- first results look promising
- still some work to be done (performance, multithreading ...)

Preliminary results:
- tests done on Dual-Xeon nodes
- GE via cross-cable (no switch!)
- 100 single measurements for each measurement point
- transfer time and consumption measured using time
- using cached files
network+protocol performance comparison: read/write garbage to /dev/null
performance comparison: reading/writing cached file to /dev/null
True rfcp file transfers to IDE-Raid system: comparison GE and IBA

- bonnie++ write performance: ~130MB/s
- Raid performance reached with IBA
- CPU usage with IBA only half of GE
- drop of transfer rate for large file sizes (needs to be investigated)
Summary of preliminary results on dual XEON

- **RDMA write raw performance**: ~780MB/s
- **rfcp remote garbage to /dev/null**: ~540MB/s (~110 MB/s for GE)
- **rfcp remote cached file to /dev/null**: ~300MB/s (~110 MB/s for GE)
- **rfcp remote cached file to local file**: ~100MB/s (limited by Raid write)
RFIO: Conclusion and more tests

- with InfiniBand (TM), the network is not a bottle neck any longer
- transfer speed limited by XEON server architecture and I/O devices

- 64-Bit architectures: works for Itanium, earlier version was tested on Opteron
- first tests made on Itanium2 give up to 450MB/s at < 10% CPU

(credits: A. Horvath, A. v. Praag, CERN)

- some known problems still to be solved
- close collaboration with people at CERN and Karlsruhe (Jos van Wezel)

Work is in progress!
InfiniBand is a nice open standard for interconnecting computer clusters.

It offers perspectives for HPC as well as HTC computing.

4× is available and working, 12× (30Gb/s) hardware has been announced.

Contact: Ulrich Schwickerath (Ulrich.Schwickerath@iwr.fzk.de)  
Andreas Heiss (Andreas.Heiss@iwr.fzk.de)