

Erfahrungen mit einem InfiniBand-Cluster am Forschungszentrum Karlsruhe

- Technologie, Markt
- vom Testsystem (Jan. 2003) zum IWARP Cluster
- Vorläufige Ergebnisse mit den ersten drei Testrechnern
 - MPI Latenzzeit und Bandbreite
 - Eigenentwicklungen:
 - direkte Messung der Latenzzeit
 - schnelle Dateitransfers mit RFIO
- Erfahrungen
- Ausblick

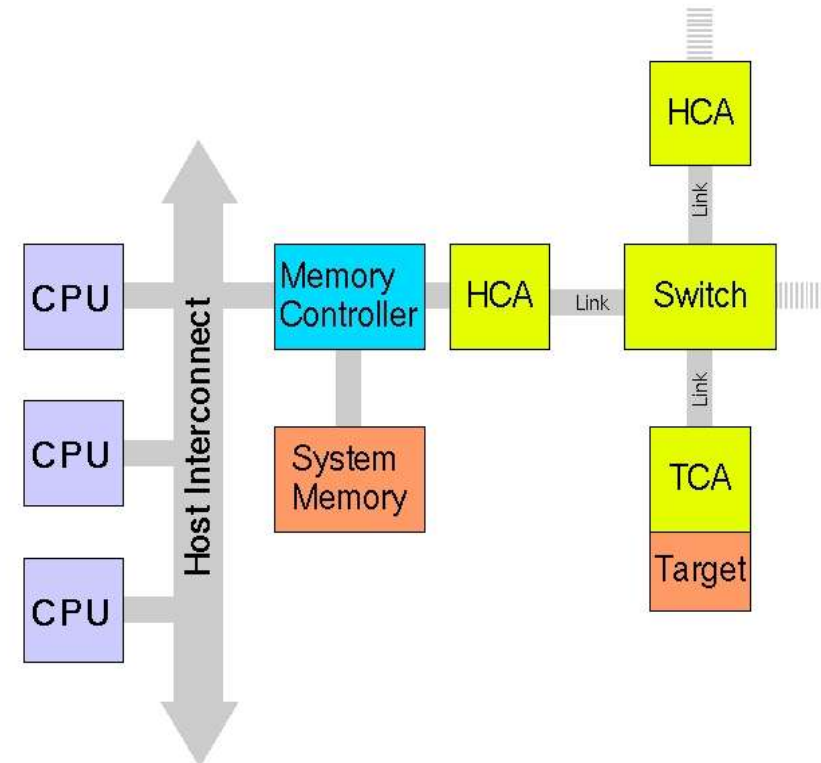
Alle Zahlen und Diagramme sind vorläufig!

Was ist InfiniBand ?

A fast interconnect technology with open specifications

Schlüsseleigenschaften:

- ◆ kanalorientiertes geschwitchtes Netzwerk niedriger Latenz
- ◆ Kupfer oder optische Verbindungen
- ◆ Geschwindigkeiten 2.5, 10 or 30 GBit/s (1x,4x,12x)
- ◆ (un)reliable and (un)connected Datentransfers
- ◆ RDMA fähig
- ◆ redundante Anbindungen möglich
- ◆ ein einziges Netzwerk für HTC and HPC Anwendungen



Anmerkungen:

- ◆ **reliable connections:** Hardware ist für Datenintegrität zuständig
- ◆ **RDMA:** Remote Direct Memory Access
- ◆ TeraScale System/Virginia (No. 3 der 500 Liste) basiert auf InfiniBand (TM)

Software

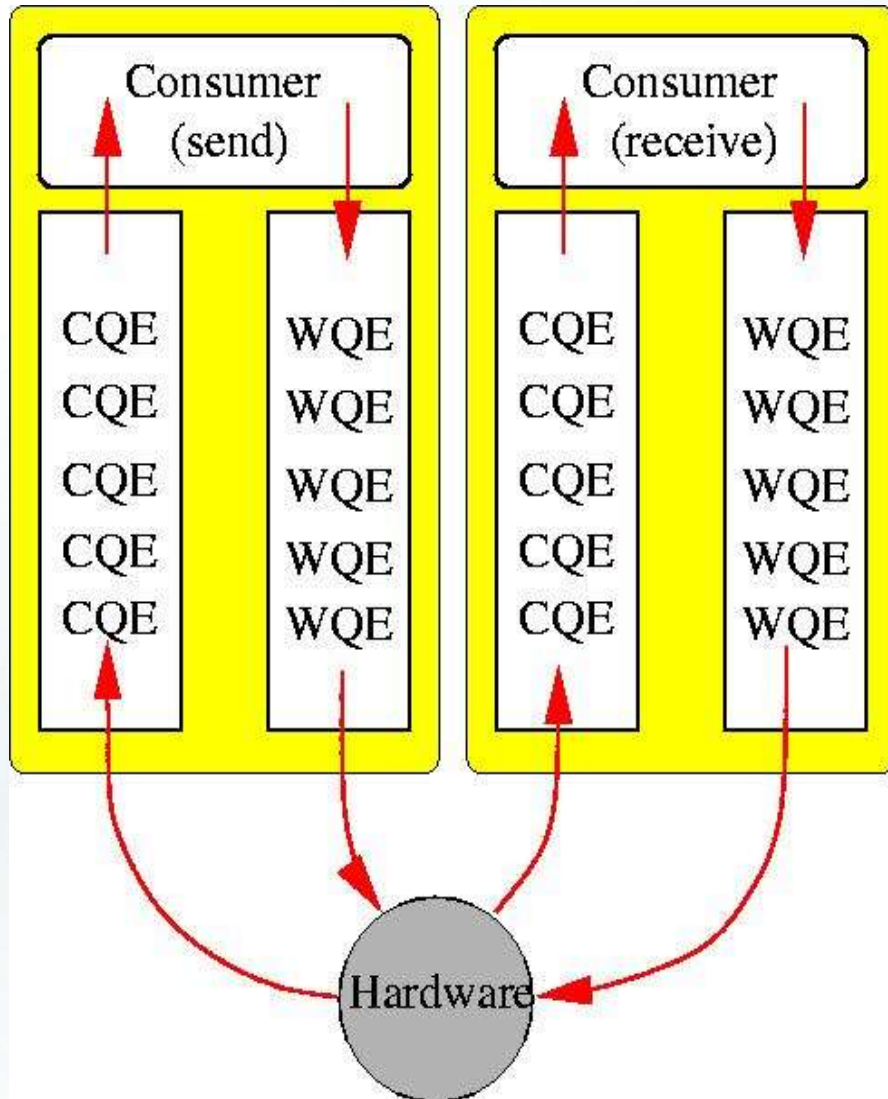
- ➔ Hardwaretreiber für verschiedene Architekturen und Betriebssysteme
(IA32, IA64, X86_64, PowerPC OS: Linux und Windows)
- ➔ Fabric Manager: mehrere Implementierungen, MiniSM, OpenSM ...
- ➔ IPoIB : Emulation von Ethernet Schnittstellen
- ➔ SRP : SCSI RDMA Protokoll: Blockorientierte Massenspeicherverwaltung
- ➔ MPI : mehrere InfiniBand Implementierungen, u.a. Ohio State University
- ➔ DAFS, DAPL, SDP and more

<http://infiniband.sourceforge.net>

Hardwareprogrammierung: Low-Level Ideen

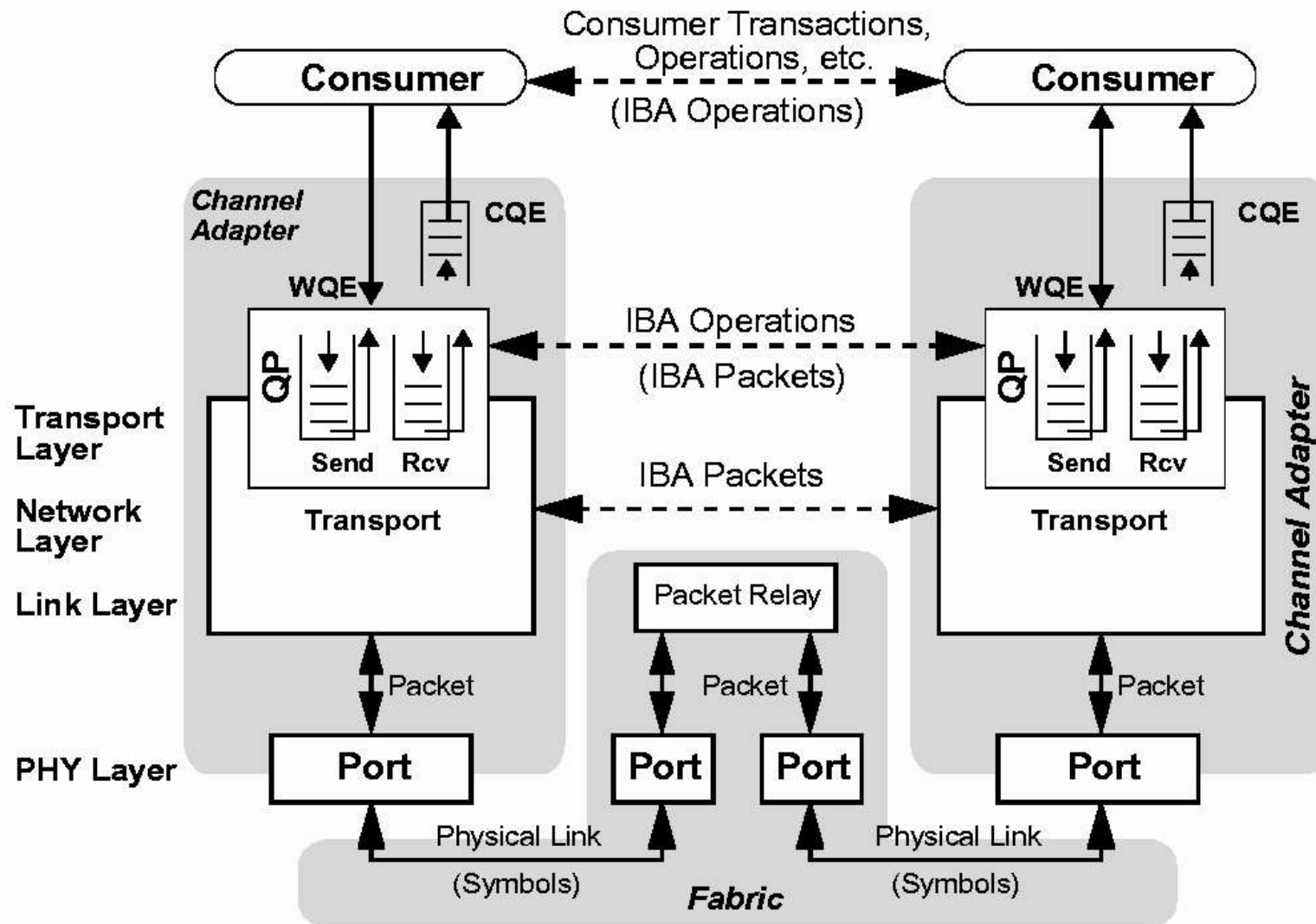
- Spezifikationen definieren Basisfunktionalität der API (Verbs)
- Implementierung dieser Verbs ist herstellerabhängig
- wir verwenden derzeit die Mellanox Verbs API (VAPI)
- Spezifikation der Verbs Funktionalität erleichtert die Portierung von Software
- InfiniBand eröffnet dem Programmierer viele verschiedene Möglichkeiten

Connections: Queues und Queue Pairs



- Work Requests gehen in "Work Request Queues" (WRQ)
- zu jeder WRQ gehört eine "Completion Queue" (CQ)
- Hardware erzeugt Einträge in der CQ
- Warteschlangen für Senden und Empfangen sind getrennt
- Je zwei Warteschlangen bilden einen QP (Queue Pair)
- Transporttyp ist (RC, UC etc) ist eine Eigenschaft der QP

Connections: Queue pair Konzept



- a QP is associated with one (or more) remote QP's
- the send queue talks to the receive queue of the remote QP
- and vice versa
- error events occur as CQE

(taken from IBTA
InfiniBand Specifications)

Marktsituation

- Chips für Host Channel Adapter: Mellanox(4x), Fujitsu (?), Intel (1x)
- Switch Chips: Mellanox, Agilent/RedSwitch, ...

- 4x HCA's von verschiedenen Anbietern (meist Mellanox basiert)
- Preise: fallende Tendenz, nach Rabatten fragen lohnt sich
- Switches: 4x, neuerdings 12x, bis ~100 Ports
- HCA's: 2Ports mit je 4x, 133MHz 3.3V PCI-X, full und low profile boards (nur ein port in eine Richtung voll nutzbar!)
- neuerdings: HCA's (2 4x Ports) für PCI-Express
- derzeit nur Kupfer, ab Sommer (?) auch optische Verbindungen erhältlich

Infiniband-Projekt des IWR: Fabric Setup (seit Jan. 2003)

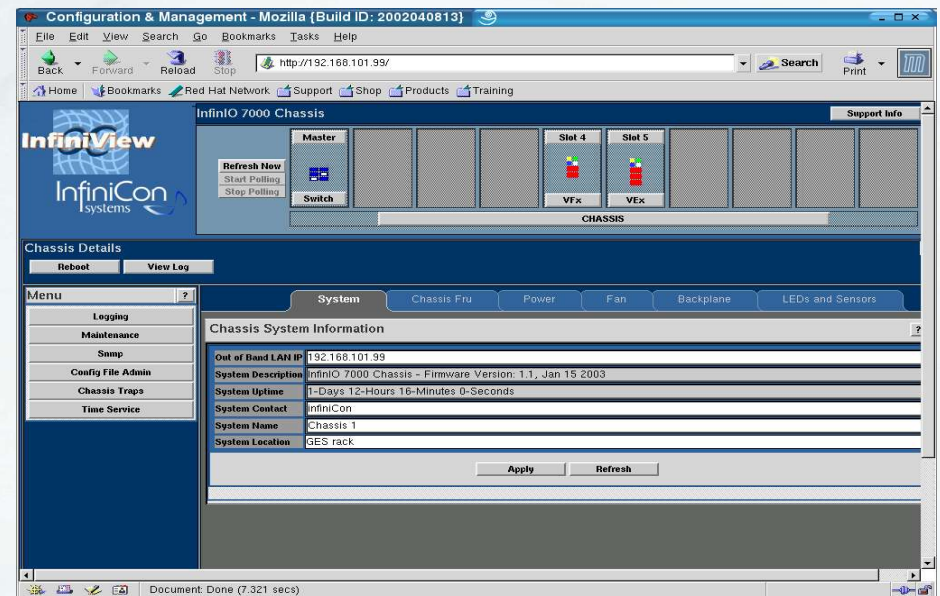
Infiniband Fabric Hardware:

- InfinIO 7000 chassis mit 4X Backplane
- 4X Switch Modul, 6 externe Ports
- 2 Port 2Gb FC auf Infiniband Modul (TCA)
- 3 Port 1Gb-Ethernet auf Infiniband Modul (TCA)
- Hersteller: InfiniCon Systems



Infiniband Fabric Software:

- ICS firmware version 1.1
- MiniSM/ICS/Lane15 SM im Test
- ICS InfiniView Chassis
Kontrollprogramm



Ergebnisse mit dem ersten Testsystem:

- Dual Xeon 2.4 GHz
- 512MB und 1 GB RAM
- Tyan Thunder i7500 and Tiger i7501 Boards
- Intel E7500/E7501 chipset
- Fast (i82550) and Gb(i82544GC) Ethernet
- PCI-X 133MHz 3.3V
- 3 InfiniServ 7000 HCA's
- 1 Mellanox Cougar HCA

ausgestellt am Linuxtag 2003 in Karlsruhe



Software:

- RH Linux 7.3
- Kernel 2.4.18-27.7x with Lustre Patches
- ICS InfiniHost™ 1.1beta
- ICS subnet manager
- ICS MPI
- Mellanox SDK 0.120 and 0.20
- MPICH 1.2.2.2 (Ohio-State Univ. 0.91)

IP over Infiniband (IPoIB) mit ICS - Software

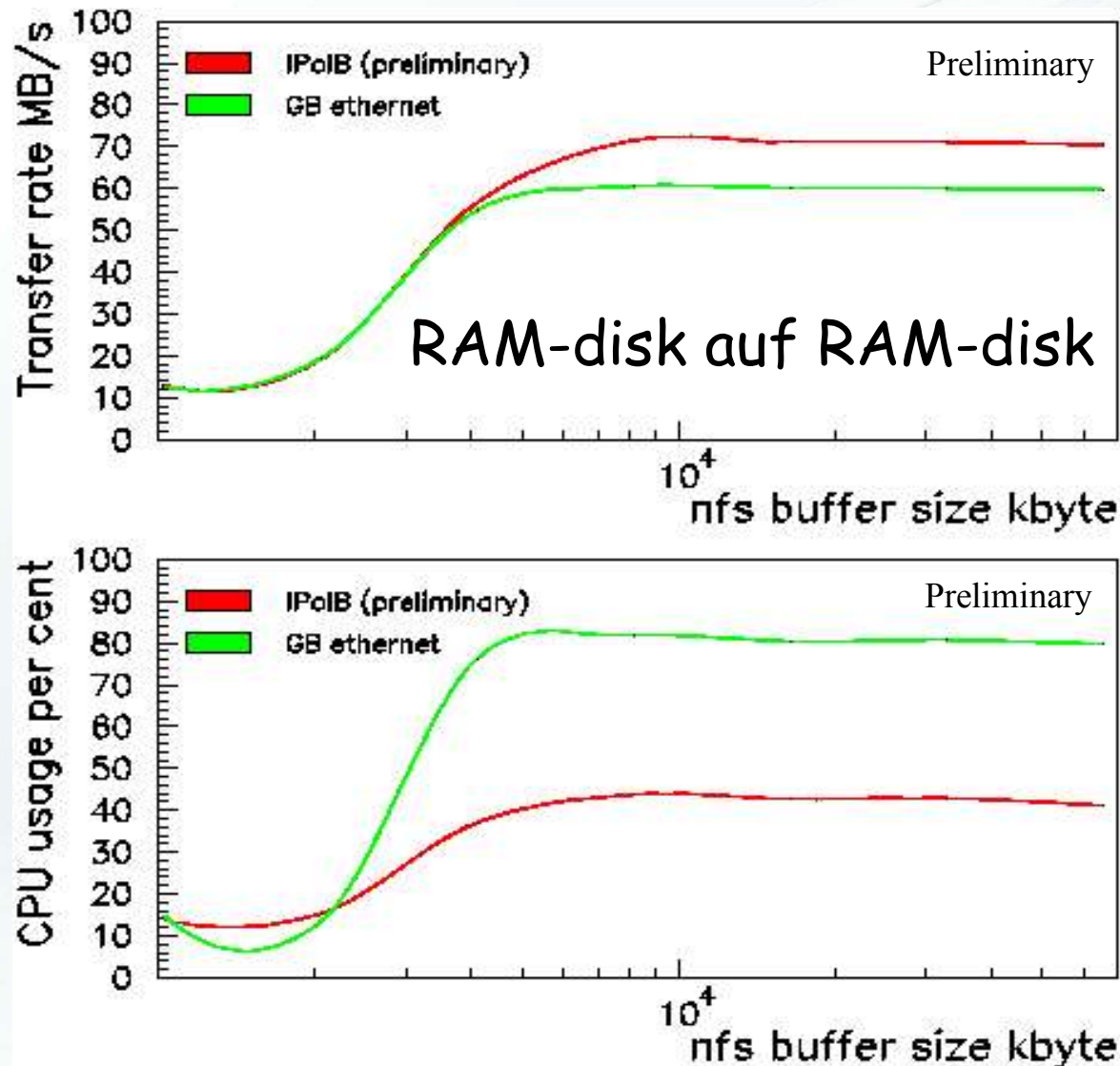
Idee: Emulation eines virtuellen Ethernet - Gerätes

- Software - emulierter Ethernetadapter
- Kein physikalischer Ethernet Anschluss!
- Transparent für Netzwerk-Applikationen wie nfs, ftp, http, rfio ...
- Performance - Einbußen durch Software- Overhead

Messergebnisse: (von Frühling 2003)

- IPerf: TCP 1.51 GBit/sec
- IPerf: UDP 2.1 GBit/sec
- vsftp read (von RAMDISK): 88MB/sec

IP over Infiniband (IPoIB) mit ICS - Software, frühe Ergebnisse



- Vergleich mit Gigabit Ethernet
- Beta-Release Software
- 256MB Testdatei, Zufallsdaten
- RAM Disk nfs exportiert
- Schreibzugriff auf RAM disk

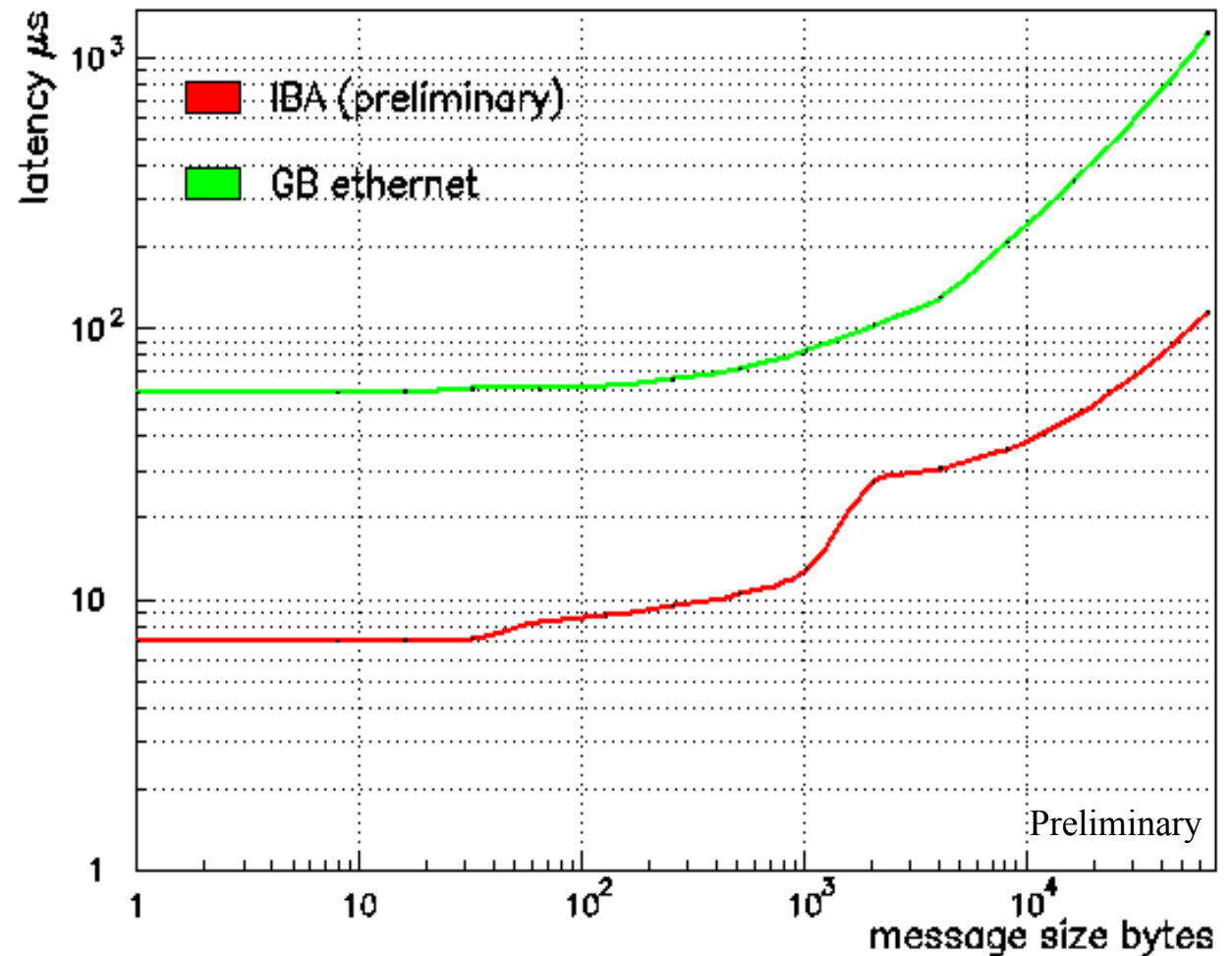
- halbe CPU - Last mit Infiniband
- etwa 30% höherer Durchsatz

Latenzzeit mit OSU MPI Testprogramm

- ICS MPI, basiert auf MPICH 1.2.22 OSU,
- ICS Infiniband - Treiber
- Vergleich mit on-Board GE (switched)

GigaBit Ethernet: etwa $60\mu\text{s}$

Infiniband : ca. $7\mu\text{s}$

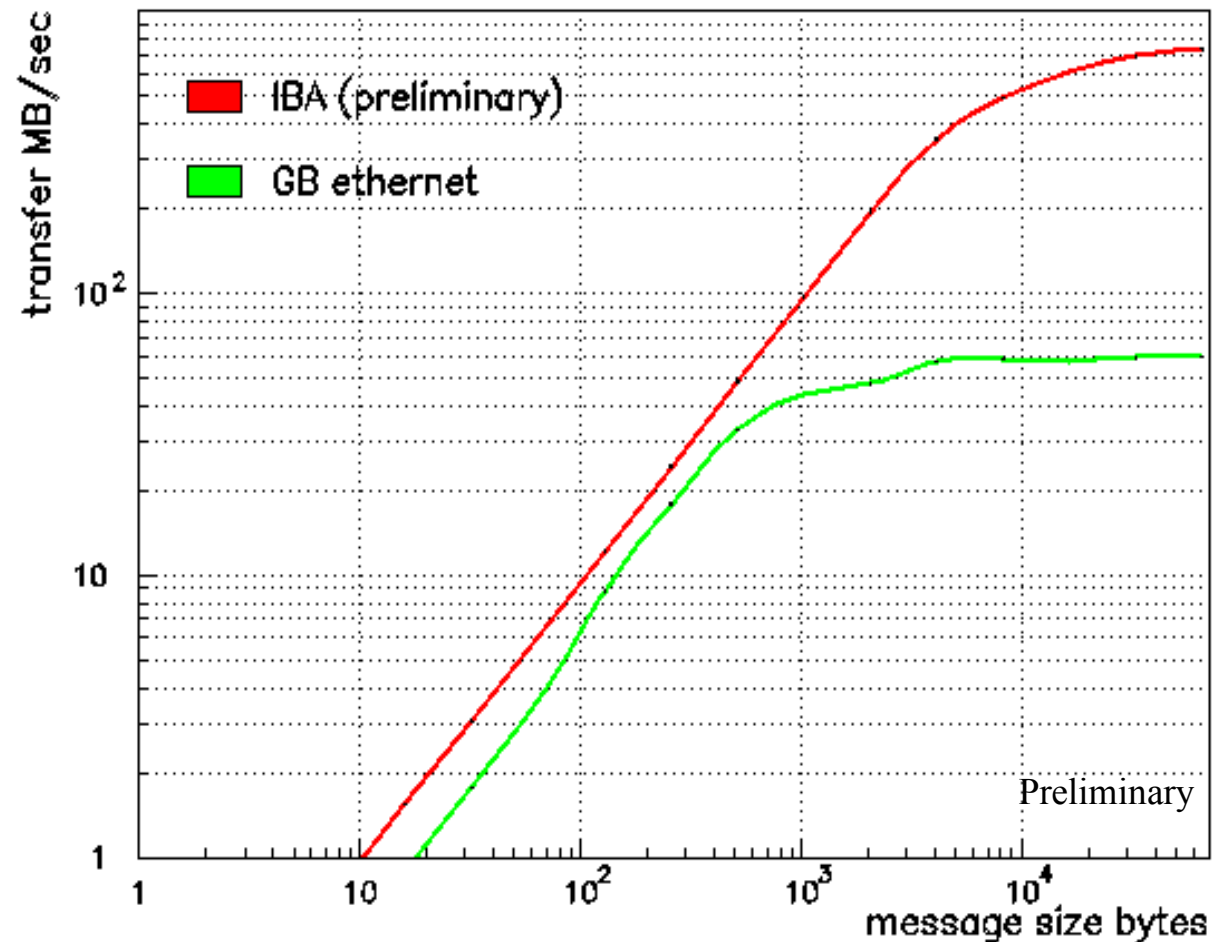


Bandbreitenmessung mit OSU MPI Testprogrammen

- ICS MPI, basiert auf MPICH 1.2.22 OSU
- ICS Infiniband - Treiber
- Vergleich mit GE (switched)

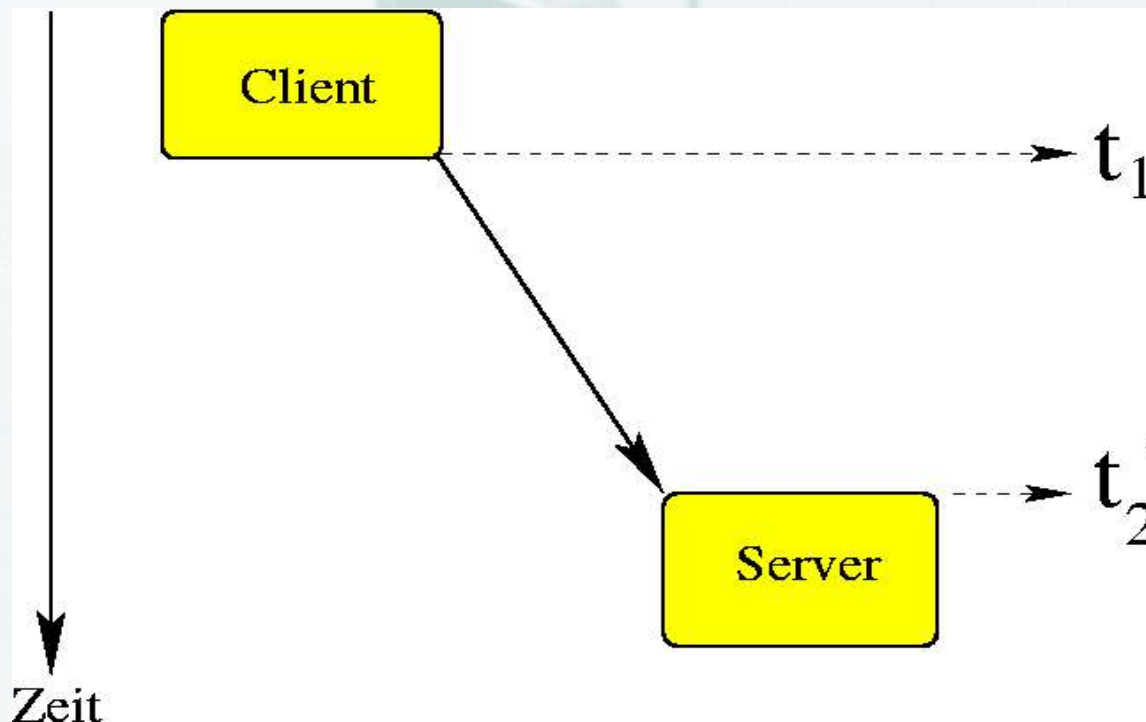
GigaBit Ethernet: bis zu 60MB/s

Infiniband : bis zu 780MB/s

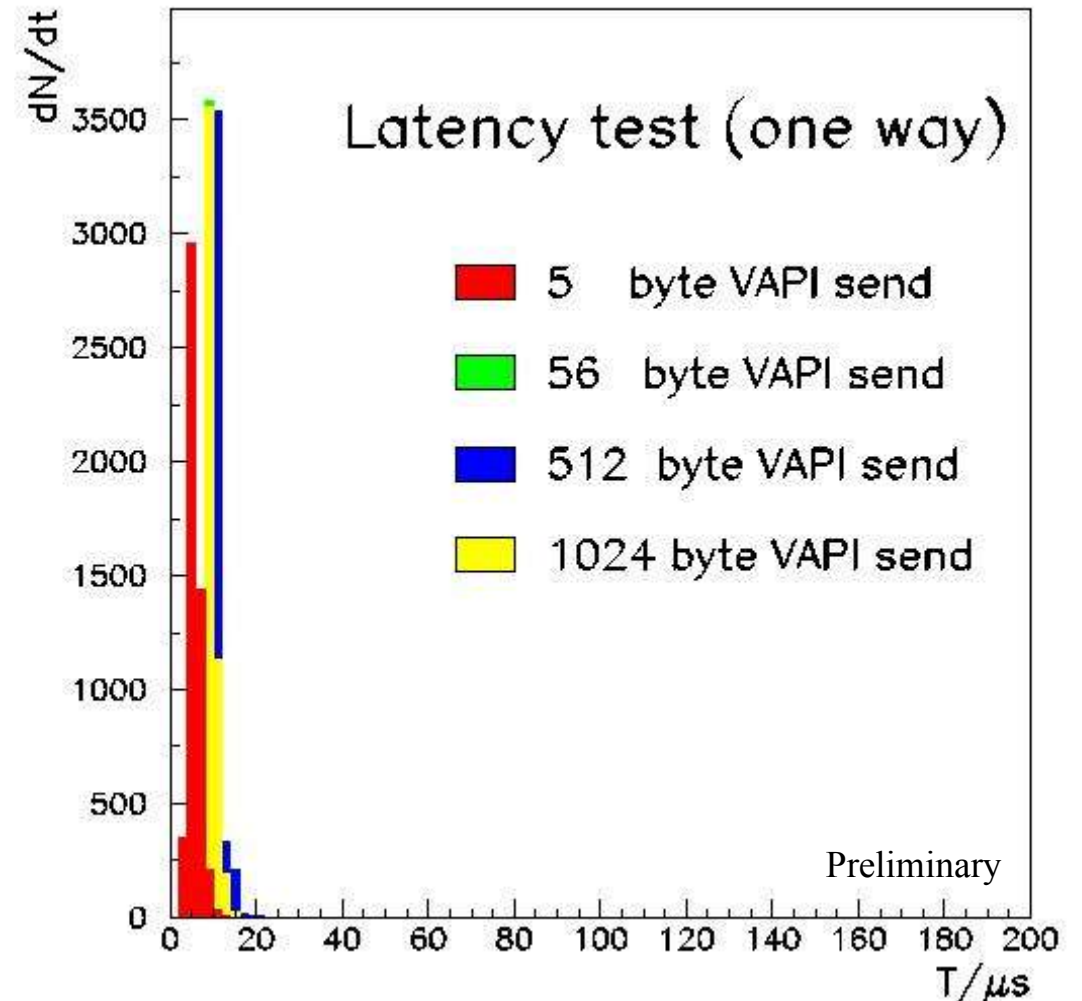
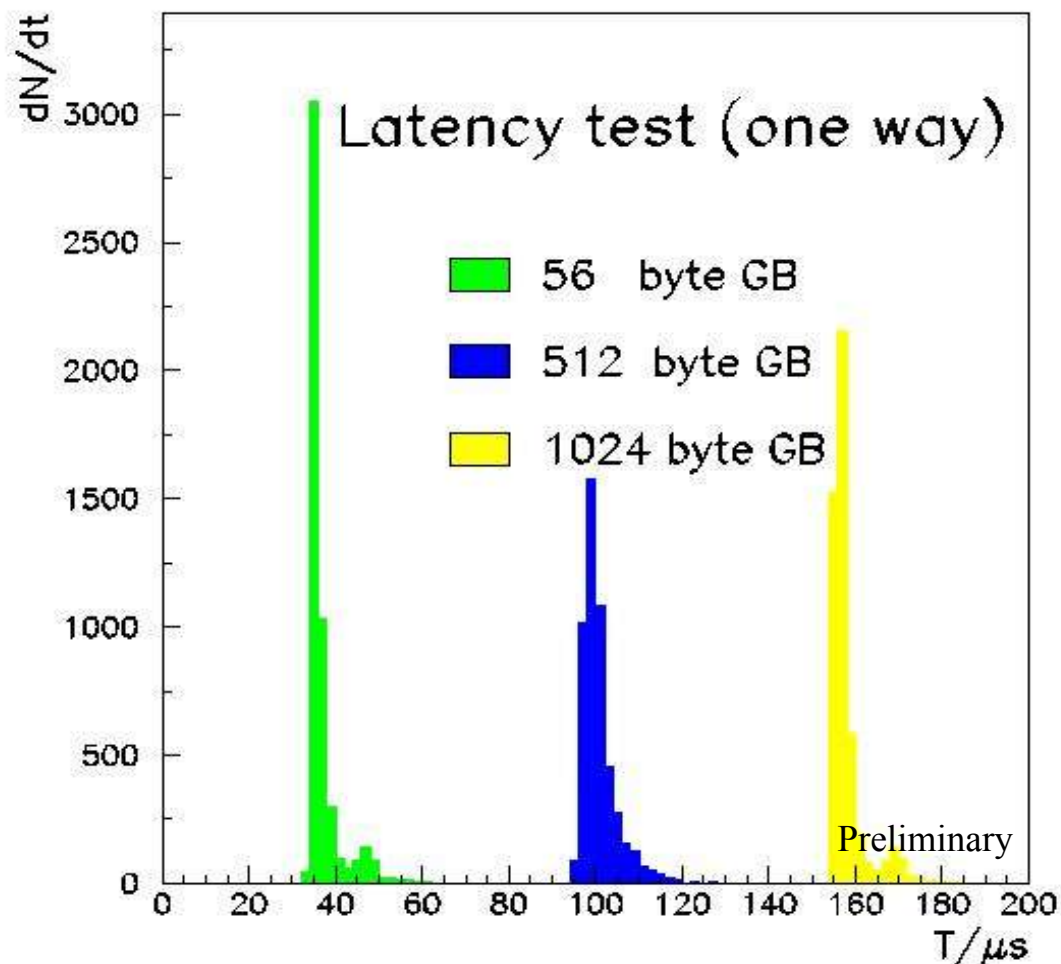


Eigenentwicklung: Latenzzeitmessungen

- Idee: direkte Messung der Signallaufzeit $t_2 - t_1$
- Erfordert exakte Synchronisation der Systemuhren



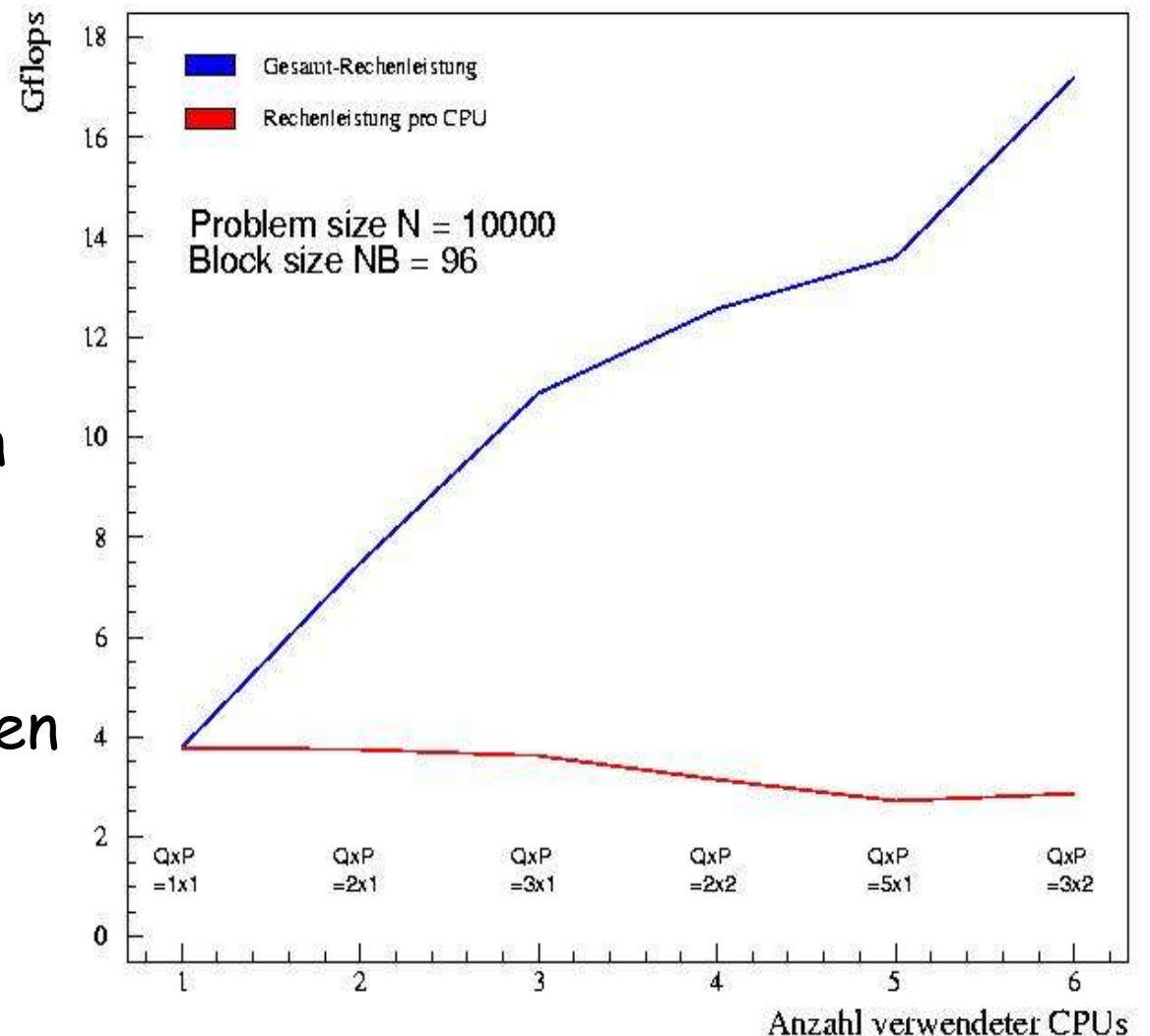
"One way" Latency - Vergleich GE and Infiniband



Weiteres Testsystem - Ausbau: IWARP Projekt

Projektziele:

- Skalierbarkeitstudien
- Prototyp für größere Installation
- Benutzerbetrieb als MPI
Parallelrechencluster
- Testsystem für Eigenentwicklungen



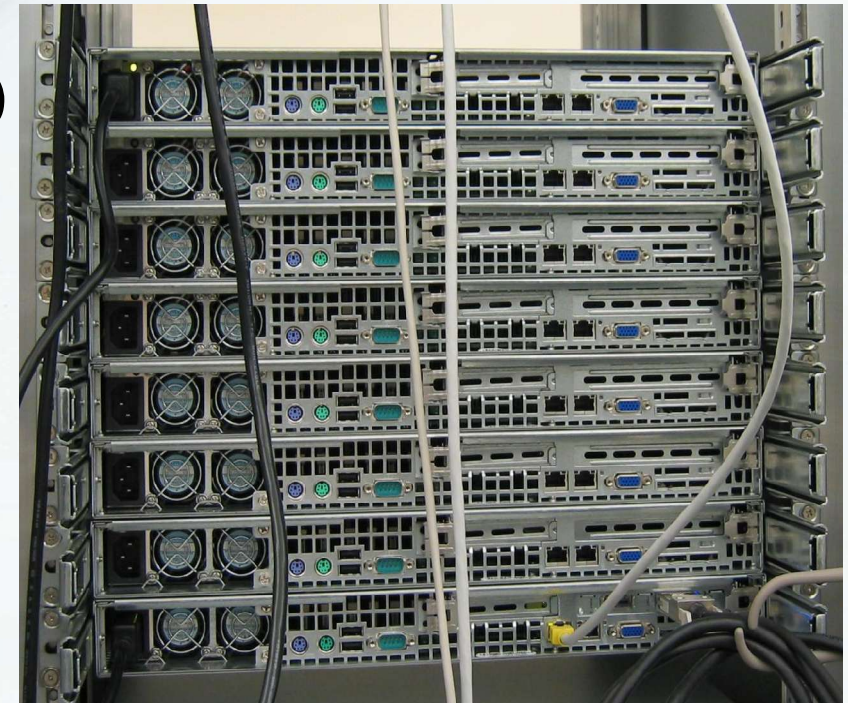
Weiterer Testsystem - Ausbau: IWARP Cluster Projekt

Hardware setup:

- 8 Rechenknoten, 2.4GHz Dual Xeon, Supermicro - Boards
- InfiniCon InfiniServ7000 4X Infiniband-Karten
- Je 2 GB Hauptspeicher
- 16 Port Mellanox Switch (Referenzdesign)

Status:

- Seit September in Betrieb
- Erste Tests: MM5 und HPL Benchmark



Weiterer Testsystem - Ausbau: IWARP Projekt

Cluster-Setup:

- Installation der Rechenknoten mit LCFGng
- LCFGng Server in User-Mode-Linux Umgebung
- Rechenknoten in privatem Netzwerk (FE und IPoIB)
- Zugang nur über Vorrechner
- Open-PBS, Ganglia Monitoring

Software Konfiguration:

- Kernel.org 2.4.21 (bald 2.4.25)
- SDK 1.01, MiniSM oder Lane15 SM (OpenSM?)
- MPI OSU 0.91 (bald 0.92) mit IFC und GCC
- MPI2 Implementierung der TU Chemnitz
- InfiniBand-RFIO
- NFS home directories

Übersicht derzeitige Hardware

13x Dual Xeon 2.4GHz Knoten

16 port 4x-InfiniBand Switch (Mellanox)

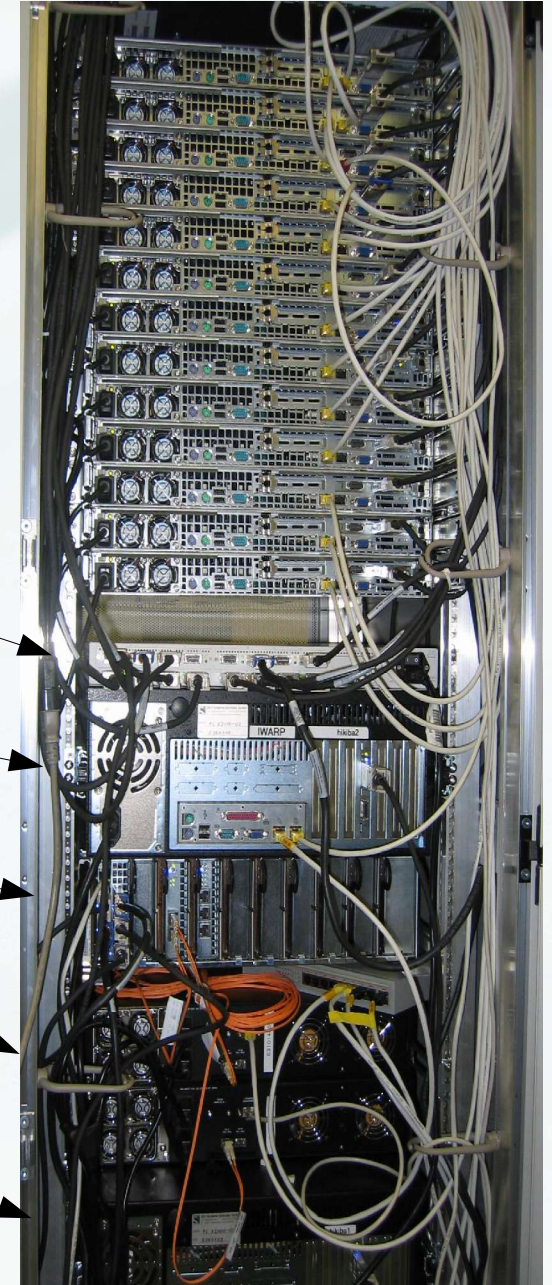
Vorrechner (IWARP) für logins

InfinIO-7000 (FC, GE, Switch)

2x IDE RAID Systeme mit 2GB FC

Testknoten (Dual Xeon 2.4GHz)

Schrankmanager/Subnetmanagerknoten



Erfahrungen:

MPI:

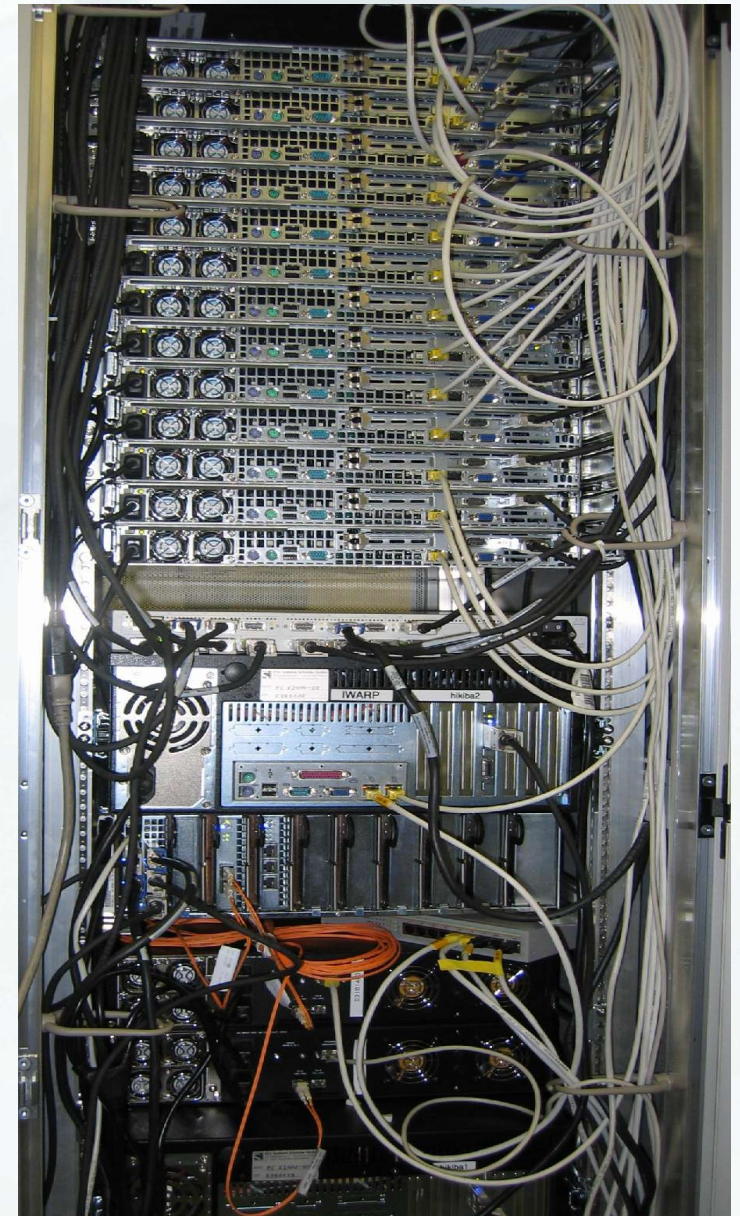
- HPL Benchmark mit 24 CPU's ergibt ca. 70 GFlops
- MPI mit OSU9.1 gibt Probleme bei shared memory
- MPI mit OSU9.2 behebt diese
- HPL Benchmark kritisch/gut für Systemtests

Hardware:

- bislang keine Probleme/Ausfälle aufgetreten

Treibersoftware:

- Mellanox: Quellen vorhanden, restriktive Lizenzen
- InfiniCon: bislang keine vollständigen Quellen



Eigenentwicklung: rfio über Infiniband

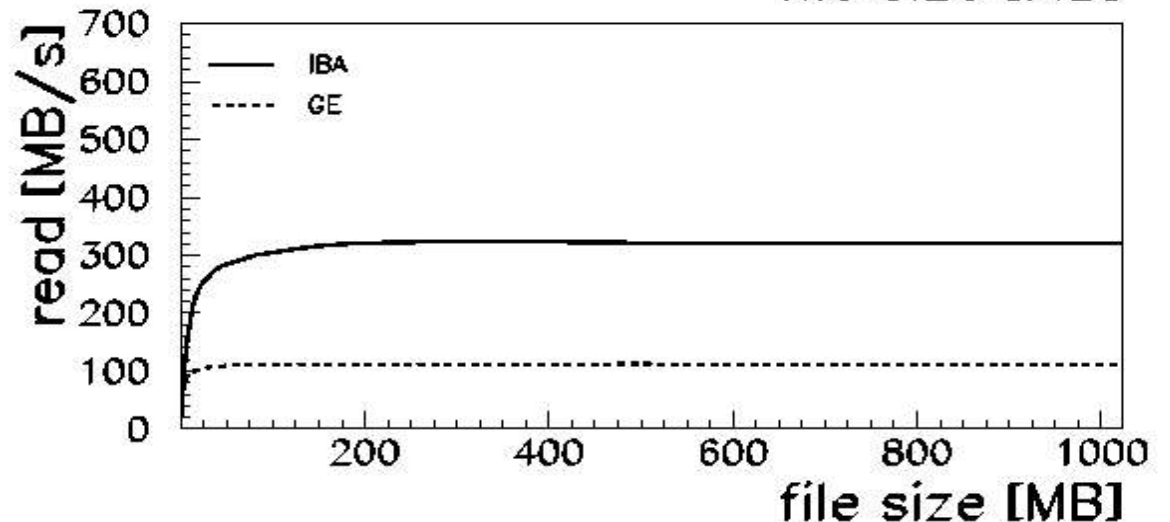
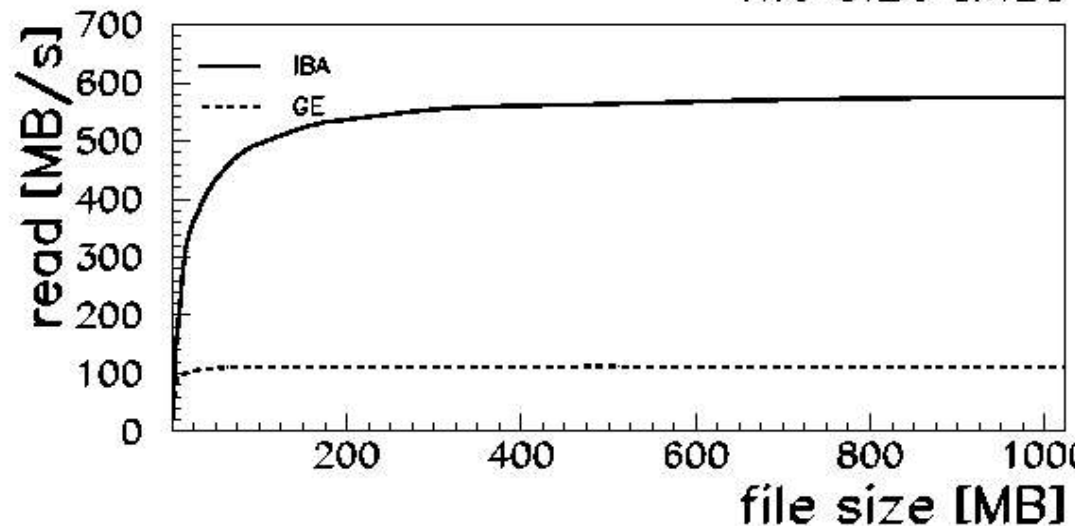
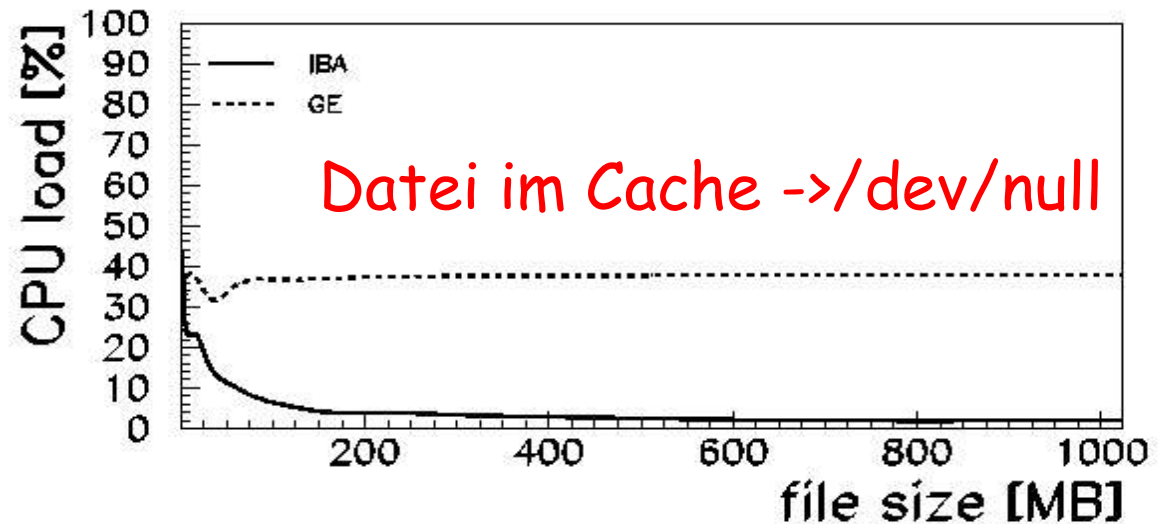
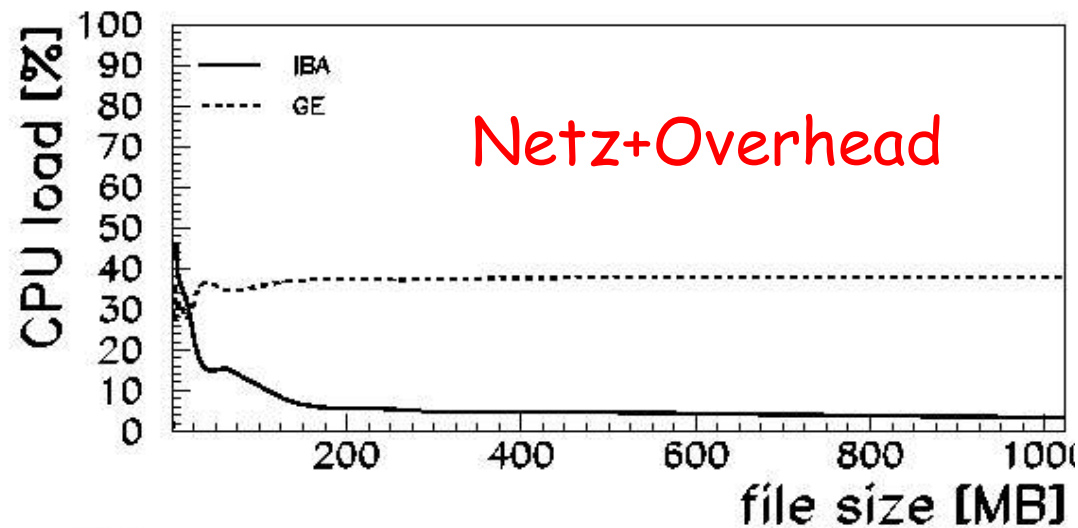
- RFIO:
- effizientes Protokoll für Zugriff auf Files auf entfernten Rechnern
 - entwickelt am CERN (seit 1990, SHIFT Projekt)
 - heute Teil des CASTOR (HSM System) Projekt

- Idee:
- Verwendung von RDMA und RC Technik zur Datenübertragung
 - Adressierung/Kontaktaufbau weiterhin über Ethernet (Transparenz!)
 - Zusammenarbeit mit CERN (Gruppe Ari van Praag) und CASTOR Entwickler

Vorteile:

- Transparent für eventuelle Anwender
- Nutzung bereits bestehender Interfaces zu Programmen wie ROOT

Erste Ergebnisse (ACAT03)



Messergebnisse: (rfcp Dateitransfers)

- RDMA raw performance: ca. 780MB/s
- remote read (Netzwerk + RFIO Overhead): ca.: 550MB/s
- remote read (Cached -> /dev/null) : ca.: 300MB/s
- echte Transfers: limitiert durch Plattenperformance (ca. 120MB/s)
- GE: immer limitiert durch Netzwerkperformance

Interpretation:

- Netzwerk (InfiniBand) KEIN Flaschenhals mehr !
- sehr niedrige CPU Last insbesondere bei grossen Dateien
- Speicheranbindung im XEON Rechner ist Flaschenhals
- Ergebnisse vom CERN auf Itanium2 (gleicher Code):

Cached->/dev/null 450MB/s

Einige Erfahrungen ...

Hardware:

- InfiniBand Kupferkabel sind empfindlich (eg. gegen Biegung)
- Steckkarten haben bisher keine Probleme gemacht
- Serverknoten: PCI-X mit 133MHz über Risercard ist nicht trivial!

Firmware:

- Firmware-updates auf HCA's mit InfiniCon Software ist trivial
- Firmware-Version und Hostsoftware müssen zueinander passen
- Firmware-update mit Mellanox SDK ist etwas umständlich
- Firmware-update unseres Switches war sehr umständlich

Hostsoftware:

- InfiniCon liefert nur Binärdateien
- ... neuerdings aber auch eine Source Code Lizenz
- Mellanox SDK enthält die kompletten Quellen
- ... und läuft im Prinzip auch auf InfiniCon HCA's (Firmware Problem!)

MPI: • sehr empfindlich auf Paketverluste (Kabel!)

Lizenzpolitik:

- bislang restriktive Lizenzpolitik von Mellanox (NDA etc)
- starke Unterstützung des InfiniBand SF Projektes
- inzwischen Release des SDK unter Wahlweise **GPL** oder **BSD** Lizenz
- <ftp://ftp.mellanox.com>

Zusammenarbeit der Software:

- Software/Firmware verschiedener Hersteller arbeitet nur bedingt zusammen
- HCA - Firmware nur bedingt austauschbar
- InfiniBand SF Software wurde noch nicht getestet (steht aber aus)
- Tests der Anbindung von Block - Geräten unter Mellanox/SF steht noch aus
(“InfiniBand SAN”)

Ausblick:

- MPI: Neue Version (OSU9.2) mit Shared Memory Support
- upgrade auf Kernel 2.4.25 beinahe abgeschlossen
- ev. Linux Version Update, Kernel 2.6 Tests
- InfiniBand SF Software: OpenSM statt InfiniCon SM
- Portierung der Eigenentwicklungen auf IBAL (Zeit ?)

weitere Zukunft:

- Einbindung des IWARP Clusters in CampusGRID Projekt am Forschungszentrum
- Aufbau eines weiteren 64bit Clusters mit InfiniBand (Opteron oder Itanium2)
- Tests von Cluster - Filesystemen, etwa LUSTRE

Zusammenfassung

- zukunftsweisende, offene Technologie
- RDMA mit 10GB/s schon jetzt zu erschwinglichen Preisen
- 12x (30GB/s) und optische Verbindungen sehr bald erhältlich
- Kinderkrankheiten werden weniger und weniger
- Hemmschuh Lizenzpolitik schwindet

Closed session Teil: Erfahrungen mit Firmen

Mellanox:

- Treibersoftware komplett als Quellen vorhanden
- Notwendigkeit des Abschlusses eines NDA
- Notwendigkeit des Kaufes eines HCA direkt von Mellanox
- restriktive Lizenzpolitik bis Ende vorige Woche
- Dokumentation ist eher nicht vorhanden
- Bei Problemen antworten Mitarbeiter von Mellanox, und bemühen sich zu helfen (was nicht immer gelingt)
- Firmware updates: bei unserem Switch bislang nicht gelungen

Closed session Teil: Erfahrungen mit Firmen

InfiniCon:

- Hardware HCA's Mellanox basiert
- hohe Rabatte für Forschung und Lehre
- kommerzielle Kunden im Visier, weniger für Bastler wie wir
- keine Treiberquellen
- Treiberupdates auf Anfrage im Prinzip möglich
- Support:
 - mehrere kostenlose Hardwareupdates, auch unangekündigt
 - Telefonkonferenzen, auch mit Entwicklern
 - effektiver Trainingskurs bei InfiniCon