Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Erfahrungen mit InfiniBand (Forschungszentrum Karlsruhe, IWR)

- 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

http://www.fzk.de/infiniband

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Was ist InfiniBand ?

A fast interconnect technology with open specifications

Schlüsseleigenschaften:

- kanalorientiertes geswitchtes 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)

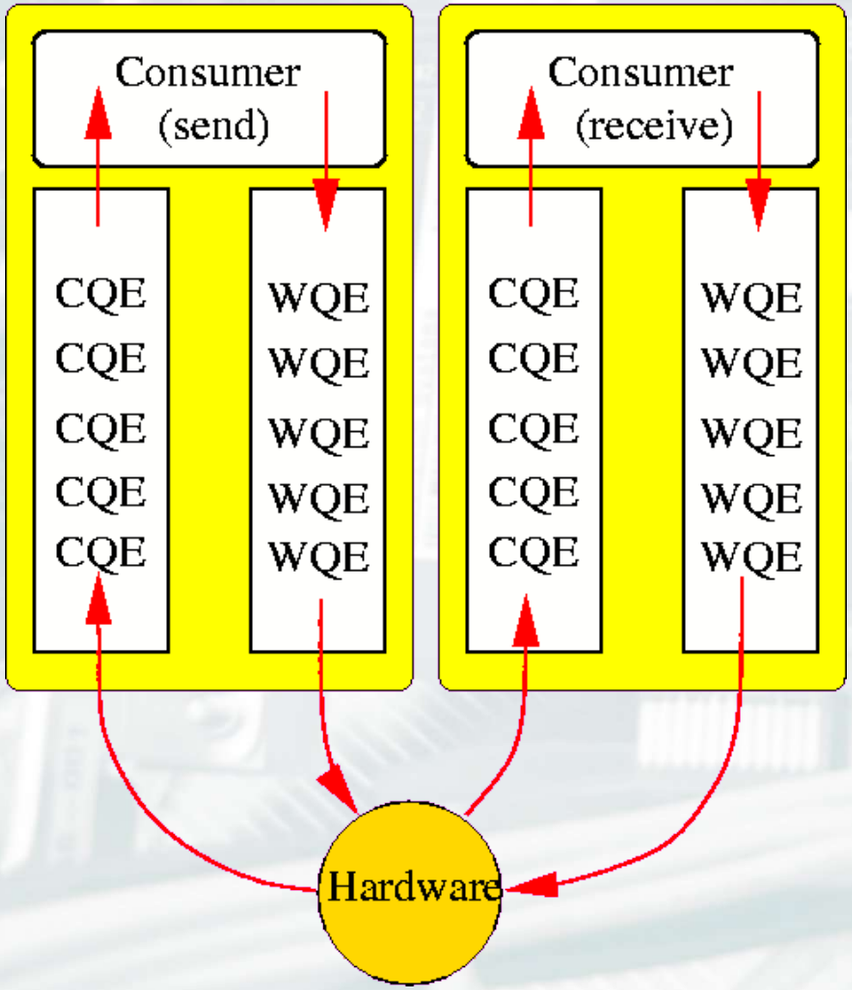
Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Programmierung der Hardware: VERBS

- Spezifikationen definieren Software Layer zur Programmierung
- lediglich Funktionsklassen und ihre Funktionsweise werden definiert
- jeweilige Implementierung dieser VERBS ist herstellerspezifisch
- Kanalorientierung: getrennte Queues zum Senden und Empfangen

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

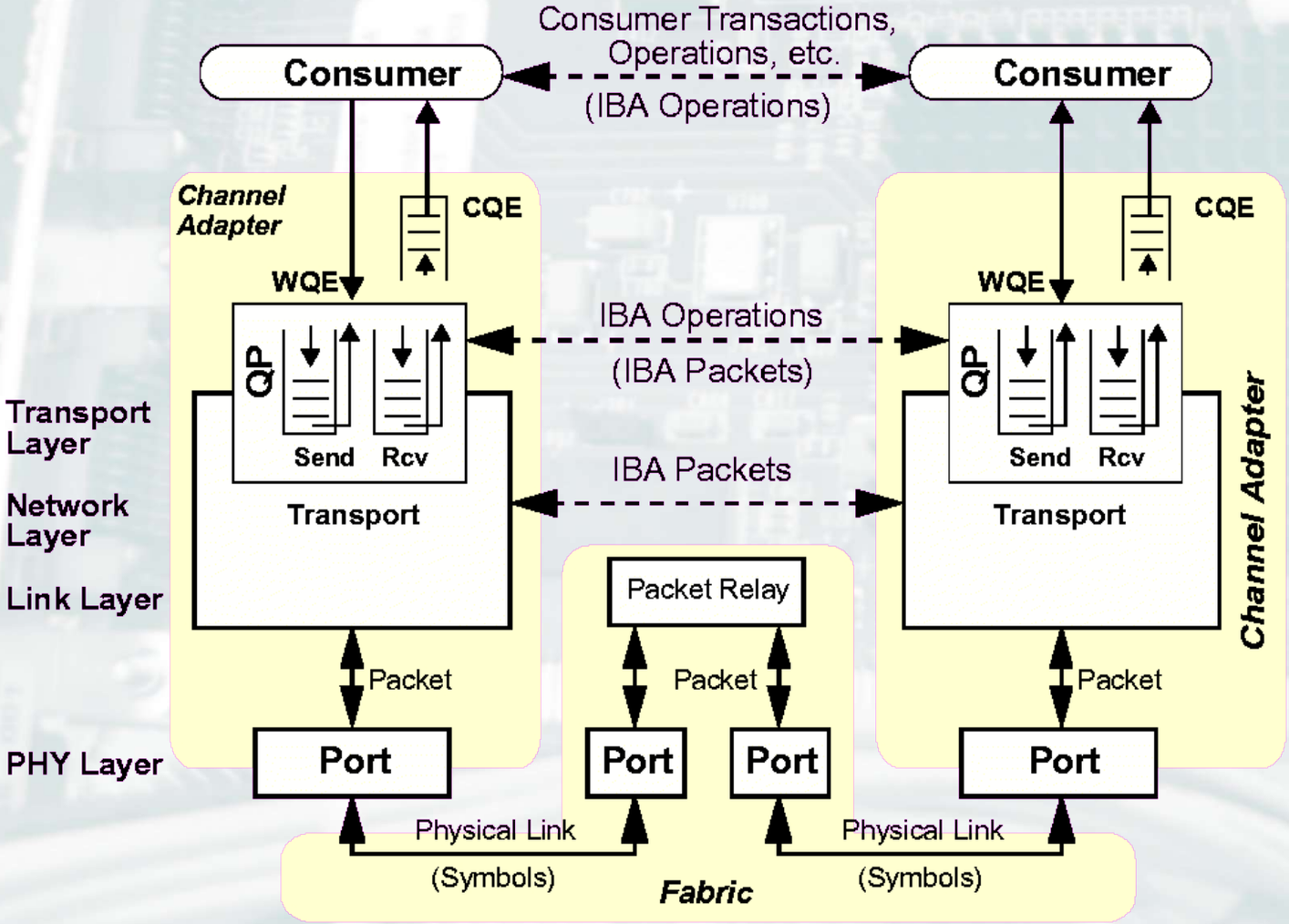
Kanalorientierte Verbindungen: Queues und Queue Pairs



- 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

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Connections: Queue pair Konzept

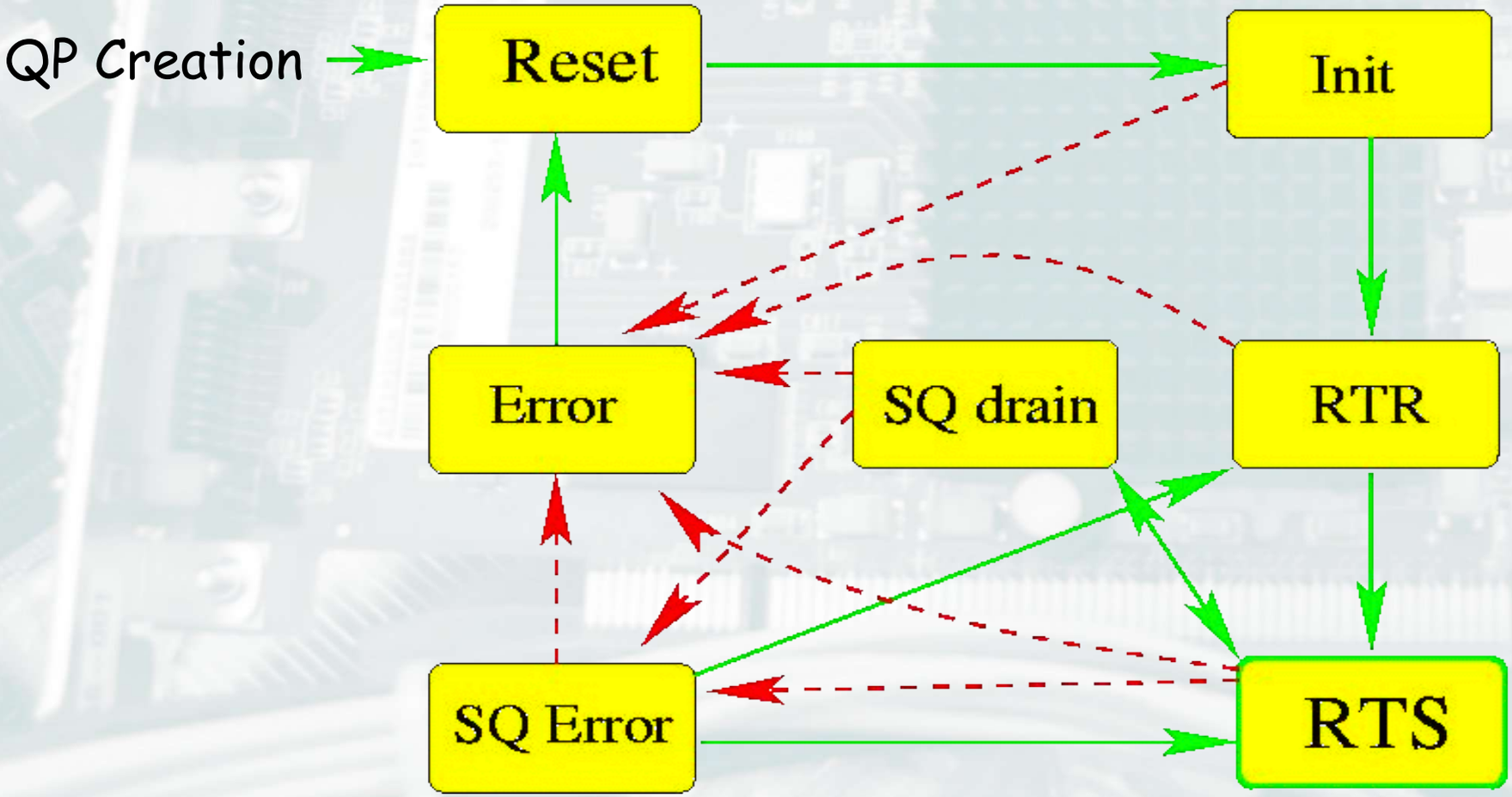


- a QP is associated with one (or more) remote QP's
- the send queue talks to the receive queue of the remote QP
- and vice versa
- error events occure as CQE

(taken from IBTA InfiniBand Specifications)

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Queue Status: Überblick (aus Spezifikationen)



Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

upper level protocols (snapshot)

- 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
- SDP : Socket Direct Protocol
- NFS über RDMA : (selbsterklärend).
- DAFS, DAPL, und andere ....

http://openib.org http://infiniband.sourceforge.org

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Marktsituation

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

- 4x HCA's von verschiedenen Anbietern (meist mit Mellanox Chipsatz)
- Preise: fallende Tendenz, nach Rabatten fragen lohnt sich
- Switches: 4x, neuerdings mit 12x Ports, bis >100 Ports pro Switch
- 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
- Kupfer, neuerdings auch optische Verbindungen erhältlich (teuer)

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

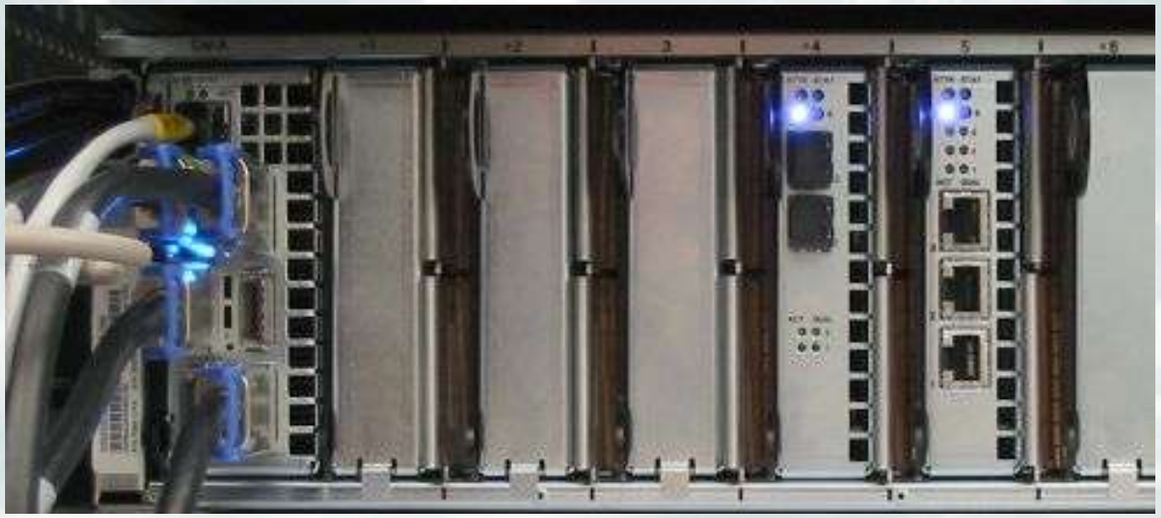
Infiniband-Projekt des IWR: Fabric Setup (seit Jan.2003)

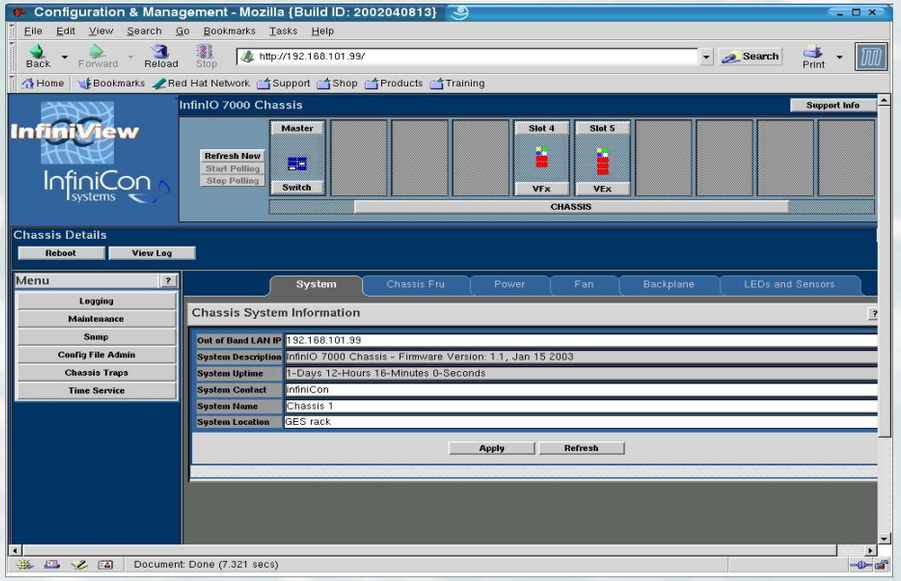
Infiniband Fabric Hardware:

- 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





Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

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)

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Latenzzeit mit OSU MPI Testprogramm

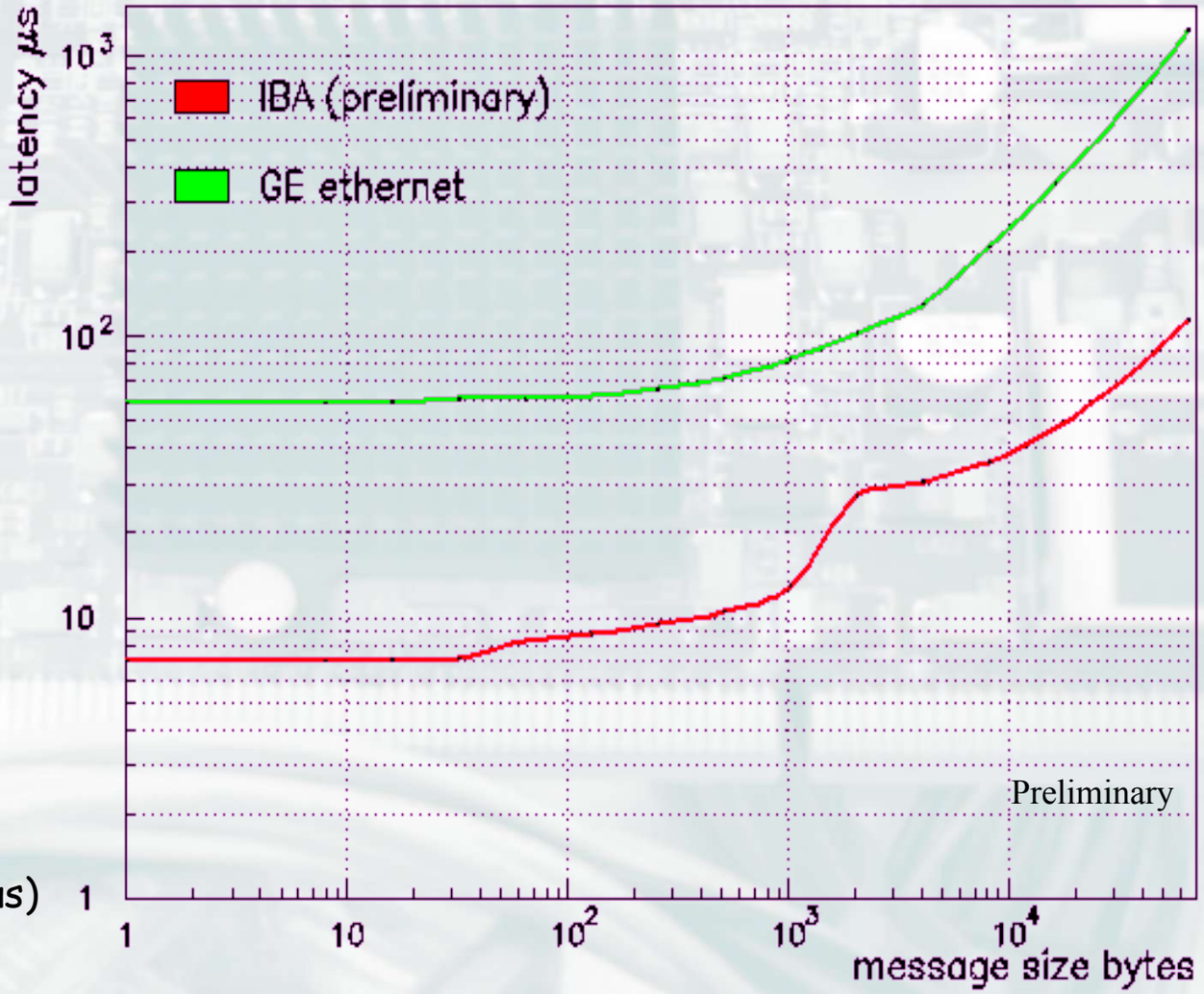
- ICS MPI, basiert auf MPICH 1.222 OSU,
- ICS Infiniband – Treiber
- Vergleich mit on-Board GE (Switched)

GigaBit Ethernet: etwa 60µs

(30µs vom Switch)

Infiniband : ca. 7µs

(auf IA64 und X86_64: Latenzzeit 5.5-6µs)



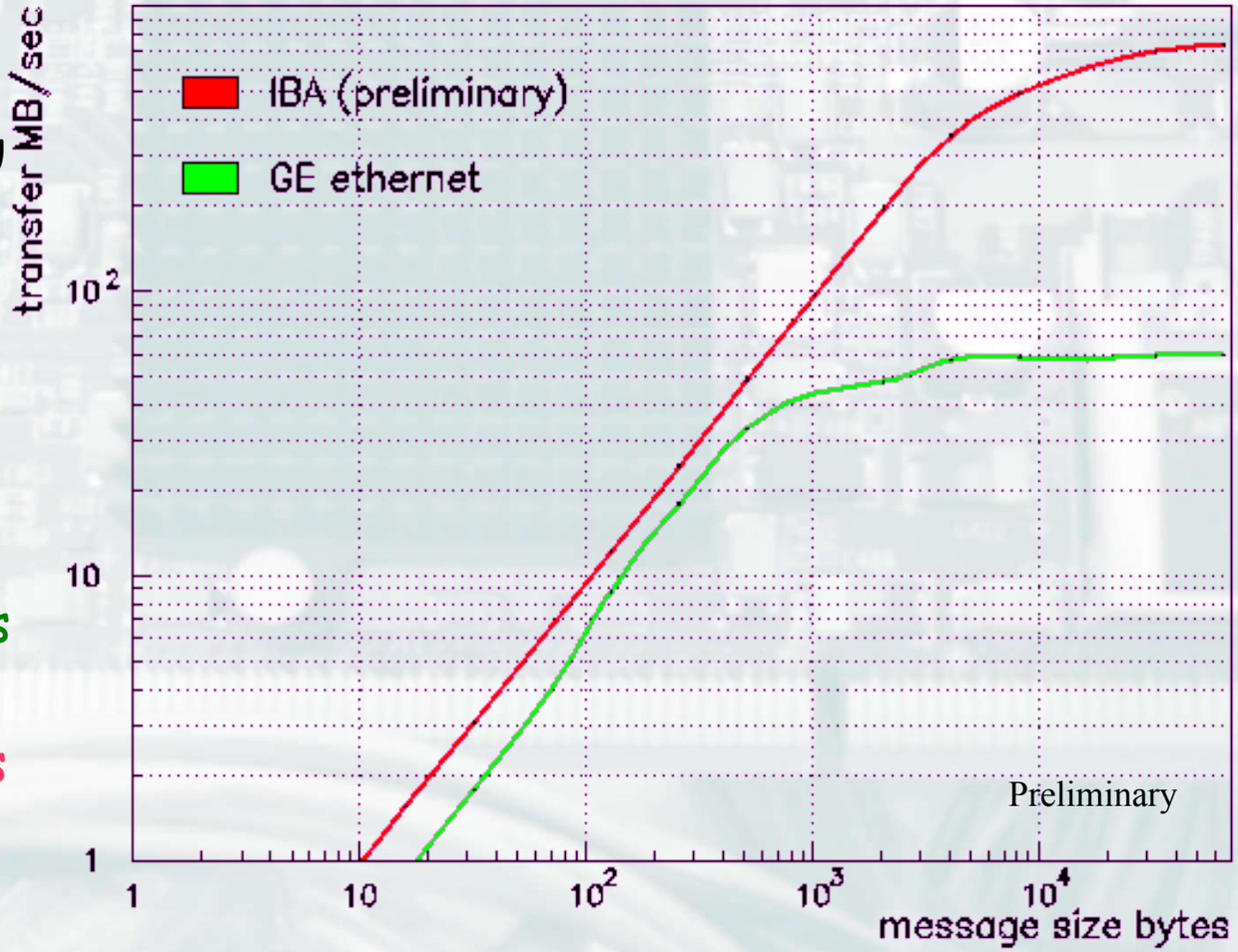
Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Bandbreitenmessung mit OSU MPI Testprogrammen

- ICS MPI, basiert auf MPICH 1.222 OSU
- ICS Infiniband – Treiber
- Vergleich mit GE (switched)

GigaBit Ethernet: bis zu 60MB/s

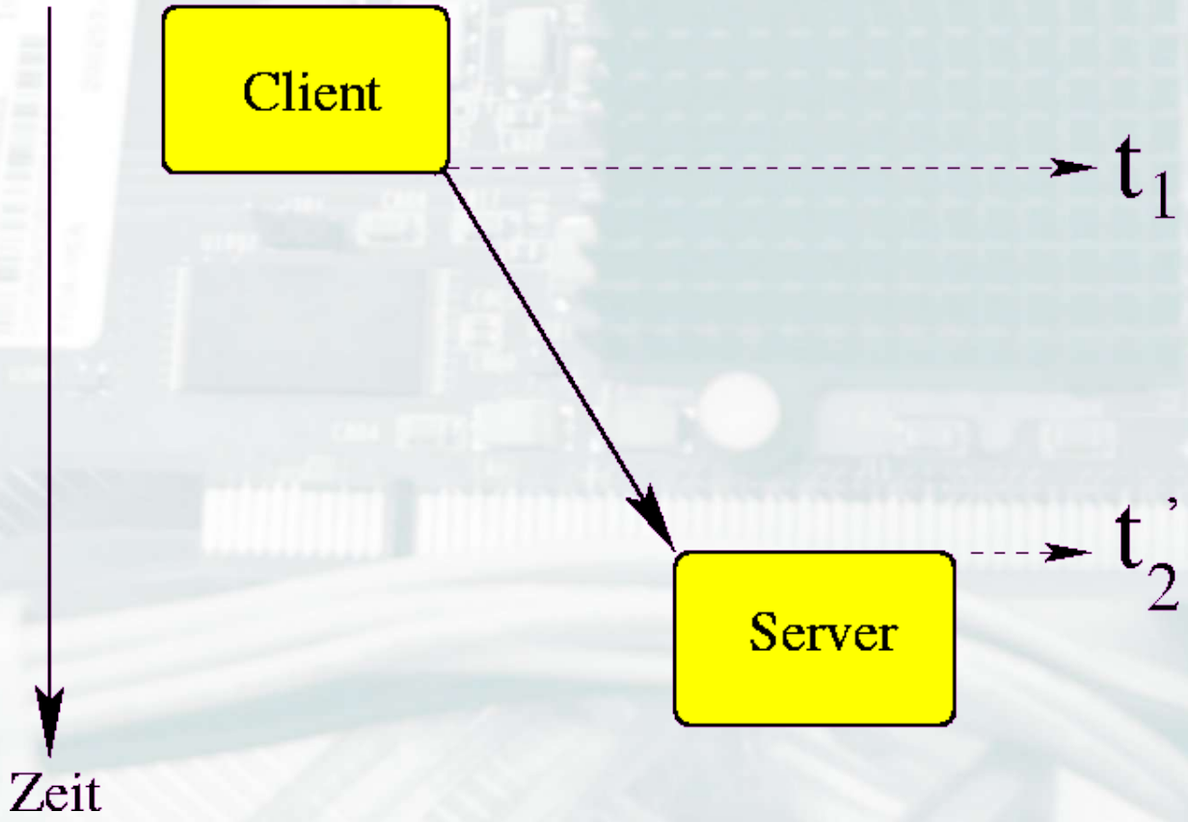
Infiniband : bis zu 780MB/s
(x86_64: bis 830MB/s)



Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

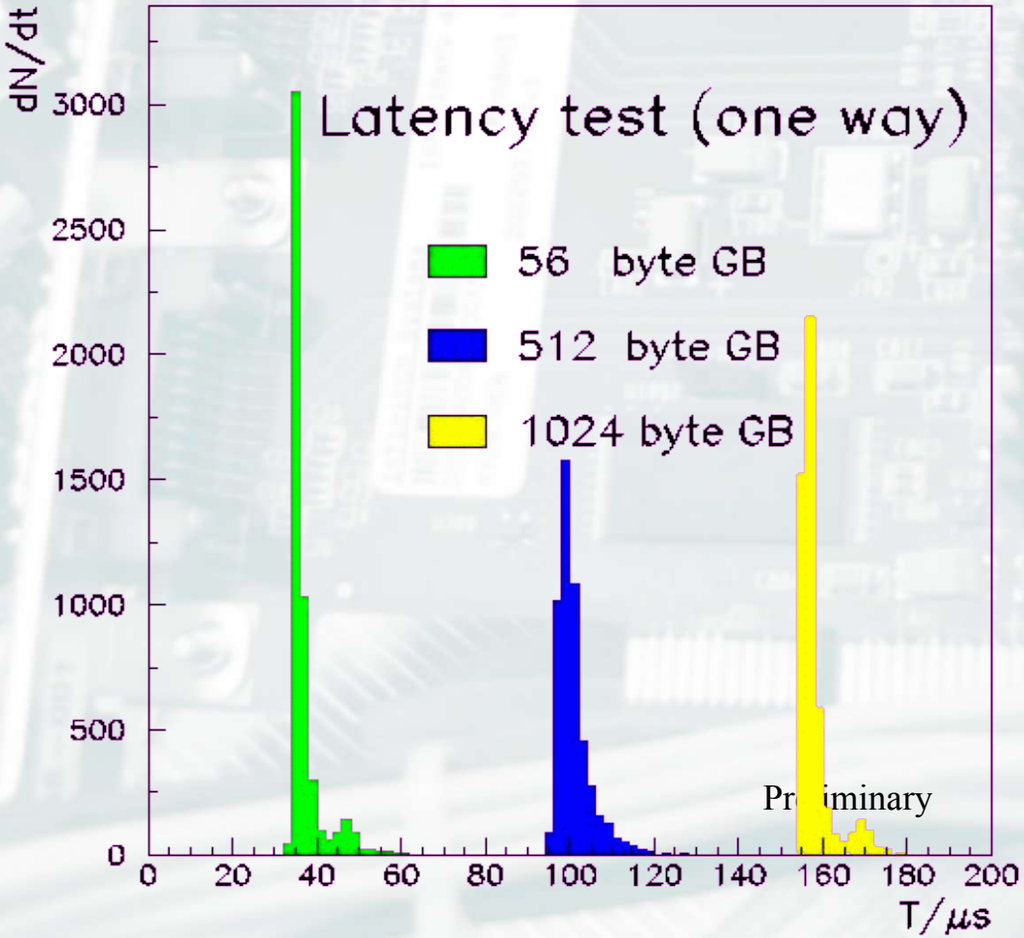
Eigenentwicklung: Latenzzeitmessungen

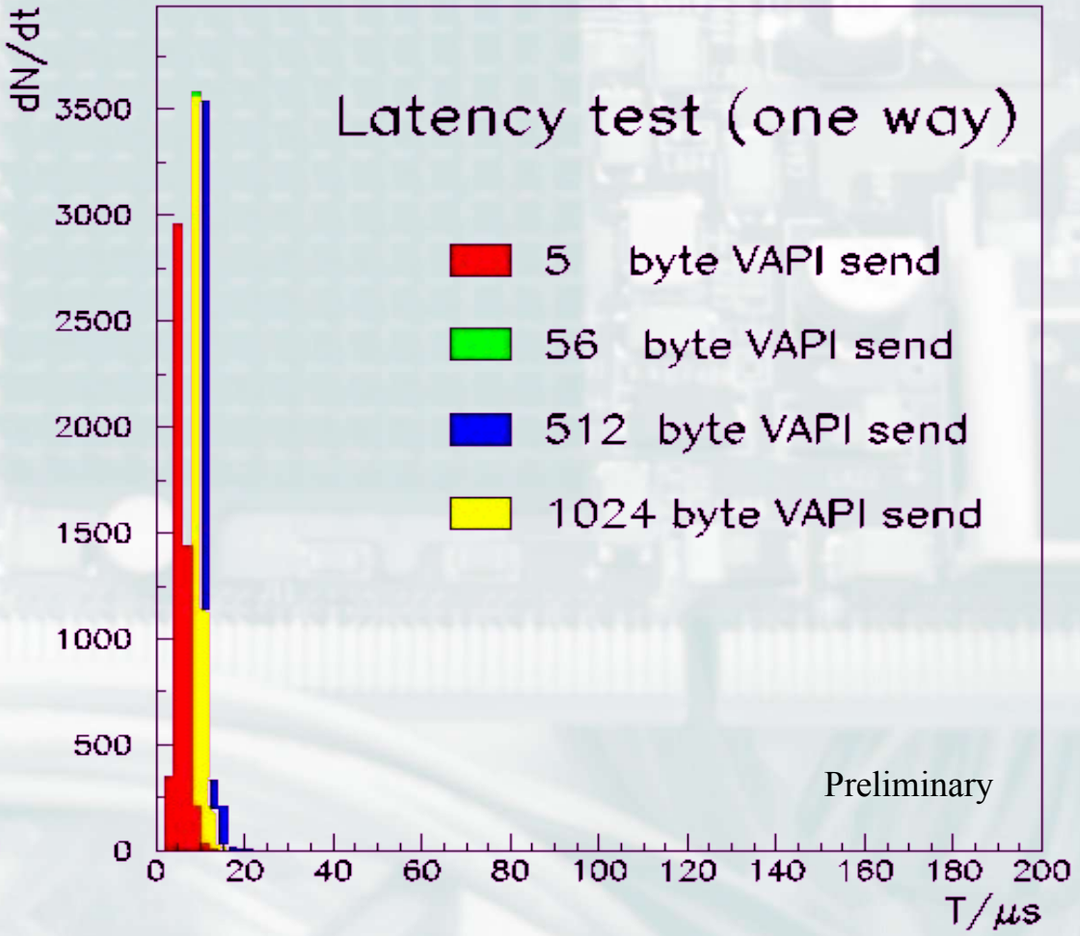
- Idee: direkte Messung der Signallaufzeit t2 – t1
- Erfordert exakte Synchronisation der Systemuhren



Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

"One way" Latency – Vergleich GE and Infiniband





Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Weiterer Testsystem – Ausbau: IWARP Projekt

Projektziele:

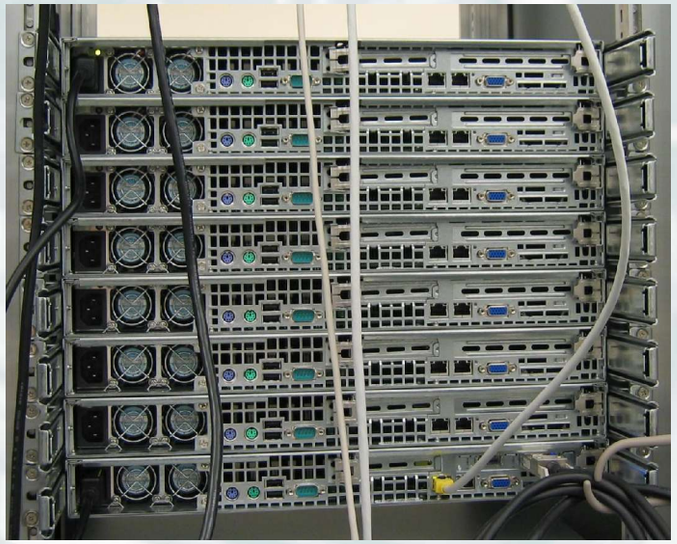
- Skalierbarkeitstudien
- Prototyp für größere Installation
- Benutzerbetrieb als MPI Parallelrechencluster
- Testsystem für Eigenentwicklungen

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Weiterer Testsystem – Ausbau: IWARP Cluster Projekt

Hardware setup:

- 13 Rechenknoten, 2.4GHz Dual Xeon, 533MHz FSB, Supermicro – Boards
- 1HE, PCI-X 133MHz 64Bit (3.3V) Slot mit Riser Card
- Je 2 GB Hauptspeicher, 60GB IDE local disk
- InfiniCon InfiniServ7000 4X Infiniband-Karten
- 16 Port Mellanox Switch (Referenzdesign)



Status:

- Seit September in Betrieb
- Erste Tests: MM5, HPL Benchmark, Lattice QCD

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

IWARP Test Cluster 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 Vanilla 2.4.25
- Mellanox OS-SDK , MiniSM, OpenSM (Test)
- MPI OSU 0.92 mit IFC (FuL) und GCC
- InfiniBand-RFIO (Eigenentwicklung)
- NFS home directories (über fast-ethernet)

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Überblick Hardware

13x Dual Xeon 2.4GHz Knoten

16 port 4x-InfiniBand Switch (Mellanox)

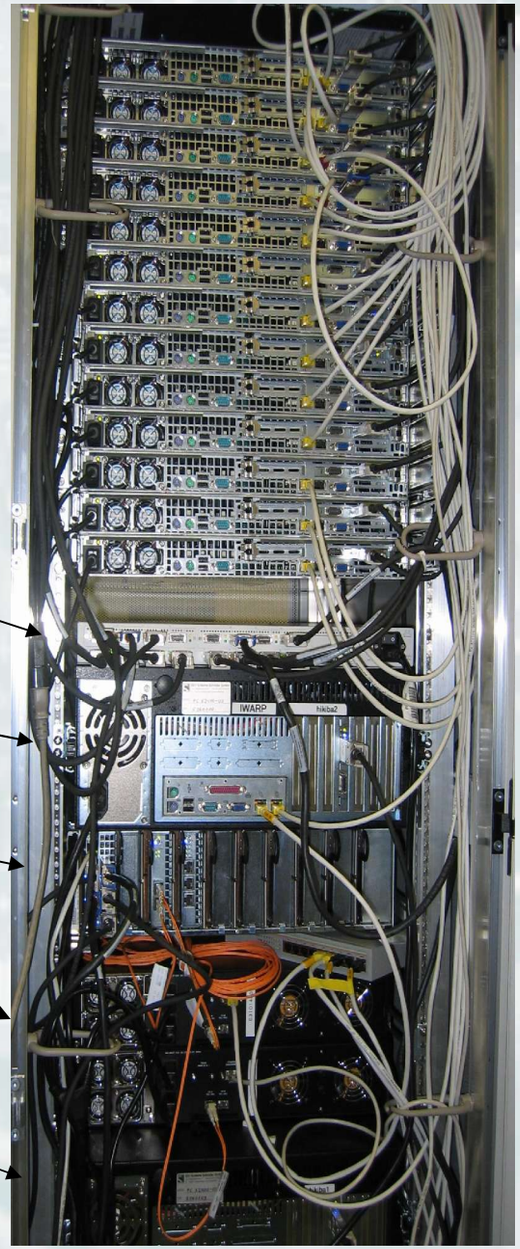
Vorrechner (IWARP) für logins

InfinIO-7000 (FC, GE, Switch)

2x IDE RAID Systeme mit 2GB FC

Testknoten (Dual Xeon 2.4GHz)

Schrankmanager/Subnetmanagerknoten



Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Erfahrungen:

MPI:

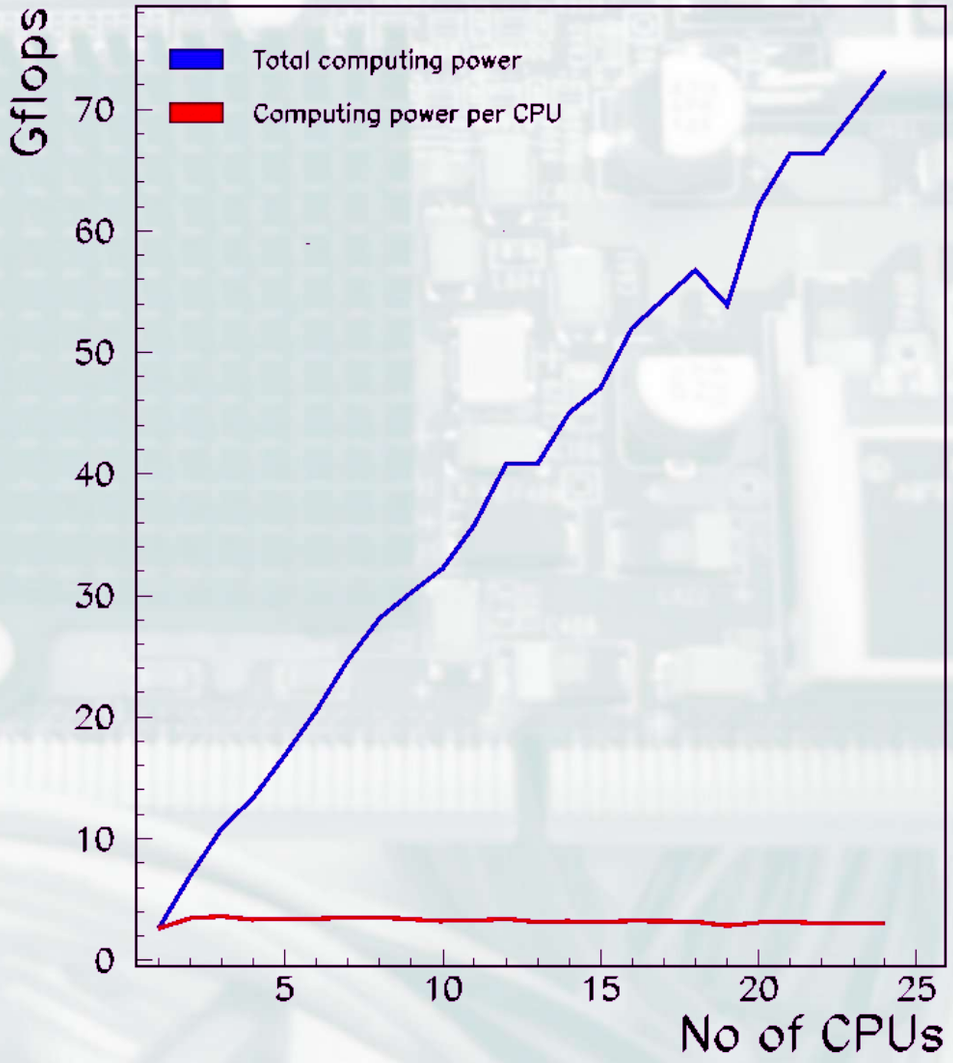
- HPL Benchmark (26 CPU's) 92 GFlops
- ca. 74% der theoretischen Performance
- gute Skalierbarkeit

Hardware:

- kleinere Kabelprobleme (gelöst)
- bislang keine Probleme/Ausfälle aufgetreten

Treibersoftware:

- Inzwischen OpenSource (BSD oder GPL)



Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

HPL Benchmark: 92.6 Gflop/s

$R_{max} / R_{peak} = 74\%$

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
W10R2C4        46000    96     2    13             700.66              9.262e+01
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) =        0.0012610 ...... PASSED
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) =        0.0027576 ...... PASSED
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) =        0.0005016 ...... PASSED
================================================================================
```

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Eigenentwicklung: rfio über Infiniband

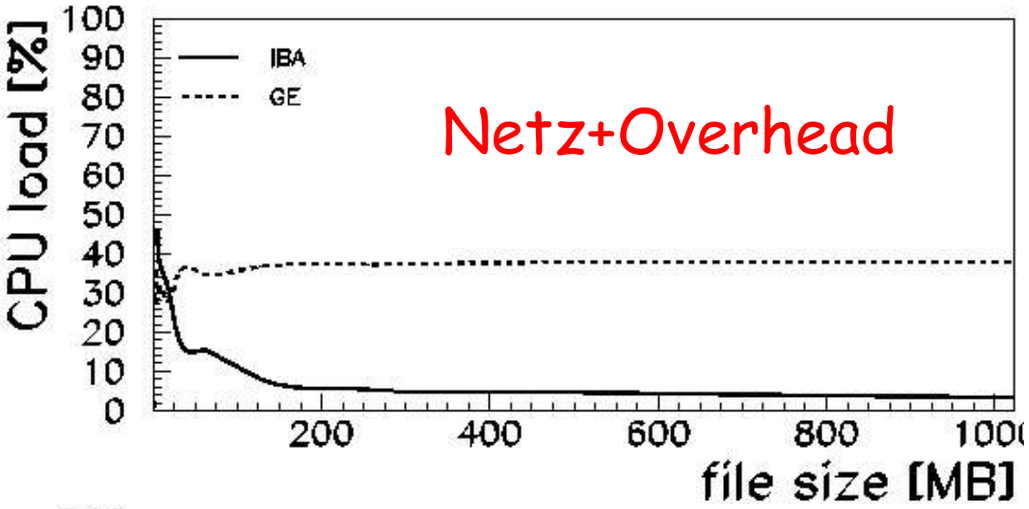
RFIO:
- effizientes Protokoll für Zugriff auf Files auf entfernten Rechnern
- entwickelt am CERN (seit 1990, SHIFT Projekt)
- heute Teil des CASTOR (HSM System) Projekt

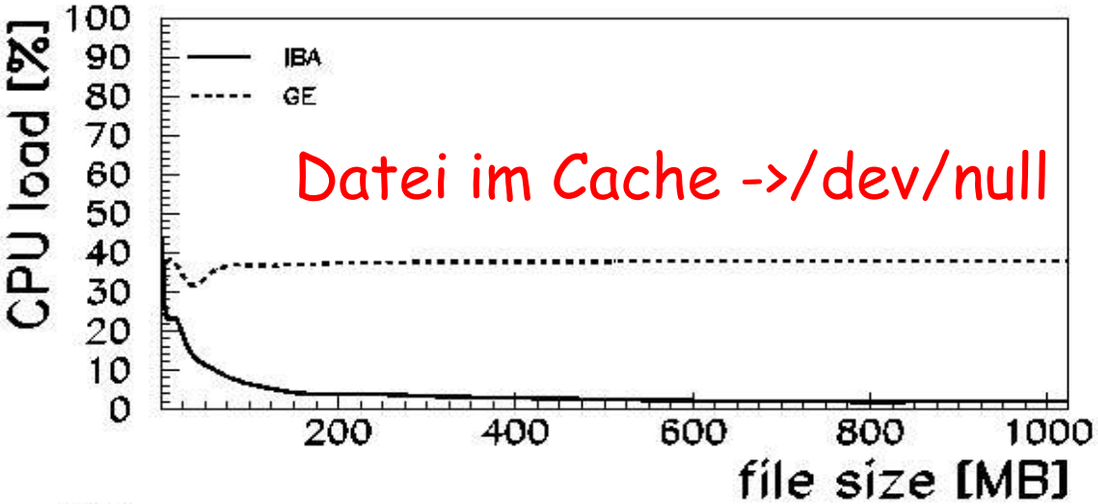
Idee:
- Verwendung von RDMA und RC Technik zur Datenübertragung
- Adressierung/Kontaktaufbau weiterhin über Ethernet (Transparenz!)
- Zusammenarbeit mit CERN (Gruppe Ari van Praag) und CASTOR Entwickler

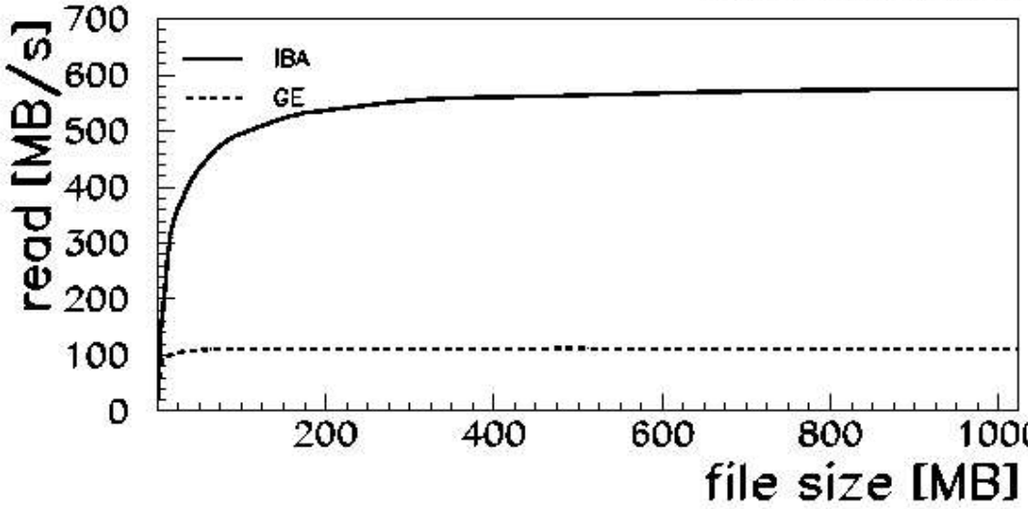
Vorteile:
- Transparent für Anwender
- Nutzung bereits bestehender Interfaces zu Programmen wie ROOT

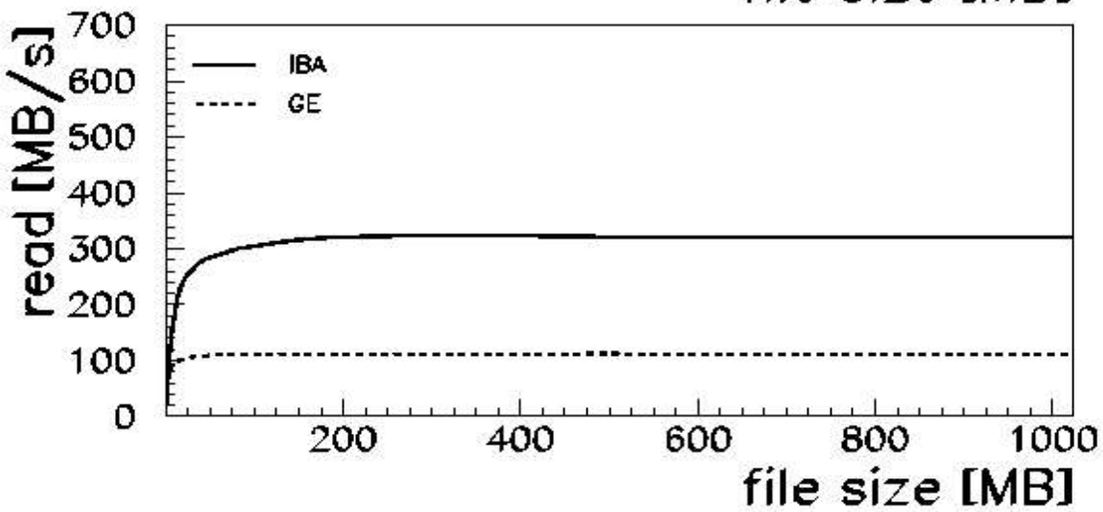
Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Erste Ergebnisse (ACAT03)









Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

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 580MB/s

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

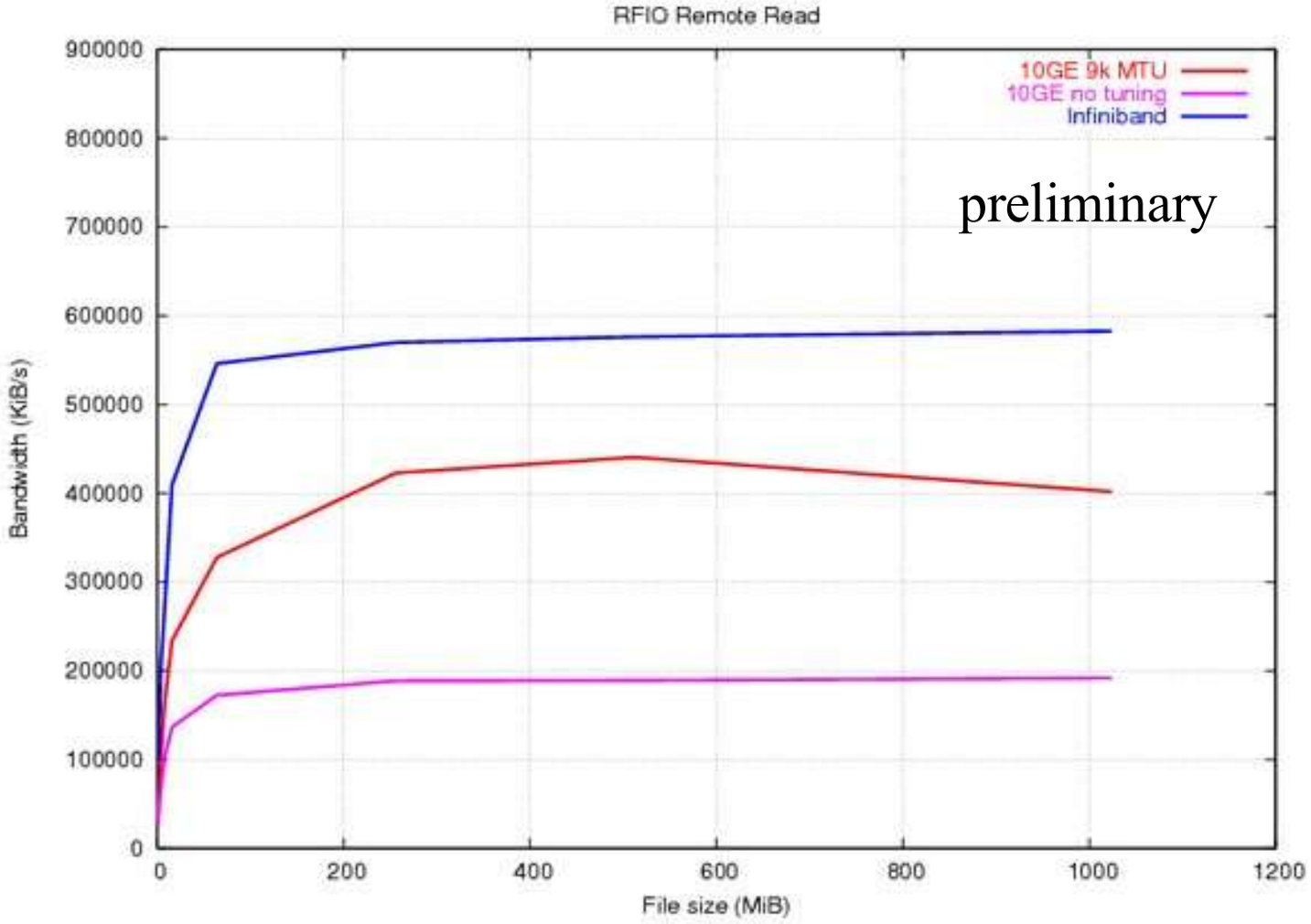
Vergleich mit 10 Gbit/s Ethernet: A.Horvath, CERN

10GE "heavy tuning" Parameter:

```
ifconfig eth2 mtu 9000
sysctl -w net.ipv4.tcp_sack=0
sysctl -w net.ipv4.tcp_timestamps=0
sysctl -w net.core.rmem_max=524287
sysctl -w net.core.wmem_max=524287
sysctl -w net.core.optmem_max=524287
sysctl -w net.core.netdev_max_backlog=300000
sysctl -w net.ipv4.tcp_rmem="10000000 10000000 10000000"
sysctl -w net.ipv4.tcp_wmem="10000000 10000000 10000000"
sysctl -w net.ipv4.tcp_mem="10000000 10000000 10000000"
sysctl -w net.ipv4.tcp_tw_recycle=1
sysctl -w net.ipv4.tcp_tw_reuse=1
```

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

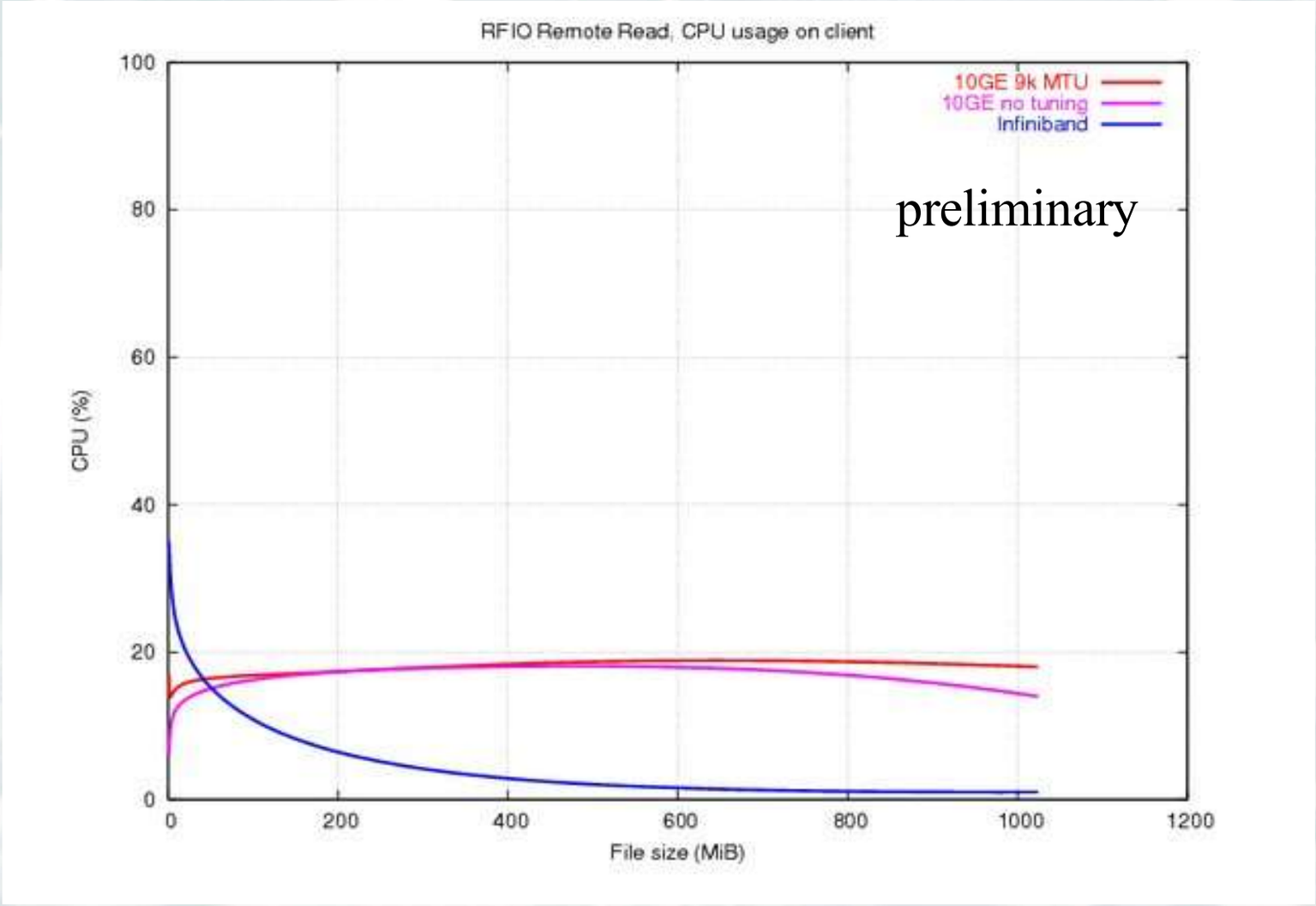
rfcp: READ Performance Vergleich mit 4x (10Gb/s) IBA mit 10GE



Credits: A. Horvath (CERN)

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

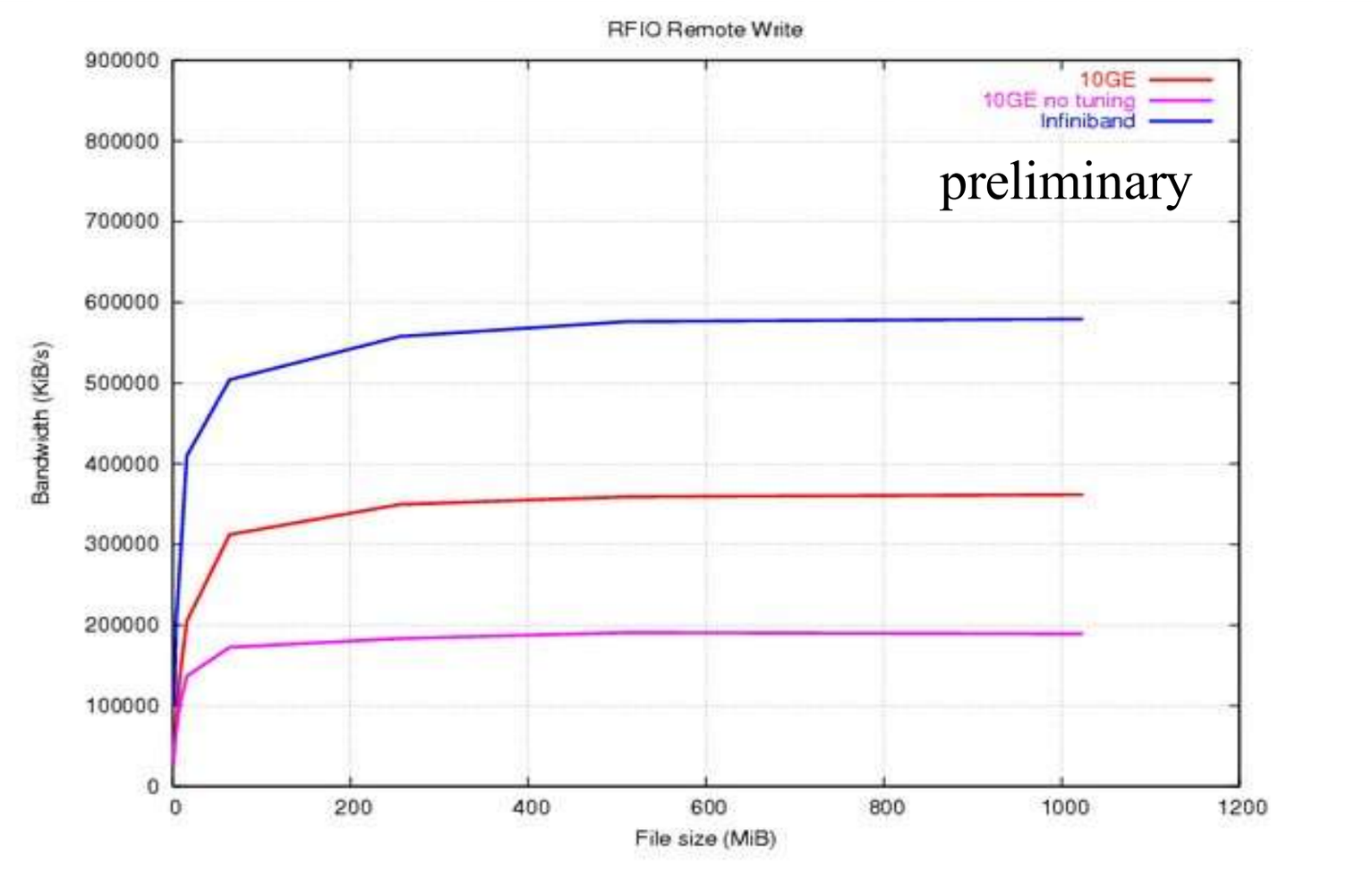
rfcp: READ Performance Vergleich mit 4x (10Gb/s) IBA mit 10GE



Credits: A. Horvath (CERN)

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

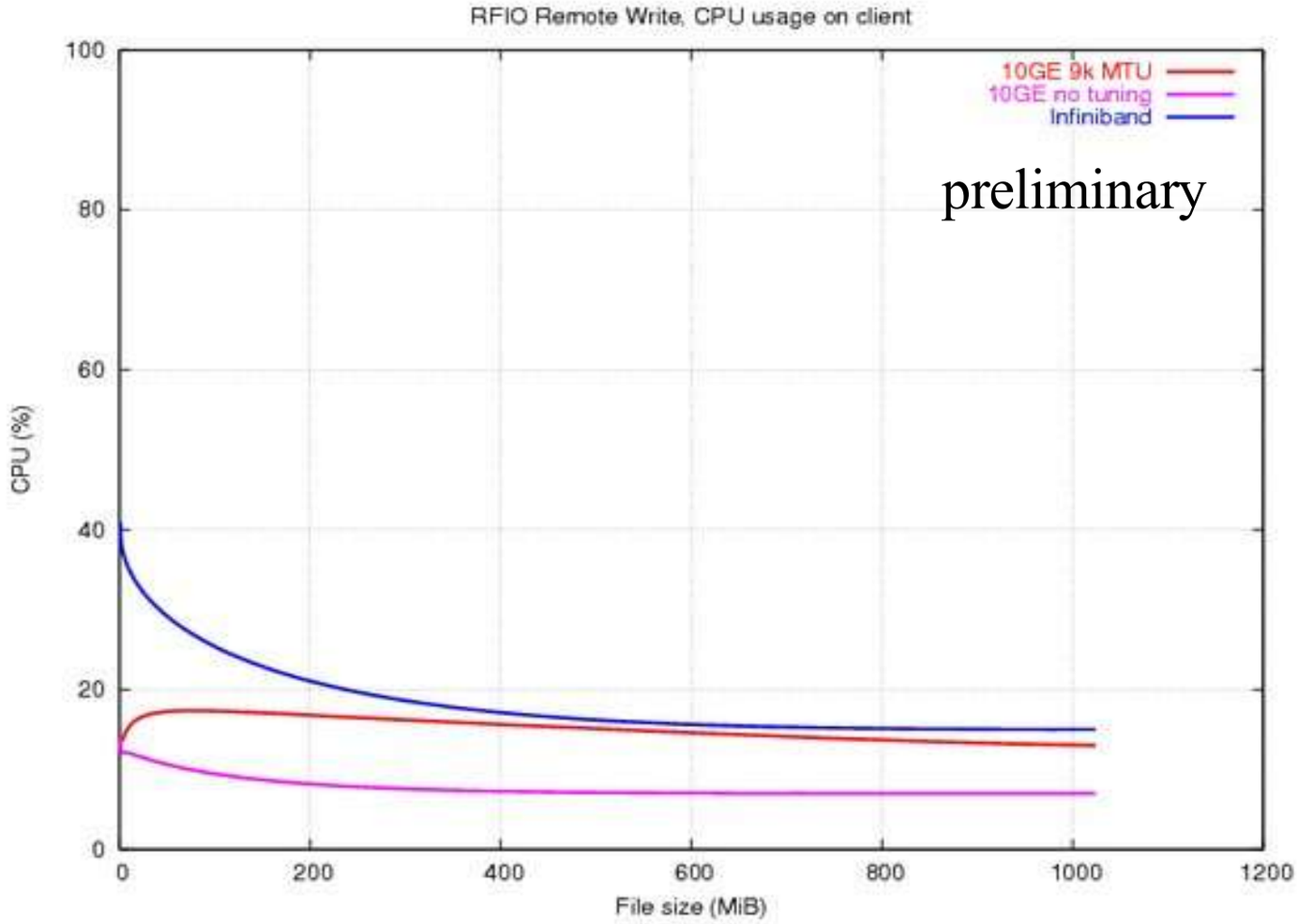
rfcp: WRITE Performance Vergleich mit 4x (10Gb/s) IBA mit 10GE



Credits: A. Horvath (CERN)

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

rfcp: WRITE Performance Vergleich mit 4x (10Gb/s) IBA mit 10GE



Credits: A. Horvath (CERN)

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Einige Erfahrungen

- Probleme mit einigen Kabeln
- PCI-X (133MHz) über Risercard können nicht alle Boards
- MPI über SMP könnte besser sein
- Lizenzpolitik war ziemliches Ärgernis für Eigenentwicklungen und Kernelupdates
- Seit Umbau im Winter läuft IWARP stabil und ohne größere Probleme
- Performance bei User-Applikationen:
  - Lattice QCD Applikationen teilweise sehr schöne Skalierbarkeit
  - andere (parform) nicht so toll (wird untersucht)
  - MM5 Benchmark: erste Tests ok, sollte wiederholt werden

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Lizenzpolitik:

- bislang restriktive Lizenzpolitik einiger Hersteller (NDA etc)
- inzwischen OS Versionen fast aller Hersteller erhältlich
- starke Unterstützung für InfiniBand SF Projektes und openib.org
- Bespiel: Mellanox SDK unter Wahlweise GPL oder BSD Lizenz
- ftp://ftp.mellanox.com

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

IWARP als PC Cluster im CampusGrid

- IWARP Cluster Projekt wird in den Userbetrieb überführt
- erreichbar innerhalb des CampusGrid Projektes für MPI Jobs
- derzeit 24 CPU's (12 Knoten) im Betrieb
- Umbau und Umzug sind abgeschlossen

Ausblick:

- Aufbau eines Opteronclusters mit InfiniBand im CampusGrid

Forschungszentrum Karlsruhe GmbH, IWR, Postfach36 40, 76021 Karlsruhe
Dr. Ulrich Schwickerath

Zusammenfassung

- zukunfsweisende, offene Technologie
- RDMA mit 10GB/s schon jetzt zu erschwinglichen Preisen
- 12x (30GB/s) und optische Verbindungen bald erhältlich
- Kinderkrankheiten sind vorhanden, werden aber weniger
- Hemmschuh Lizenzpolitik erledigt