Diskless Cluster und Lustre – Erfahrungsbericht zum CHiC

Frank Mietke, Torsten Hoefler, Torsten Mehlan
und Wolfgang Rehm

Fakultätsrechen- und Informationszentrum (FRIZ) /
Professur Rechnerarchitektur
Technische Universität Chemnitz

UNIX Stammtisch 24.04.2007

TECHNISCHE UNIVERSITÄT CHEMNITZ

Inhaltsverzeichnis

Hochleistungsrechnen

- Zusammenlegen von:
  - Prozessorleistung
  - Hauptspeicherkapazität
- High Performance Computing (HPC)
- High Throughput Computing (HTC)
- Anwendungen: Simulation und Datenverarbeitung
- Top500 Liste (öffentlich)

HPC und Linux in Chemnitz

- **C**hemnitzer **Li**nux **C**luster – CLiC (2000)
- **C**hemnitzer **H**ochleistungs **Li**nux **C**luster – CHiC (2007)

221,6 GFlop/s

8,21 TFlop/s

HPC Markt

- Mainstream Cluster
- Cutting-Edge Cluster
- Bleeding-Edge Cluster
- Vision

HPC ist Testfeld für neue Technologien

Projektstart

- Erste Gespräche Mitte 2003
- CHiC-Konsortium (24 Professuren)
- HBFG Antrag April 2004
- Antragssumme 2,64 Mio Euro
- Genehmigung, Ausschreibung und Aufbau 2006

Ziele des Projektes

- Mindestens 1000 Prozessoren
- Hochgeschwindigkeitsnetzwerk
- Eigenbau/-design
- Open Source Software
- Verbesserung der Stabilität des Systems (Festplatte, Netzteil, Hauptspeicher)
- „Geringer” Leistungsverbrauch ($<$ 200KW)

Bauprojekt

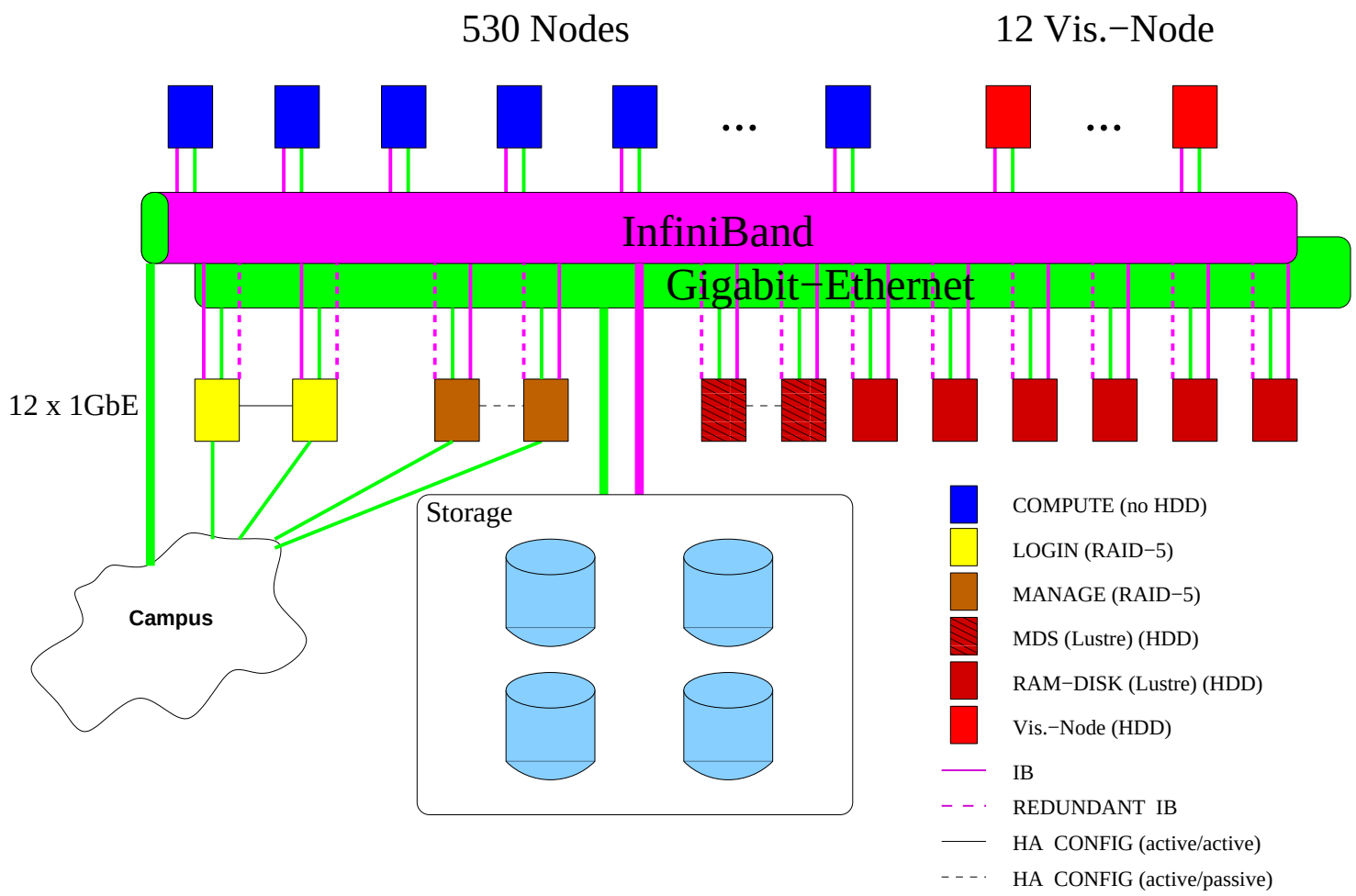
- Begrenzte Kühlleistung im Serverraum (–> wassergekühlte Schränke)
- Tragkraftproblematik
- Raumrekonstruktion ca. 1,7 Mio Euro

Raumaufbau

Clusteraufbau

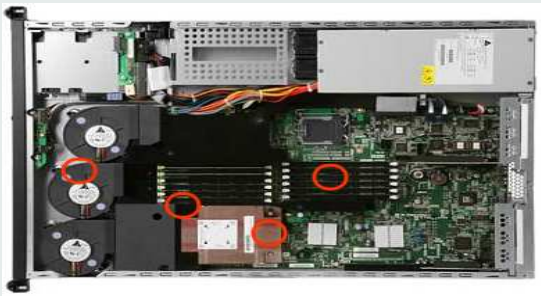
- 1 Monat Aufbauzeit
- 21,6 Tonnen Material (Schränke + Inhalt)
- 4200 Muttern und 4600 Schrauben notwendig
- 4900 Kabel mit 9800 Steckverbindern (8km Länge)
- 300 Manntage Aufwand

Clusterarchitektur



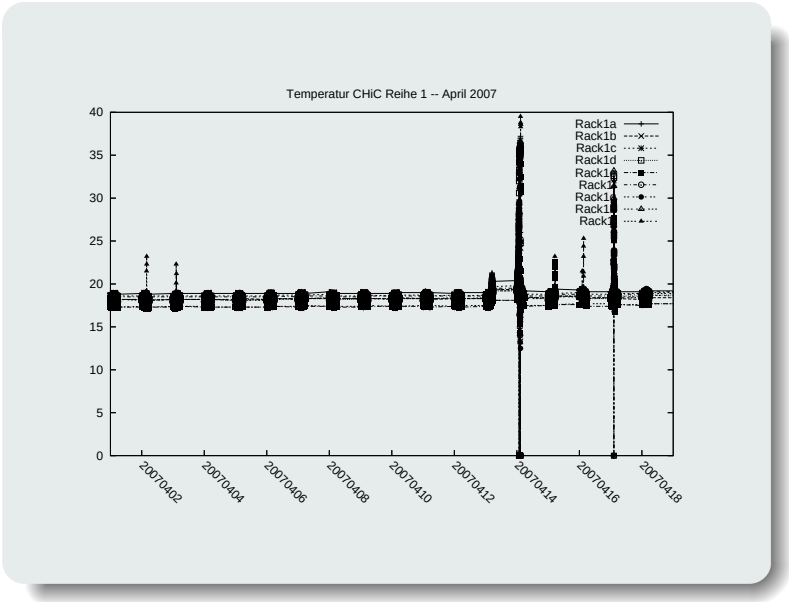
Cluster Hardware

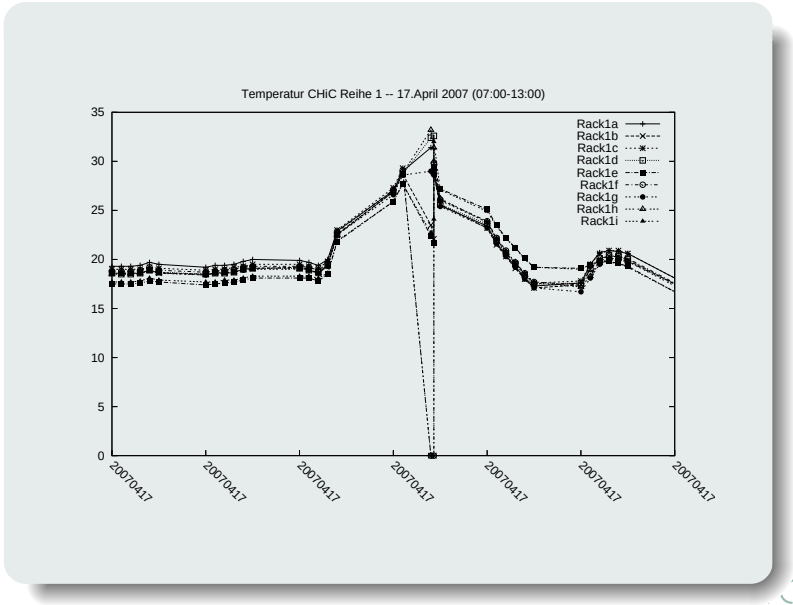
- IBM Server Technologie
- AMD Opteron Dual Core Rev. F
- Voltaire InfiniBand 10Gbit/s
- 288-Port IB-Switche
- NVIDIA Quadro FX 4500 X2



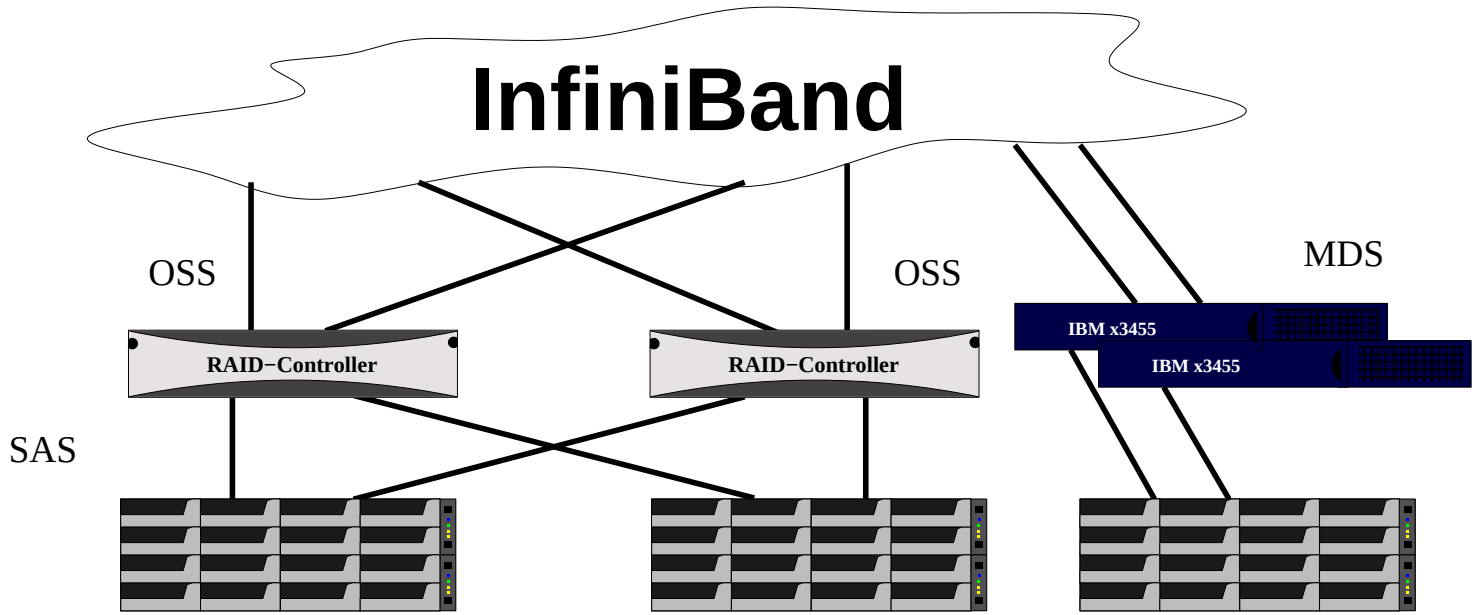
Cluster HW – Erfahrungen

- 5 Netzwerke (IPMI und PXE, IPMI und APC)
- Sehr stabile Knoten- und Infrastrukturtechnik
- Probleme IPoIB HA (ARP)
- InfiniBand-Ethernet Gateway (HW-Probleme)
- **Aber:** kompetenter Support



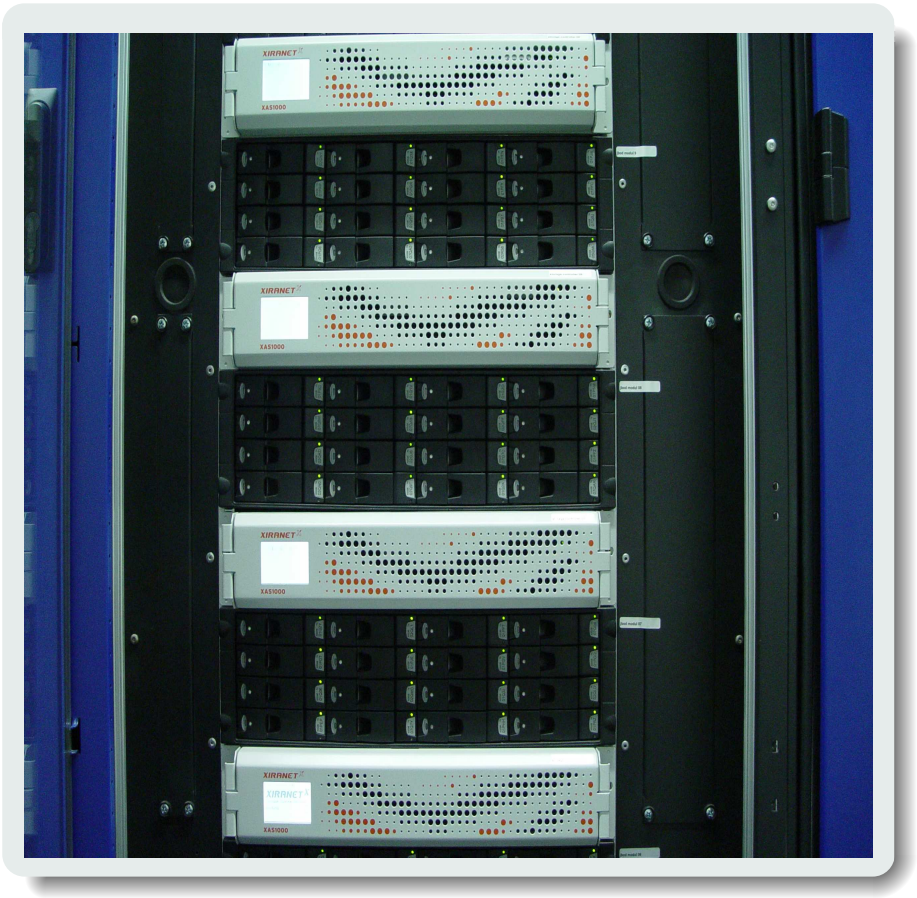


Storagearchitektur



5x

Storage Hardware



- 160x SATA (OSS)
- 16x SAS (MDS)
- 3,5GB/s (write)
- 1,7GB/s (read)
- RAID-5

Erfahrungen Storage Komplex

- Leistungfähige und stabile Hardware
- Lustre-Setup in 20 Minuten
- MDS/OSS HA funktioniert
- Fehlende RAID-6 Unterstützung
- Performancewerte Lesen moderat

Software–Umgebung

- Scientific Linux 4.4
- Open Fabcris Enterprise Ed. 1.1
- Lustre 1.6 Beta 7
- Open MPI und MVAPICH
- GNU Compiler, EKOPath Compiler
- LoadLeveler Batchsystem

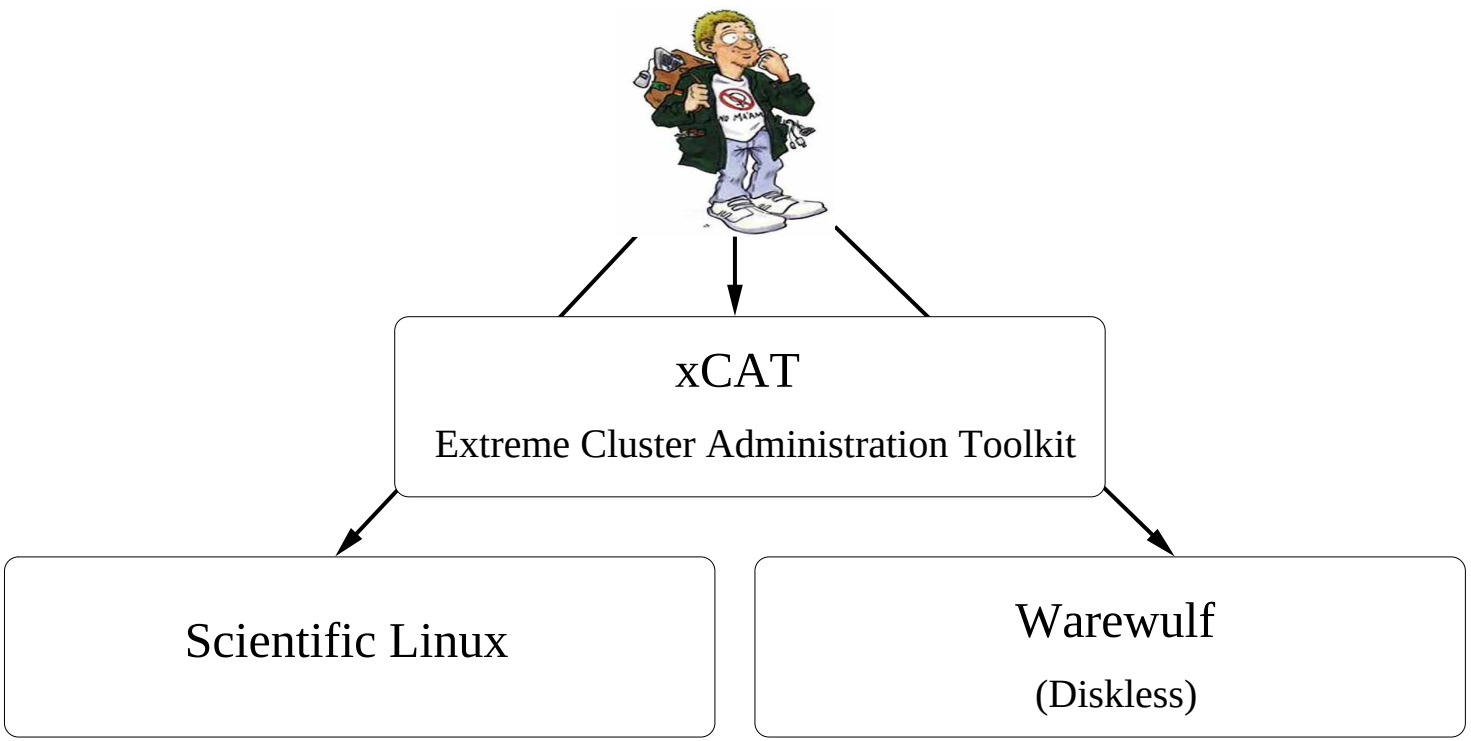
Modules Werkzeug

- Besseres Handling mit Umgebungsvariablen
- Einfache Handhabung / Erstellung von Modulen

```
$ module avail

$ module show <modulename>

$ module load/unload <modulename>

$ module initadd/initrm <modulename>
```

Management

- xCAT (www.xcat.org)
- Warewulf (www.warewulf-cluster.org)
- Scriptsammlungen/Konfigurationsdateien
- schlechte Dokumentation



Schritte zur Inbetriebnahme (xCAT) – 1/3

1. Definition des Clusters:
   - Knotennamen, Netzwerke usw.
   - 20+ Konfigurationsdateien
   - XX angepasste Scripte
2. Konfiguration der Switchtechnik:
   - Ethernet
   - InfiniBand (Switch + Gateway)
   - Terminal Server
   - Global Console Server
3. Sammeln der MAC-Adressen (Ser.Cons. oder Switch):
   - `$getmacs <noderange>`
   - `$makedhcp -allmac`

Schritte zur Inbetriebnahme (xCAT) – 2/3

1. Installations- und Imagetemplates:
   - SL-4.4
   - Warewulf
2. BIOS/FW Update 500+Knoten (IPMI)
3. Warewulf Knotenimage (Größe)
4. Serverskalierung (Dienste)
   - Offene TCP-Verbindungen
5. Synchronisierung der Knoten

Schritte zur Inbetriebnahme (xCAT) – 3/3

1. Paralleles Dateisystem (Lustre)
   - Aufsetzen der RAIDs
   - Konfiguration des Metadatenserver (RAID-10 + Aktiv-Passiv HA)
   - Konfiguration der Object Storage Server (RAID-5 + Aktiv-Aktiv HA)
2. Software:
   - Compiler
   - MPIs
   - Mathe-Bibliotheken
   - kommerzielle Nutzer-Software

Monitoring

Fehler sind die Regel!

- Nagios (Client/Server)
- Zugriff über SSH/IPMI/Ser.Console/APC
- Auftretende Lastspitzen im Managementknoten
- Kommunikationsaufkommen beachten

Debugging und Profiling

- Spezielle Übersetzungen von MVAPICH1/2 und Open MPI
- mpirun mit Debugger (gdb,ddd)
- valgrind mit MPI Unterstützung
- Prozessor Performance Counter (PAPI)
- mpiP, gprof, Frysk usw.
- PathOpt2

Higher Performance Cluster

- 12+ TFlops (einfache Genauigkeit)
- www.gpgpu.org

InfiniBand Projekte

- MPICH2 InfiniBand Device (CH3)
- MySQL InfiniBand Transporter
- Switchfreie InfiniBand Topologien
- Opteron-Cell Hybride und InfiniBand
- HSM über InfiniBand (RAM -> SATA)

Danke für Ihre Aufmerksamkeit

??Clusterführung??