RECHENZENTRUM der UNIVERSITÄT MANNHEIM

UNIVERSITÄTS-RECHENZENTRUM URZ

Kopplung der bwGRiD Cluster

von

Heidelberg und Mannheim

Steffen Hau, Heinz Kredel, Sabine Richling, Rolf Bogus

ZKI AK Supercomputing, zib, FU Berlin, Oktober 2009

Zentren für Kommunikation und Informationsverarbeitung e.V.

Inhalt

- Heidelberg und Mannheim
- bwGRiD Projekt
  - Hardware, Software und Betrieb
  - Speicher-System
- Kopplung der beiden Cluster
  - Hardware: InfiniBand über DWDM
  - Performance: MPI Benchmark
  - Organisation, Betrieb und Administration
- Zusammenfassung

Universität Heidelberg

- 12 Fakultäten
- Naturwissenschaften, Medizin, etc.
- 28.000 Studenten
- Rechenzentrum:
  bisher Konzentration auf IT-Dienstleistungen
  Hochleistungsrechner in Instituten (z.B. IWR)

Universität Mannheim

- 5 Fakultäten
- Betriebs- und Volkswirtschaft
  - mit Wirtschafts-Informatik
- Sozialwissenschaften
- 11.000 Studenten
- Rechenzentrum:
  Erfahrung mit Hochleistungsrechnern

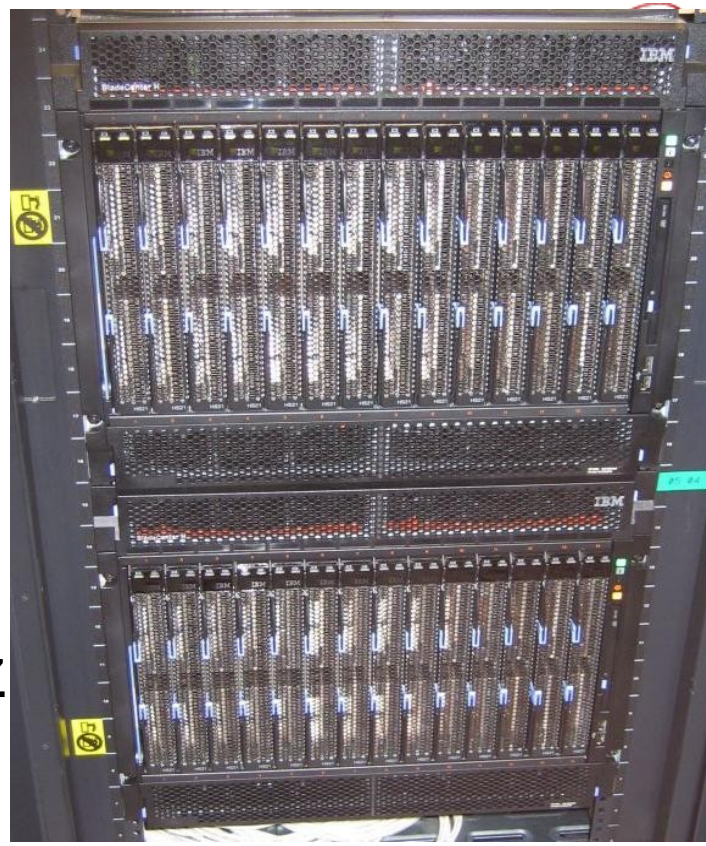
bw-GRiD Cluster

- Projektantrag vom HLRS an BMFT in 2007
- für D-Grid Infrastruktur an den Universitäten in Baden-Württemberg
- explizit als verteiltes System mit dezentraler Verwaltung
- an den Standorten
  - Stuttgart, Ulm (mit Konstanz), Freiburg, Tübingen, Karlsruhe, Heidelberg, Mannheim
  - für die nächsten 5 Jahre
- Konzeption und Beschaffung durch HLRS in 2008

bw-GRiD Ziele

- Nachweis der Funktionalität und des Nutzens von Gridkonzepten im HPC Umfeld
- Überwindung von bestehenden Organisations- und Sicherheitsproblemen
- Entwicklung von neuen Cluster- und Grid-Anwendungen
- Lösung der Lizenzproblematik
- Ermöglichung der Spezialisierung und Arbeitsteilung von Rechenzentren

Hardware

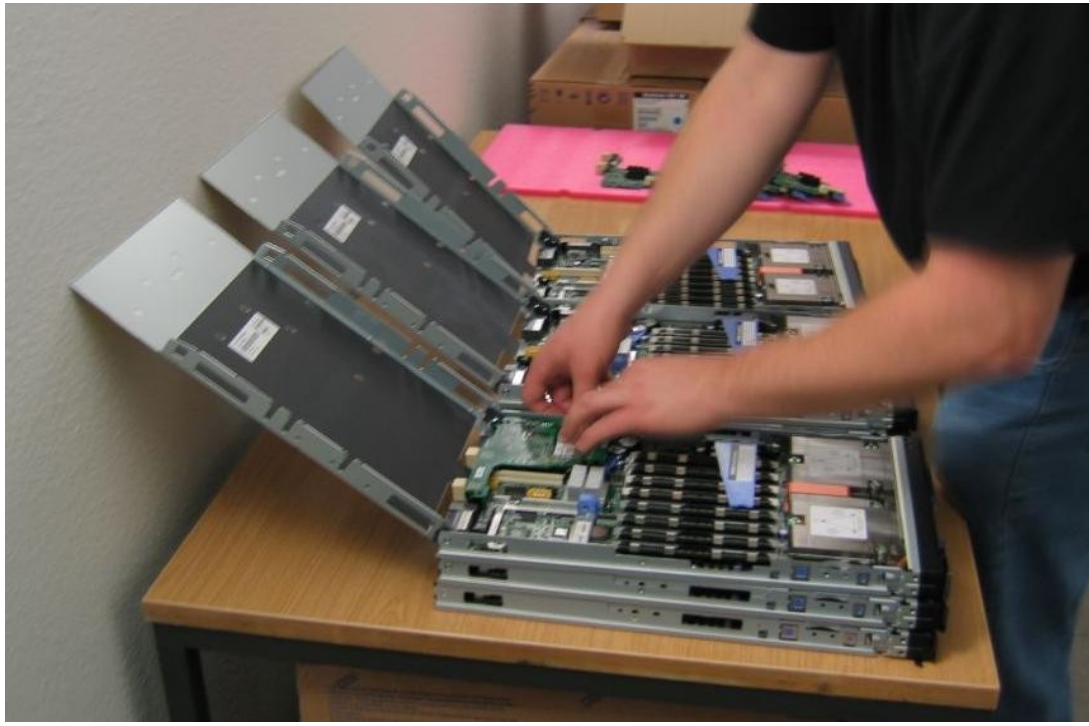


- 101 IBM Blade-Center Gehäuse
- mit je 14 IBM HS21 XM Blades
  - mit je 2 Intel Xeon CPUs, 2.8 GHz
    - mit je 4 Cores
  - 16 GB Hauptspeicher
  - keine Festplatten
  - Gigabit Ethernet
  - Infiniband Netzwerk

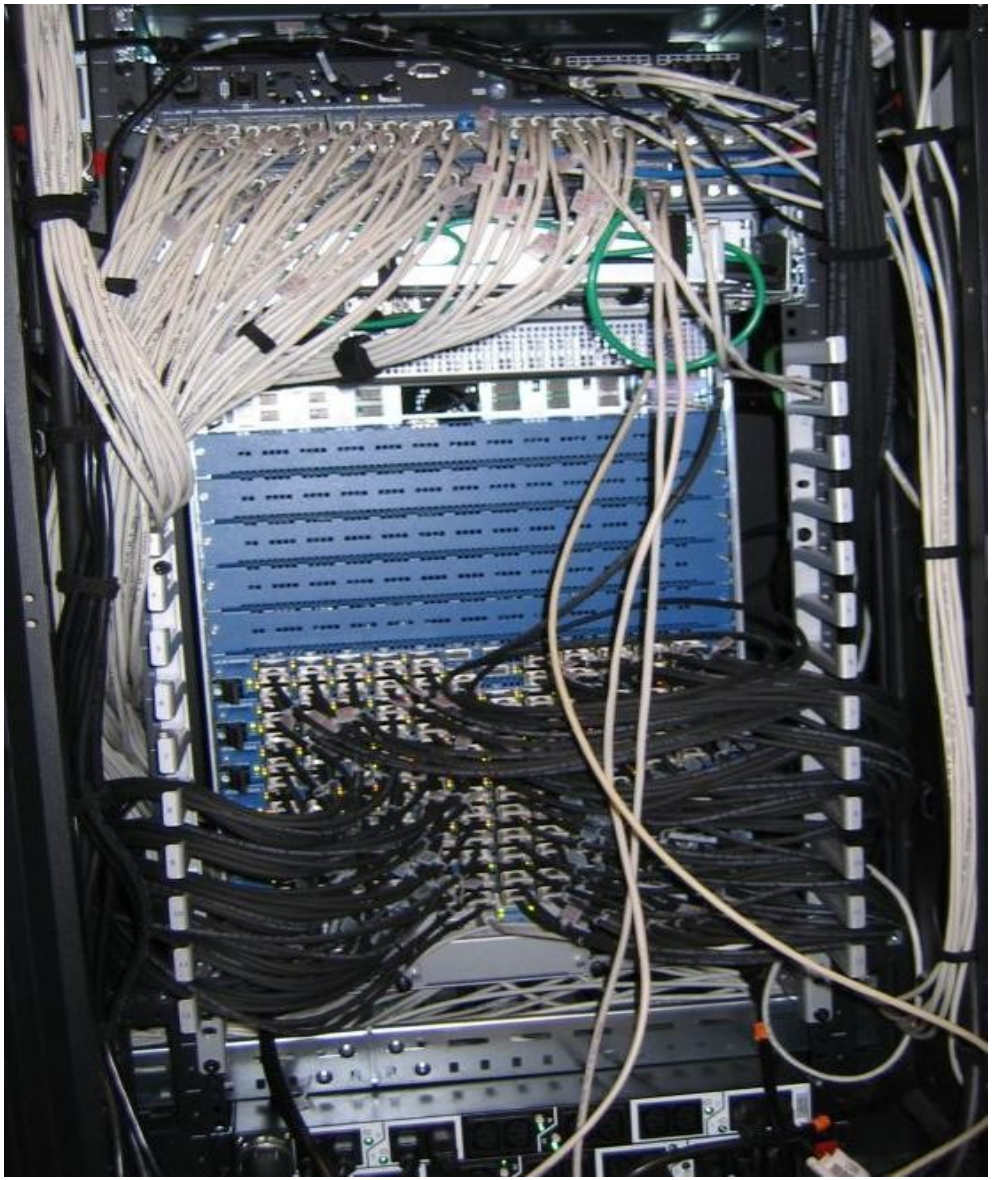
= 1414 Blades

= 11312 CPU Cores,

1120 Cores in Mannheim



InfiniBand



Software

- Scientific Linux 5.0
  - basierend auf RedHat 5.0 Quellcode
  - gepflegt u.A. vom CERN
    - dort sind ca. 140.000 Rechner für LHC im Einsatz
  - ähnlich CentOS
- Kernel und System Image vom HLRS vorbereitet
- booten über das Netz mit DHCP und TFTP
- Festplatten über NFS, jetzt auch mit Lustre
- Swap-Bereich in lokaler Festplatte
  - am Anfang minimaler Swap-Bereich über NBD

Anwendungen

	Mathematik	Biowissenschaften	Ingenieurwiss.	Wirtschaftswiss.	Chemie	Physik	Informatik
Freiburg	25% CFD		20% Mikrosystemt.				
Heidelberg	25% CAS, Ug	20% Comp. NeuroSc.					
Karlsruhe	10 % LinAlg		30% CFD, CSM				
Konstanz	x SPH	x Biochemie				x Theor. Ph, QM	x DataMining
Mannheim	15% CAS			30% CAS, Simulation			
Stuttgart			35% CFD, Comp.Mech.				
Tübingen		20% BioInfo				25% Astrophysik	
Ulm		5% BioInfo, Medizin			25% Theor. Ch, MD		

Alle: ca. 55% Betrieb, Middleware, Compiler, Tools

CFD = Computational Fluid Dynamics
CAS = Computer Algebra Systems
MD = Molecular Dynamics
QM = Quantum Mechanics

Projekte - Heidelberg

- **Theoretische Physik**
  QCD, Monte-Carlo und Molecular Dynamics
  Chromatin Folding, Bose-Einstein-Kondensate
- **IWR**
  - Molekulare Biophysik (Anionen-Transport)
  - Computational Neuroscience (mit Software UG)
  - Parallele Löser für DG-Diskretisierung
- **Physikalische Chemie**
  Electronic structure, Molecular Dynamics
  Lipids, Proteins, Many-Body-Systems
- **Sonstige**: Universitätsklinikum, Computerlinguistik

Projekte - Mannheim

- **VWL**: Statistische Analysen zur Sicherheitspolitik, Generationenbeziehungen, Währungsdaten, ...
- **BWL**: Statistische Analysen zu Manageroptionen, impliziten Kapitalkosten, Insiderhandel, ...
- **Informatik/Mathematik**: Simulation von mobilen Netzwerken, Algorithmen für Tracking, ...
- **Material- und Geowissenschaften** (Darmstadt): Nanokristalline Materialien, Kristallwachstum
- **Medizin**: Statistische DNA-Analyse
- **UB**: Automatische Klassifikation von Dokumenten

Aufbau Anfang 2008

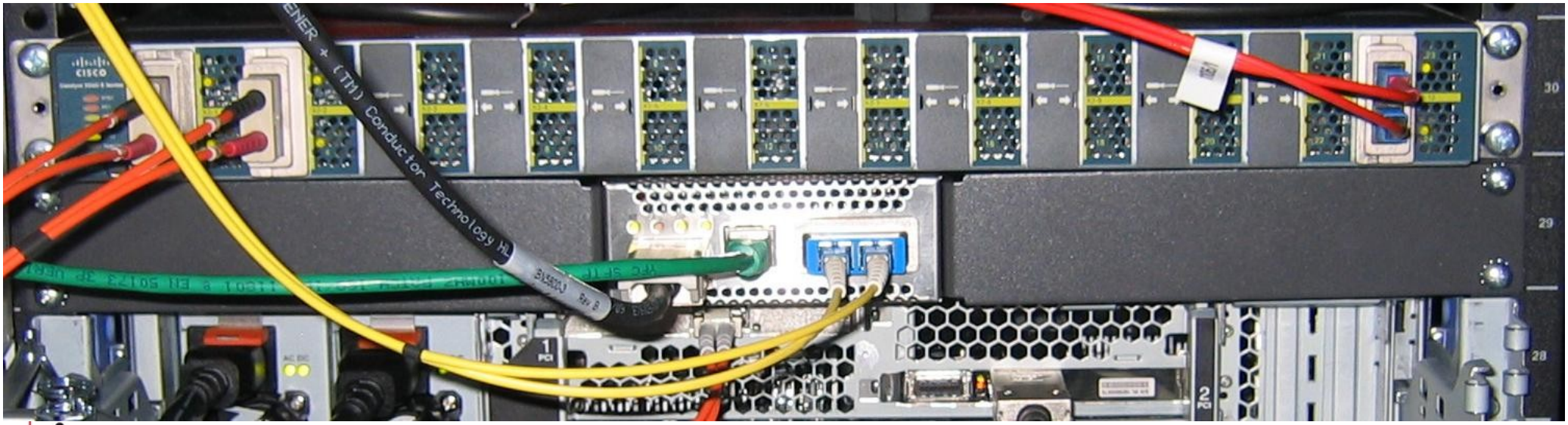
- ein gemeinsamer Admin für MA und HD
- Hardware
  - geliefert und aufgebaut, Januar bis März 2008
  - Netzwerk funktioniert, Ethernet und auch Infiniband
- Administration und Betrieb
  - Festplatten Platz für Home-Dirs über NFS-Server
  - Benutzerverwaltung ist konfiguriert, über LDAP
  - Batchsystem konfiguriert
  - in HD Nutzung der **Benutzerverwaltung und des NFS Speichers des IWR**

HP Lustre Speichersystem

- 32 TByte netto
- Paralleles Filesystem
- Seit Anfang 2009

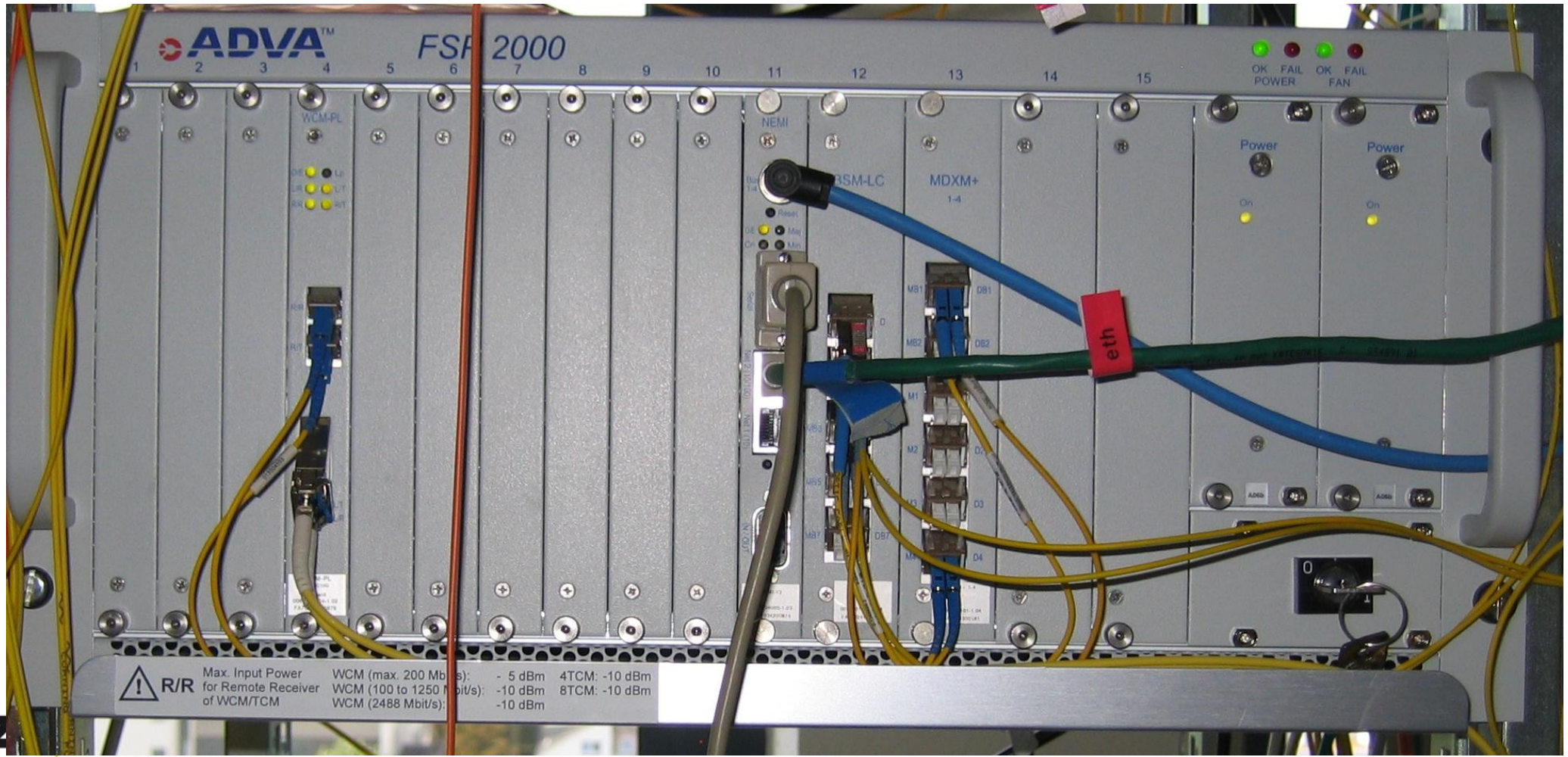
Kopplung mit Heidelberg

- Antrag in 2008
- Beschaffung und Aufbau bis Mai 2009
- in Betrieb seit Juli 2009
- InfiniBand auf Glasfaser: Obsidian Longbow



Kopplung - Hardware

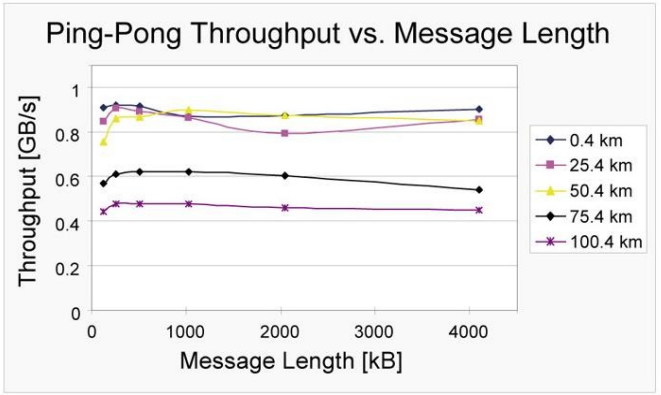
- ADVA Einspeisung DWDM Strecke

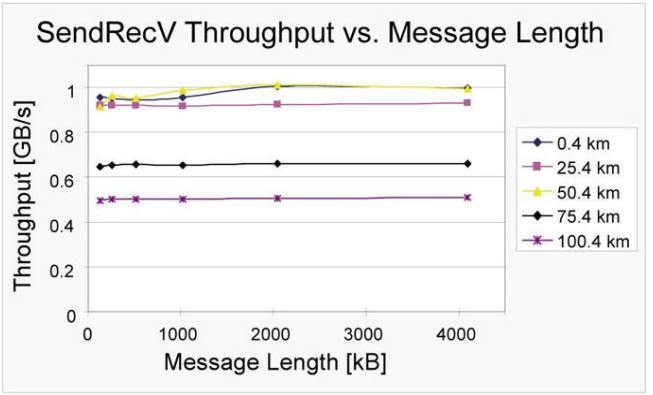


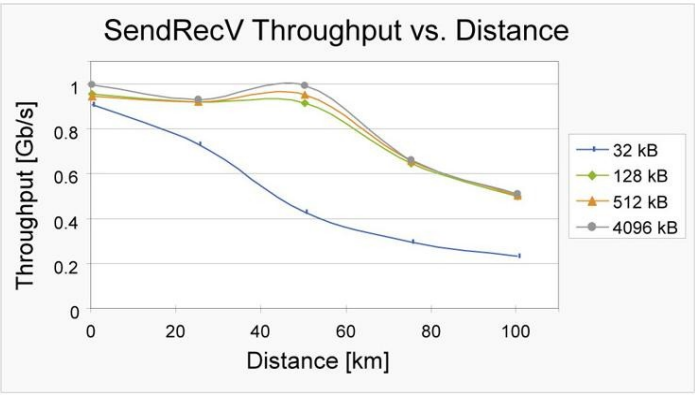
HLRS MPI Performance

- Messungen für verschiedene Entfernungen
- bis 50-60 km akzeptabel
- Bandbreite ca. 900 – 1000 MB/sec
- keine Angaben zur gemessenen Latenz

Measurement results – full InfiniBand throughput over more than 50km distance



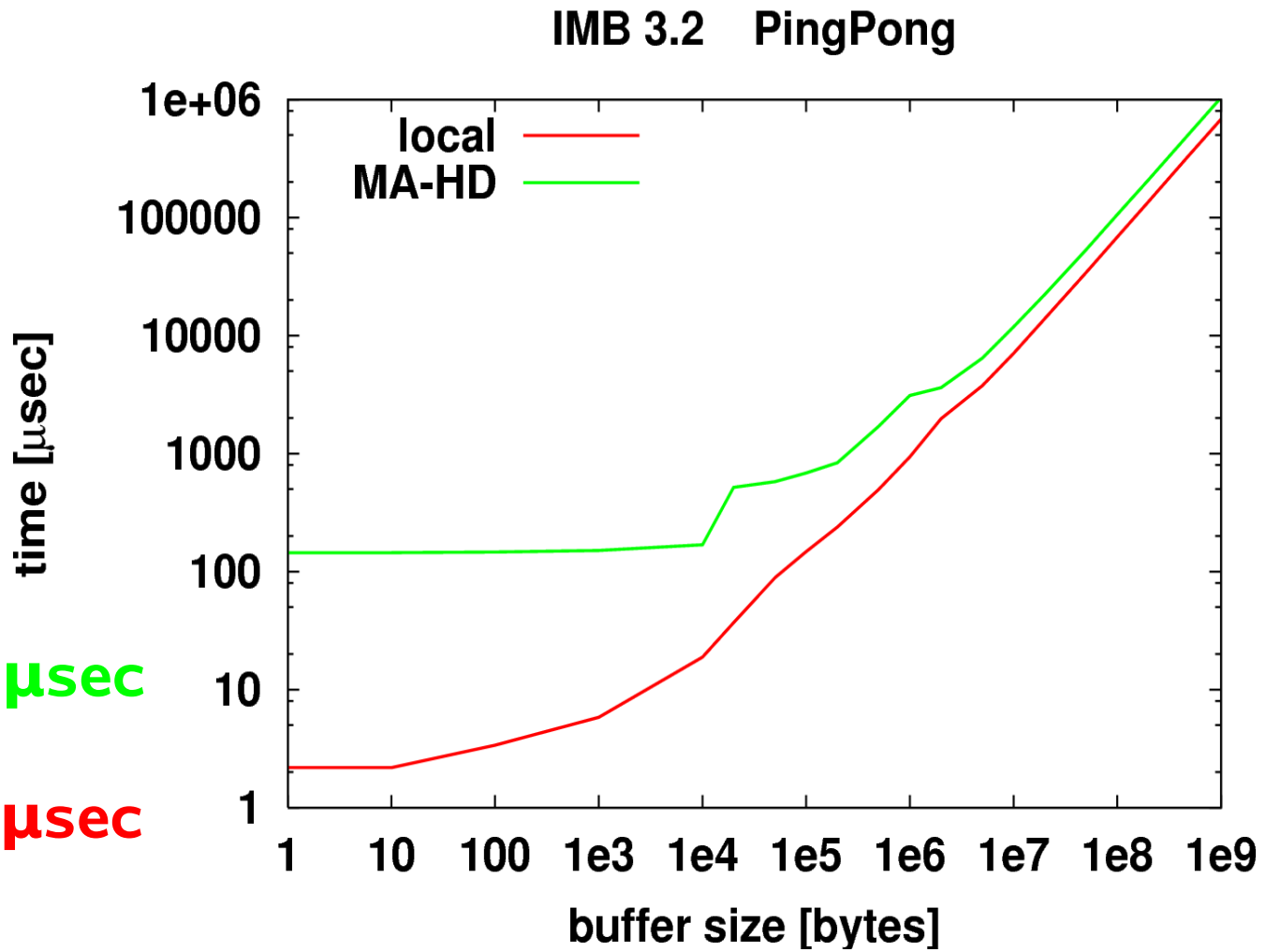




Quelle: HLRS, ADVA, Obsidian, 2008

MPI Performance (1)

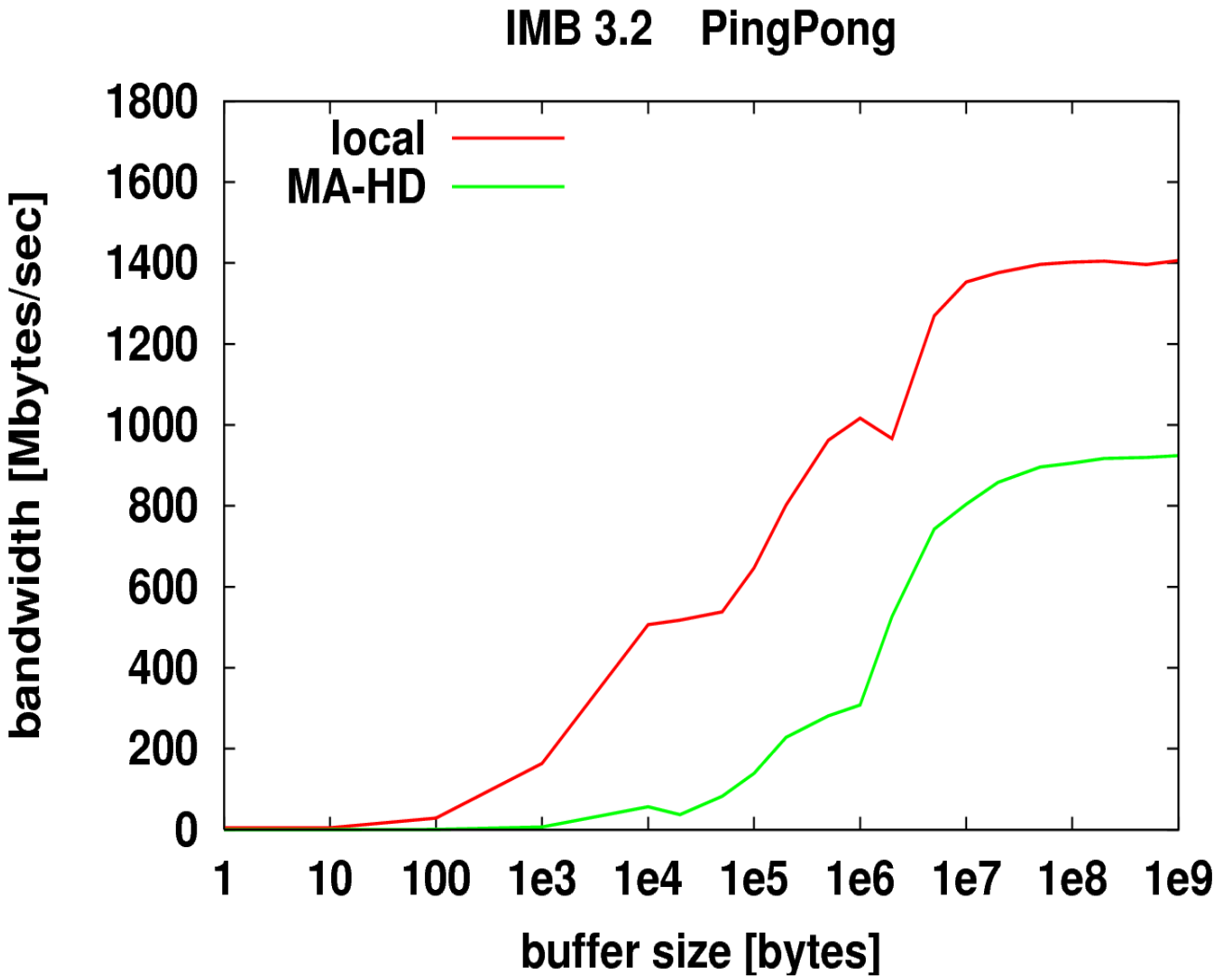
Latenz



Kopplung: 145 µsec

Lokal: ~2 µsec

MPI Performance (2)



Bandbreite

Lokal:
1400 MB/sec

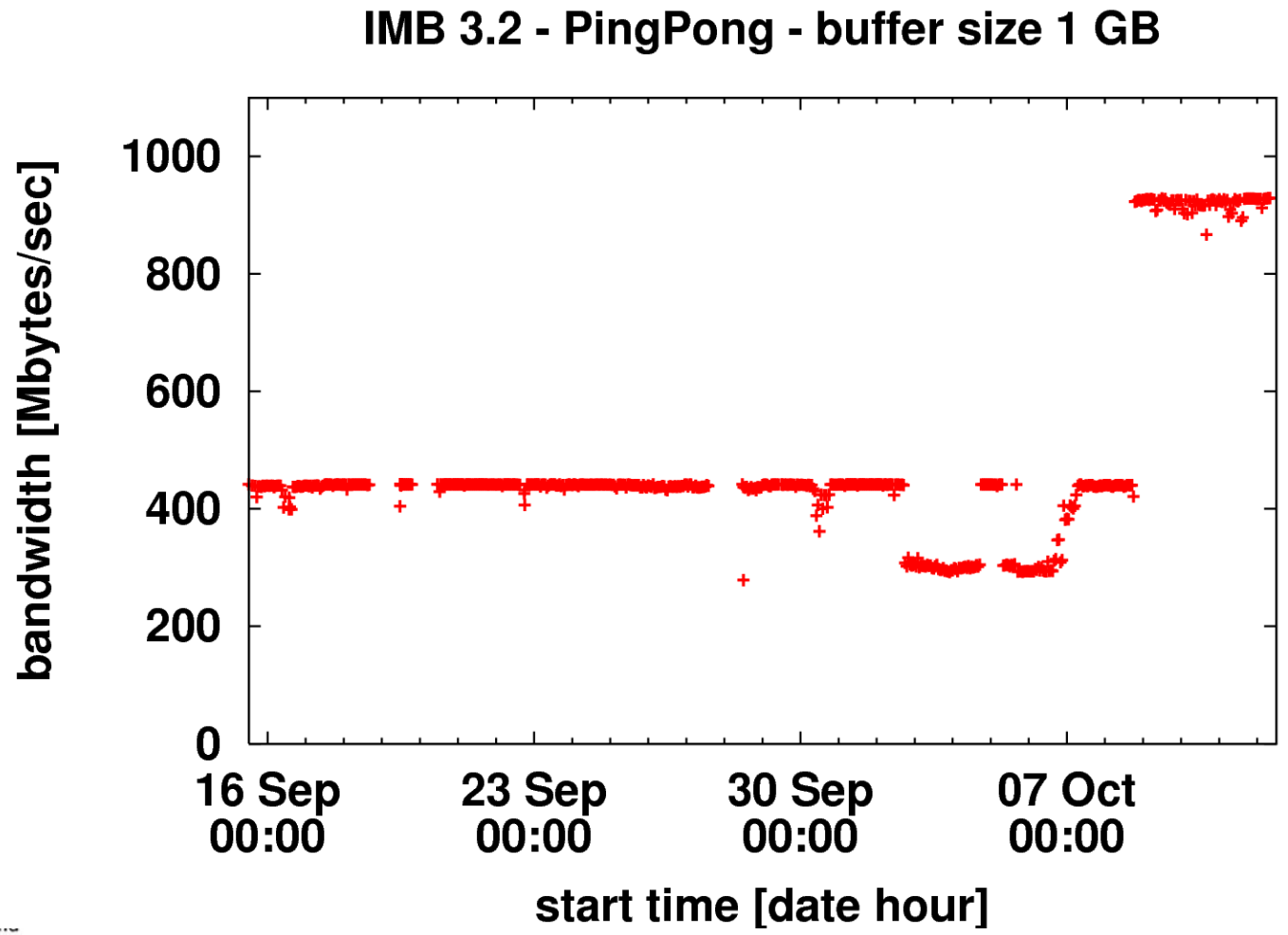
Kopplung:
930 MB/sec

Netztechnik Erfahrungen

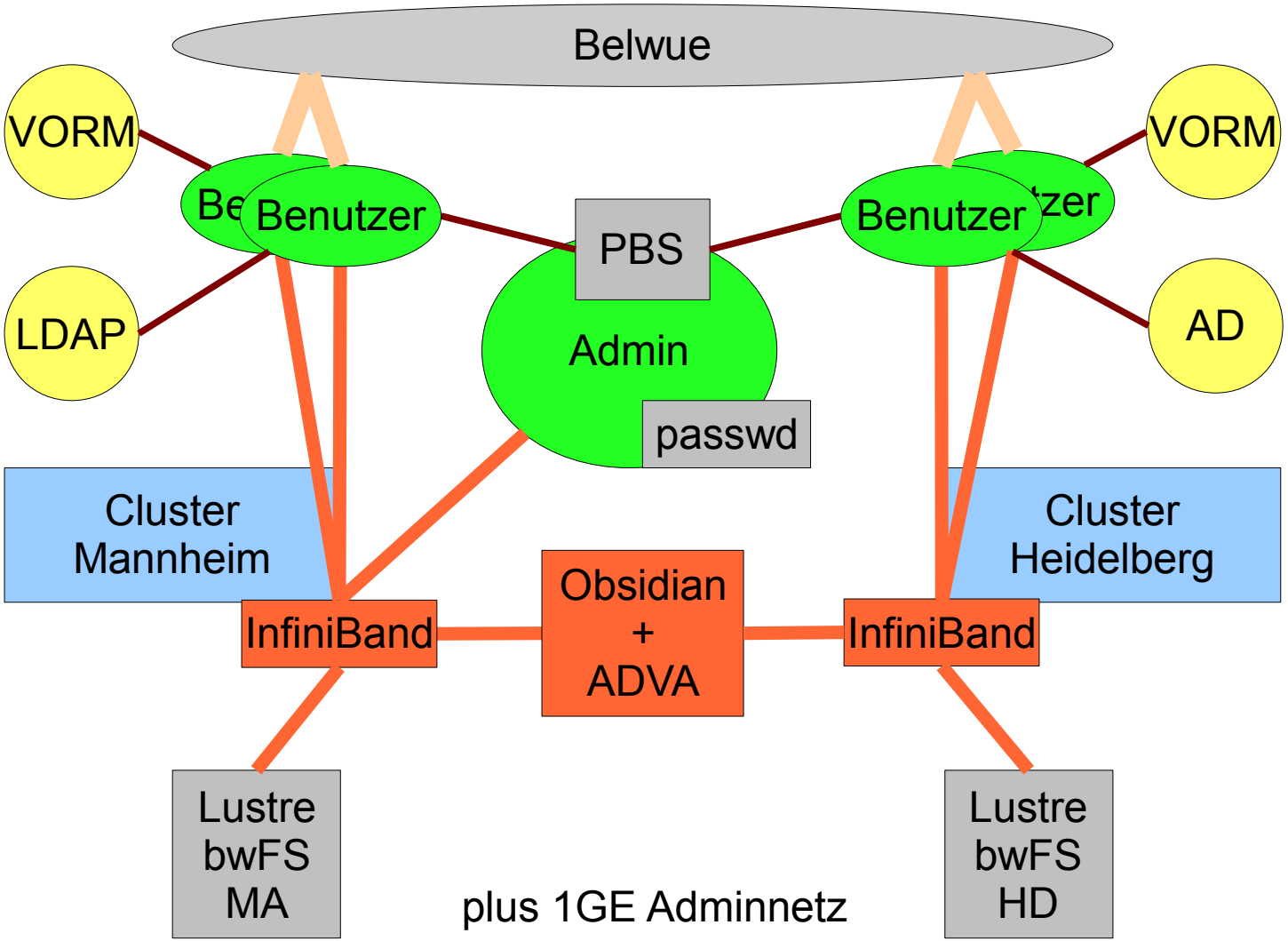
- Entfernung MA - HD beträgt 28 km (18 Luftlinie)
  - Licht benötigt 116 µs für die Entfernung
- Latenz mit 145 µs recht hoch, Licht + 30 µs
  - lokale Latenz 1.99 µs P-t-P (15 µs coll. comm.)
- Bandbreite mit ca. 930 MB/sec wie erwartet
  - lokale Bandbreite ca. 1200 - 1400 MB/sec
- Obsidian benötigt eine Lizenz für 40 km
  - hat Puffer für noch weitere Entfernungen
  - diese werden nur über die Lizenz aktiviert
  - wir waren der Meinung, Lizenz für 10 km reicht

MPI Bandbreite

- Einfluss der Obsidian Lizenz



Standorte und Komponenten



gemeinsame Administration

- nur noch ein Admin-Server, ein PBS
- 2 Zugangsrechner für SSH, 10GE an Belwü
- 2 Zugangsrechner für GSI-SSH, Globus, 10GE
- Cluster-Admin-Tools vom HLRS zur Administration der Hardware
  - MAC-Adressen, DHCP Tabelle
  - TFTP zum Booten der Kernel
  - NFS für (Admin) Software und Konfigurationen
- zusätzlich 2 extra BladeCenter für ITP (Inst. für Theor. Physik) in HD eingebaut

gemeinsame Benutzerverwaltung

- lokale Kennungen in MA und HD (die das bwGRiD benutzen) sind alle verschieden (!)
- Erzeugen der passwd und group Dateien
- Gruppen werden mit Prefix 'ma', 'hd', bzw. 'mh' für d-grid Nutzer versehen
- uidNumber +100.000 für MA, +200.000 für HD und +2.000.000 für d-grid
- Authentifizierung am Zugangsrechner
  - direkt über jeweiligen LDAP bzw. AD
  - oder mit d-Grid-Zertifikat

getrennte Storage Systeme

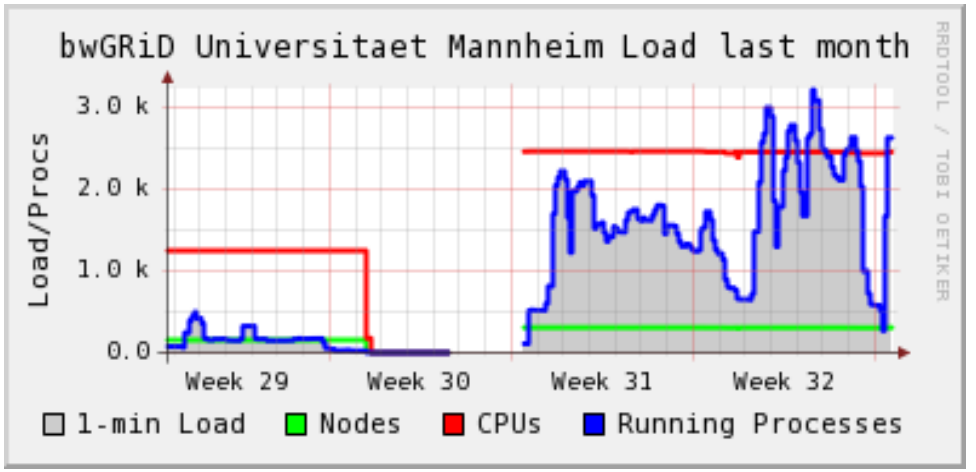
- BWFS Lustre mit 32 TB an jedem Standort
- Homes der Benutzer bleiben am Standort
- Lustre über Kreuz gemounted, über InfiniBand
  - /bwfs/hd/home, /bwfs/ma/home
  - /bwfs/ma/gridhome
- InfiniBand Performance scheint auszureichen
- Backup mit TSM etwas langsam
- Ausfall des Storage in Heidelberg
  - Nutzung des Mannheimer Storage für den gesamten Cluster

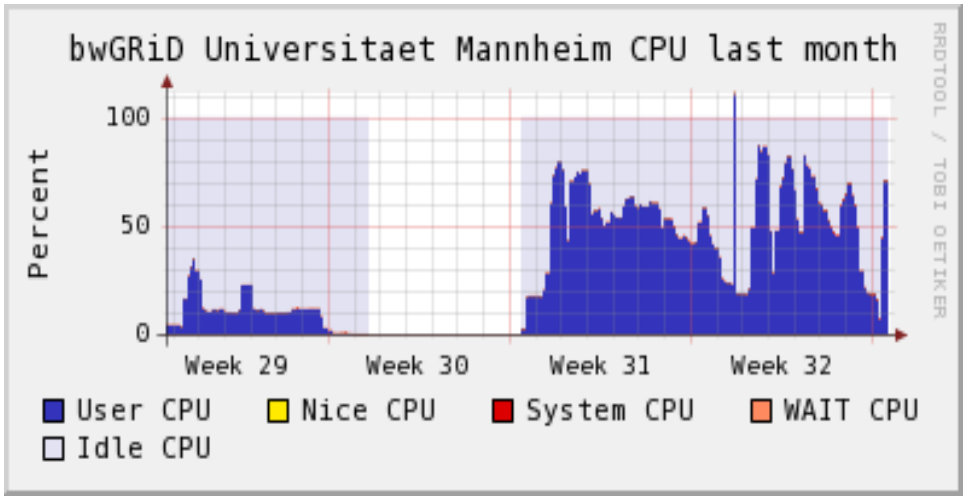
gemeinsames Batch-System

- wegen Latenz: Jobs auf einer Seite lassen
- Jobs bleiben auf einen Cluster beschränkt
  - d.h. keine Jobs mit mehr als ~140 Knoten
- PBS mittlerweile mit Moab Scheduler
- Performance von MPI mit InfiniBand bei Kommunikation über 28 km nicht ausreichend
  - aber keine Tests mit realen Workloads
  - Linpack bringt nur ca. ¼ der lokalen FLOPS
- 4 Queues: single, normal, itp, testkoppl

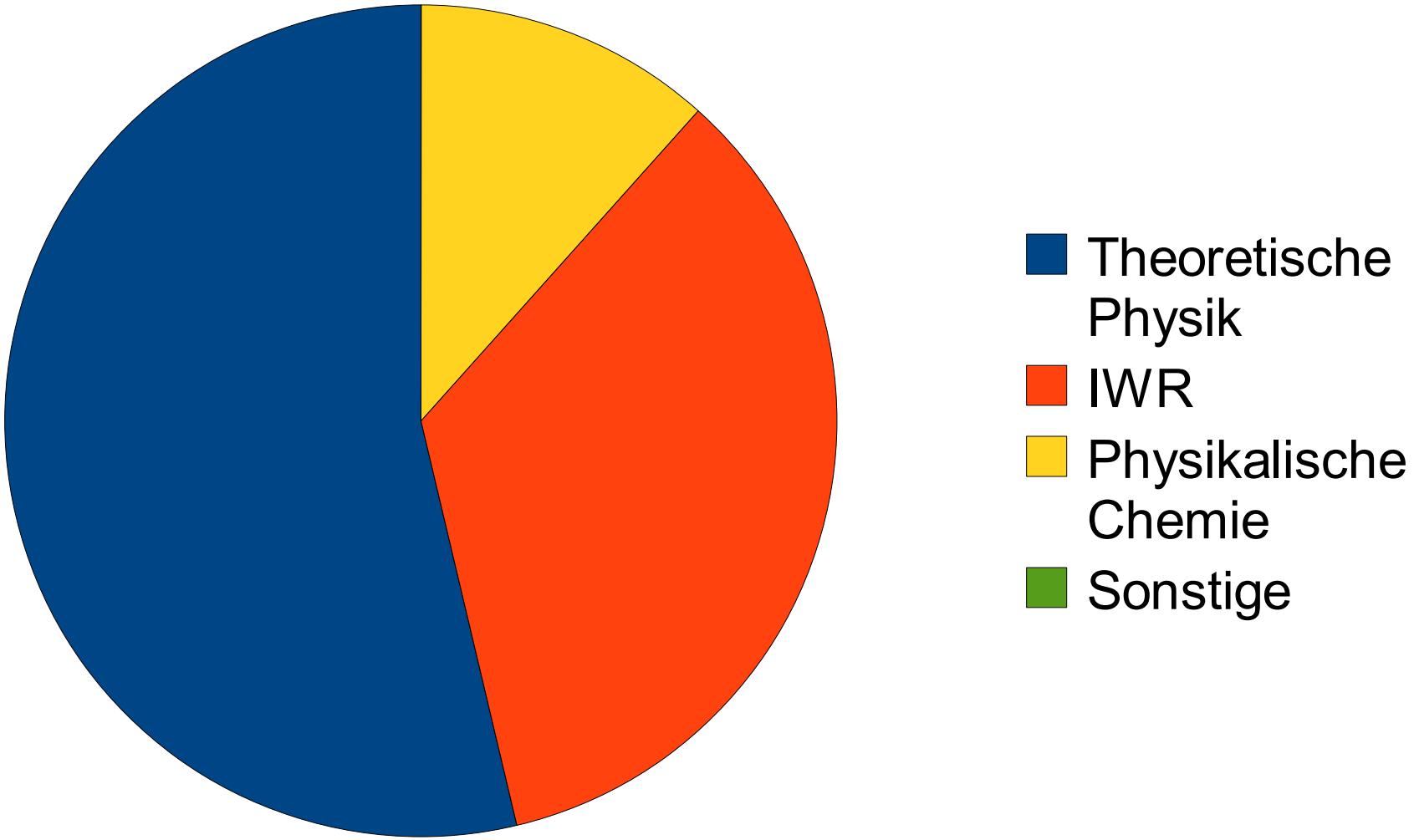
Ganglia Report

- CPU load 1-min
- Anzahl Prozesse
- User CPU Usage

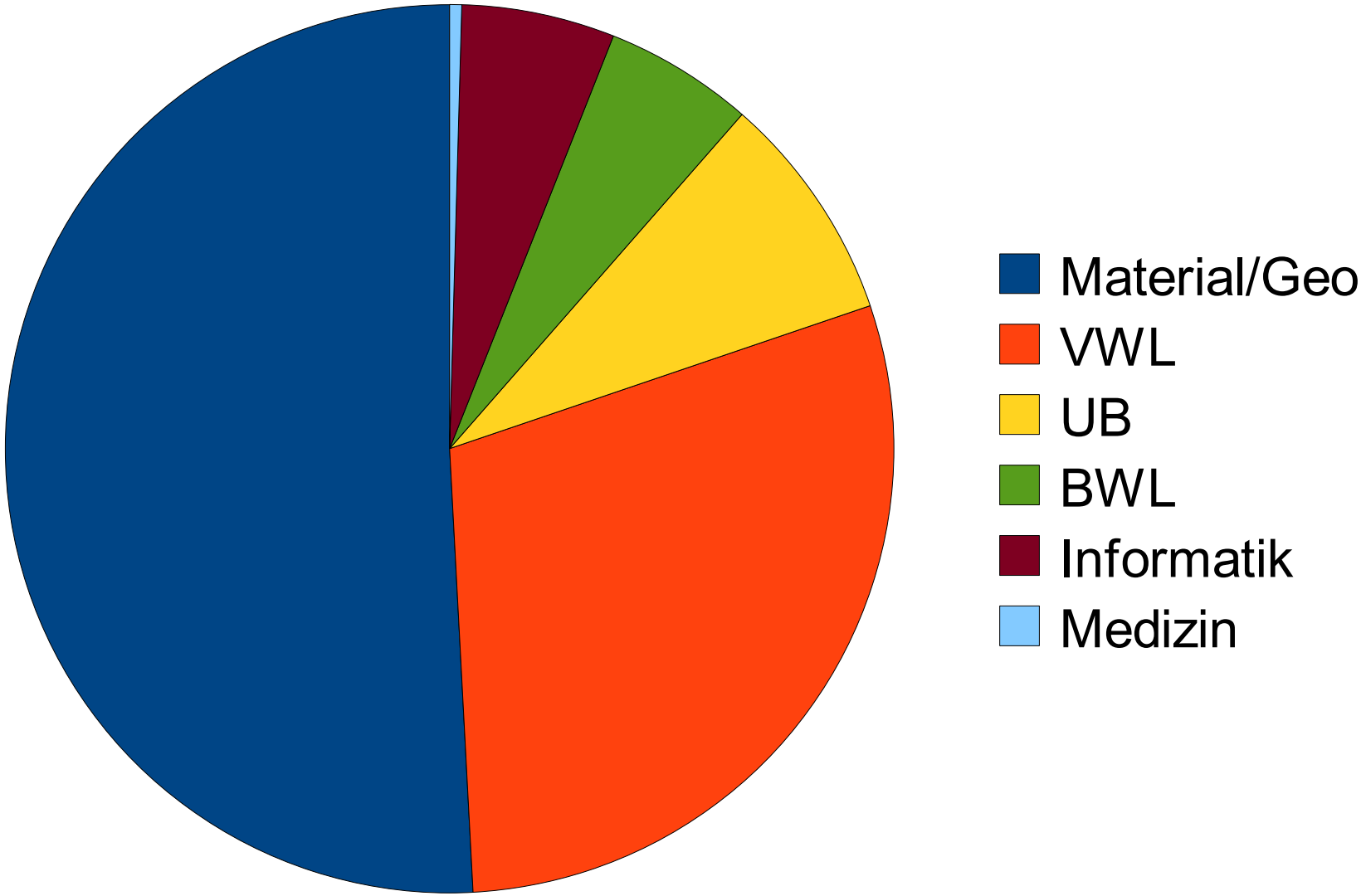




Rechenzeit 2009 - Heidelberg



Rechenzeit 2009 - Mannheim



Zusammenfassung

- Netztechnik: Obsidian, ADVA und InfiniBand funktioniert
- Latenz mit 145 µs für MPI Jobs zu hoch
- Bandbreite mit 930 MB/s ist gut
- Netzkopplung für „Single System Cluster“ Administration und Betrieb ausreichend und stabil
- guter Lastausgleich auf beide Standorte durch gemeinsames PBS
- Jobs sind auf einen Standort beschränkt

Dank an alle Beteiligten

- IWR Heidelberg
  - Bogdan Costescu, Hermann Lauer, Martin Neisen
- bwGRiD
  - Michael Schliephake, HLRS
  - Christian Mosch, KIZ Ulm
- RUM Mannheim
  - Helmut Fränznick, Rudi Müller, Gerd Rohde, Ralf-Peter Winkens
- URZ Heidelberg
  - Johannes Wilhelm, Hartmuth Heldt, Wolfgang Schrimm, Joachim Peeck

Vielen Dank für die Aufmerksamkeit!

- Fragen? Bemerkungen?

- Links
  - http://www.bw-grid.de/
  - http://web.urz.uni-heidelberg.de/server/grid/
  - http://www.uni-mannheim.de/rum/ag/zs/bw_grid_cluster/
  - http://www.hlrs.de/
  - http://www.d-grid.de/