bwGRiD Treff am URZ

Sabine Richling, Heinz Kredel

Universitätsrechenzentrum Heidelberg
Rechenzentrum Universität Mannheim

29. April 2010

bwGRiD Treff – Participants

Invited are

- Current users of the bwGRiD Clusters HD/MA
- Students and scientists interested in Grid Computing
- Members of the Universities Heidelberg and Mannheim

bwGRiD Treff – Content

- Status and Plans for bwGRiD
  - bwGRiD Cluster HD/MA
  - bwGRiD Project
  - Discussion software modules
- Lectures and/or workshops
  - Introduction Batch-System
  - Software packages and Parallelization
  - Programming in Java, Fortran, C
  - Parallelization with MPI/OpenMP
  - Grid access and Grid usage
- User contributions
  - Presentation of projects
  - Demonstration of problems/solutions
- To meet you in person

bwGRiD Treff – 29.04.2010

Agenda for today:

- The current status of bwGRiD (S. Richling)
- Details on the interconnection of the bwGRiD clusters Heidelberg and Mannheim (H. Kredel)
- Plans for the home directories and the next image (S.Hau)
- Presentation of a medicine project (L. Jahnke, Medizinische Fakultt Mannheim)
- Discussion of topics for further meetings

Current Status of bwGRiD

S. Richling (URZ)

What is a Grid? What is Grid Computing?

"Grid computing is coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations." (Ian Foster)

- A Grid is a infrastructure which integrates resources such as high performance computers, compute clusters, data storage units, networks and scientific instruments.
- Grid resources can be used for problem solving in virtual organisations.
- Virtual organisations consist of members from different institutions working together and sharing the resources.
- Grid resources remain under the control of the owner.

D-Grid Initiative

- www.d-grid.de
- Start: September 2005
- Funding: $\sim$ 50 Million Euro
- Aim: Development and establishment of a reliable and sustainable Grid infrastructure for e-science in Germany.
- Funded by the Federal Ministry of Education and Research (BMBF)

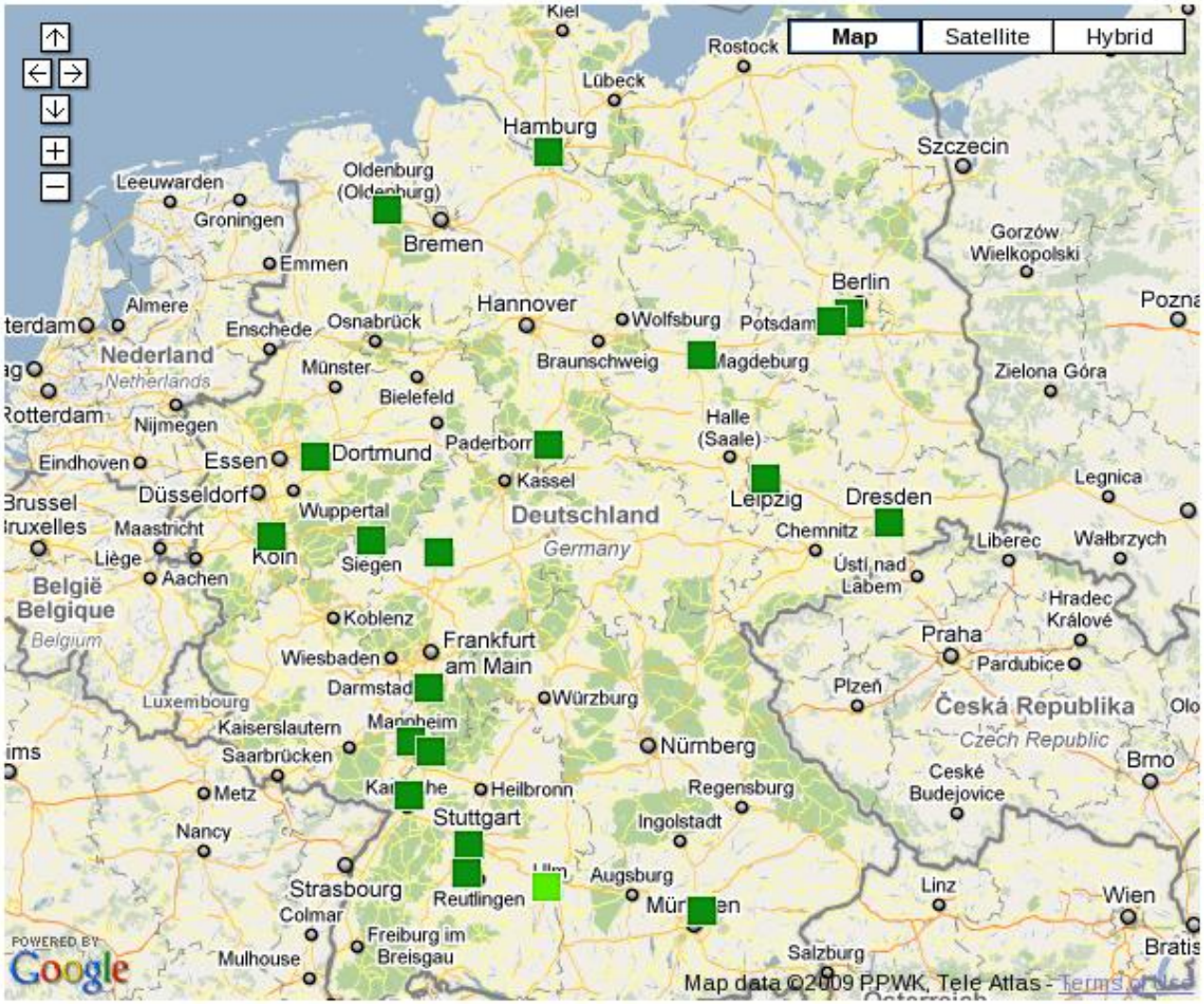
Bundesministerium
für Bildung
und Forschung

D-Grid Overview

- One project for the development of a Grid platform and for securing the sustainability of this platform:
  - D-Grid Integration Project (DGI)
- Community projects from different fields of research, e.g.:
  - High Energy Physics (HEP-Grid)
  - Astrophysics (AstroGrid-D)
  - Medicine and Life sciences (MediGrid)
  - Climate Research (C3-Grid)
  - Arts and Humanities (TextGrid)
  - Universities of BW (bwGRiD)
- More than 20 German computing centers provide resources for D-Grid.
- D-Grid supports several Grid middlewares and services, e.g.:
  - Globus Toolkit
  - UNICORE
  - gLite
  - GridSphere

D-Grid Resource Provider

Dynamisch erstellte Ressourcenanbieter-Karte



normaler Betrieb
geplante Wartung
Wartung
Fehler

Geomaint MDS Dynamic Map

bwGRiD

- www.bw-grid.de
- Community project of the Universities of BW
- Compute Clusters: Stuttgart, Ulm (Konstanz), Karlsruhe, Tbingen, Freiburg, Mannheim/Heidelberg
- Central storage unit in Karlsruhe
- Distributed system with local administration
- Computing centers focus on software in different fields of research
- Access via at least one middleware supported by D-Grid

bwGRiD - Aims

- Proof of the functionality of and the benefit from Grid concepts for the HPC community.
- Managing organisational and security problems
- Development of new cluster and Grid applications
- Solving license difficulties
- Enabling the computing centers to specialize

Access Possibilities

Important!

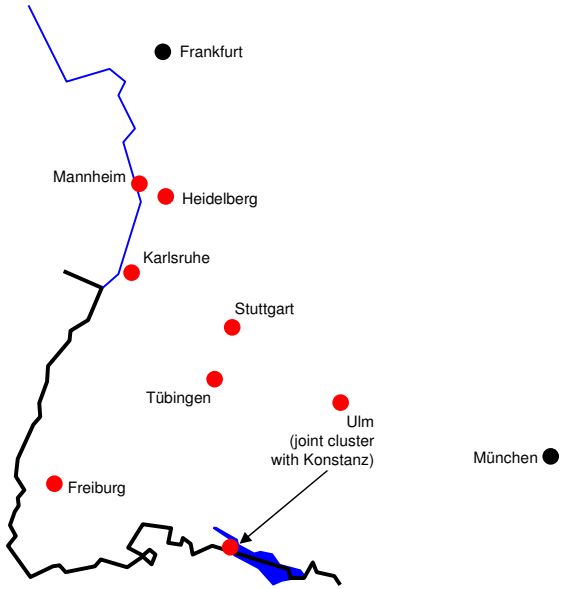
- Access with local Accounts:
  Project numbers and User IDs (URZ); User IDs (RUM)
  $\longrightarrow$ Access **only** to bwGRiD cluster MA/HD
- Access with Grid Certificate:
  Grid Certificate, VO membership, Grid Middleware
  $\longrightarrow$ Access to all bwGRiD resources

Access to bwGRiD resources

With Middleware Globus Toolkit (gsissh, GridFTP, Web Services)

- Compute cluster:
  - Mannheim/Heidelberg: 280 nodes
  - Karlsruhe: 140 nodes
  - Stuttgart: 420 nodes
  - Tbingen: 140 nodes
  - Ulm (Konstanz): 280 nodes
  - Freiburg: 140 nodes
- Central storage:
  - Karlsruhe:

128 TB (with Backup)
256 TB (without Backup)



bwGRiD Software

- Common Software:
  - Scientific Linux
  - Intel Compiler
  - Software modules prepared for distribution among the sites (e.g. MPI versions, mathematical libraries, various free software)
- Focus of bwGRiD sites:
  - Mannheim: BWL, VWL, Computer Algebra
  - Heidelberg: Mathematics, Neuroscience
  - Karlsruhe: Engineering, Compiler & Tools
  - Stuttgart: Automotive simulations, Particle simulations
  - Tbingen: Astrophysics, Bioinformatics
  - Ulm: Chemistry, Molecular Dynamics
  - Konstanz: Biochemistry, Theoretical Physics
  - Freiburg: System Technology, Fluid Mechanics

bwGRiD in Development

- Integration into D-Grid Infrastructures (VO, Middleware, Monitoring)
- D-Grid User Support Portal
- Unification of the bwGRiD clusters
- Development of bwGRiD Portals
- Improvement of the bwGRiD webpage http://www.bw-grid.de

D-Grid User Support

- Ticket-System
- D-Grid News
- Maintenance
- May 2010: Integration into European System NGI

$\Rightarrow$ `http://dgus.d-grid.de/`

Unification of the bwGRiD clusters

- Operating system, Software
- Job Queues: Default queue available at all sites
- Workspaces for temporary files and large amounts of data (User allocates, extents and deletes workspaces by himself)
- User Support: Login messages, bwGRiD manpage, module help, coordination of local and central documentation

Development of bwGRiD Portals

- Freiburg: System Technology Portal (K. Kaminsiki)
- Heidelberg: Medicine Portal (L. Jahnke, J. Fleckenstein, M. Niknazar, J. Hesser)
- Tbingen: Bioinformatics Portal (S. Storch, W. Dilling)
- Ulm: Chemistry Portal (K. Taylor, D. Benoit), Avatar (H. Lang), Basis Portal and Gatlet (B. Boegel), Project leader (C. Mosch)
- Stuttgart: Workflow Management (T. Krasikova, Y. Yudin, N. Currle-Linde)
- Karlsruhe: Engineering Portal (E. Syrjakow)

$\Rightarrow$ Demonstration of the Basis Portal.

Improvement of the bwGRiD webpage

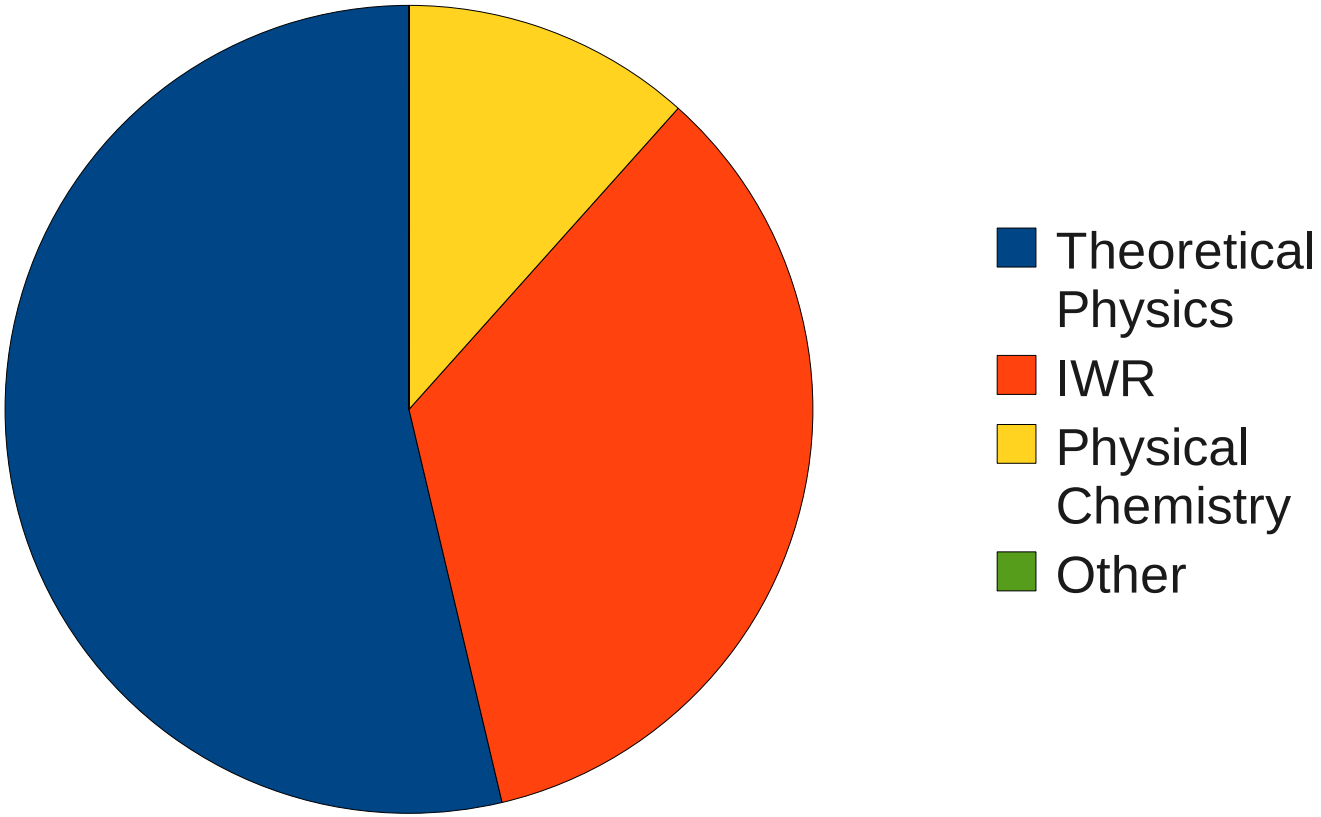
- New Layout
- More Content
- Access Information
- Portal Section
- User Project Descriptions

$\Rightarrow$ `http://www.bw-grid.de/`

User Projects 2009 – Heidelberg

- Theoretical Physics:
  QCD, Monte-Carlo and Molecular Dynamics
  Chromatin Folding, Bose-Einstein condensates
- IWR:
  Molecular Biophysics, Computational Neuroscience
  Development of parallel solvers
- Physical Chemistry:
  Electronic structure, Molecular Dynamics
  Lipids, Proteins, Many-Body-Systems

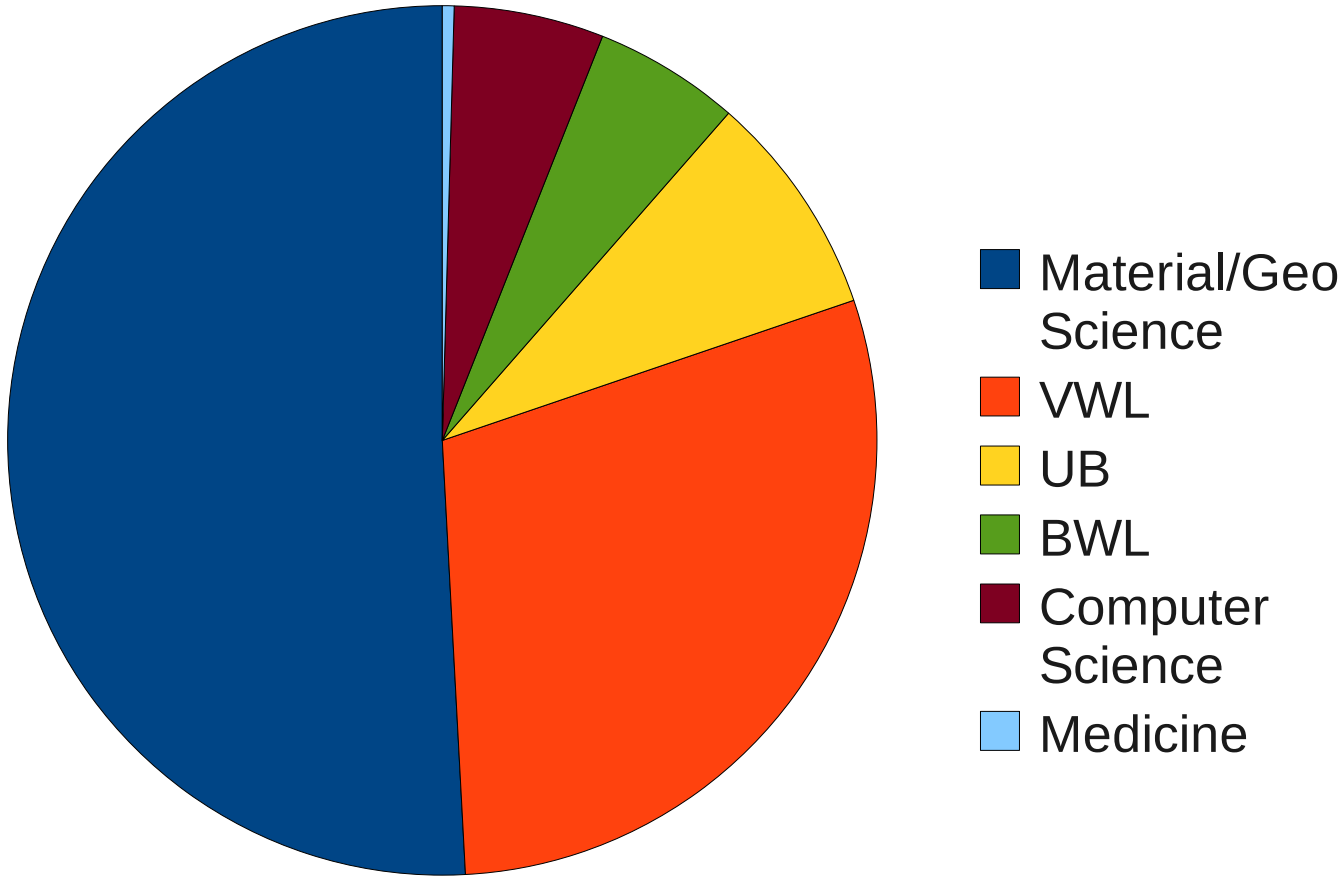
CPU Time 2009 – Heidelberg



User Projects 2009 – Mannheim

- VWL: Statistical analysis, Security Policy, Relationship between generations, Currencies
- BWL: Statistical analysis, Manager options, Implicit capital costs, Insider trading
- Computer Science: Simulation of mobile Networks, Tracking algorithms
- Material- and Geo science (Darmstadt): Simulation of nano-crystalline materials and crystal growth
- Medicine: Statistical DNA analysis
- UB: Automatic classification of documents

CPU Time 2009 – Mannheim



Your project will be published at www.bw-grid.de soon

- We prepare and send you a template containing information on your project we already have.
- You add missing information (coworkers, pictures, links, publications) and send it back to us.
- We collect the improved project descriptions and send it to Konstanz.
- Updates are possible at any time.

Planned for May/June 2010

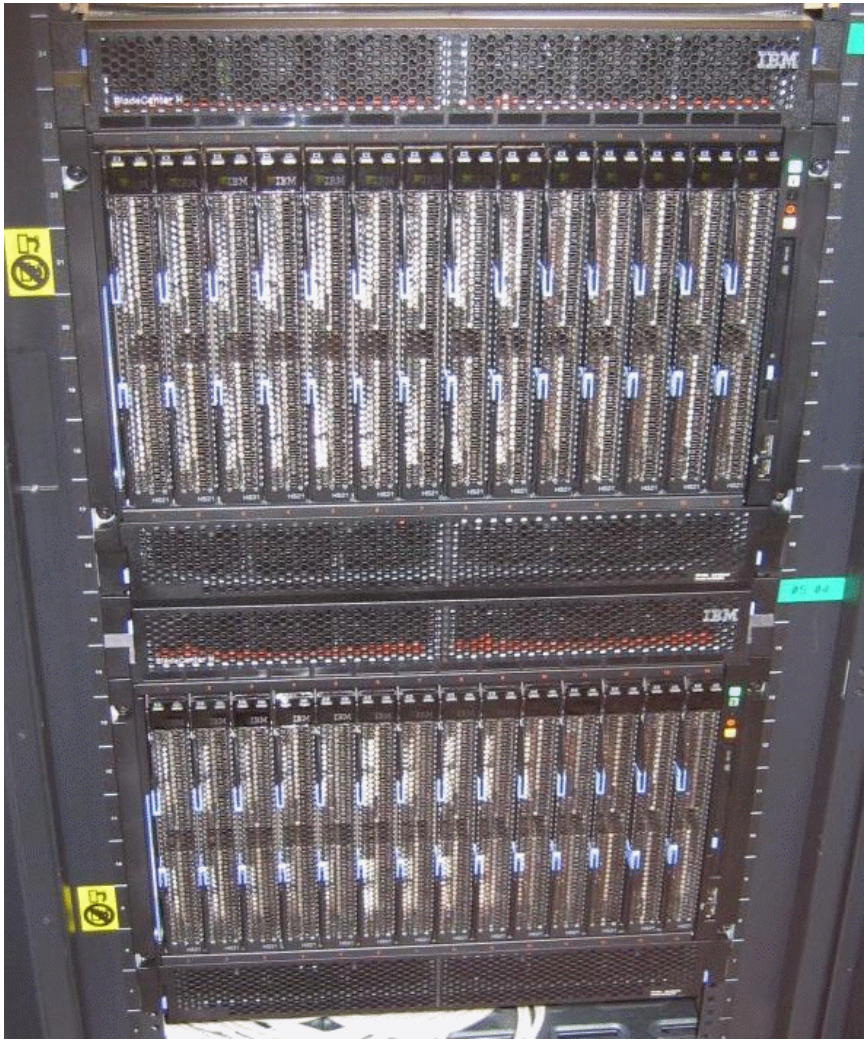
Interconnection of the bwGRiD clusters Heidelberg and Mannheim

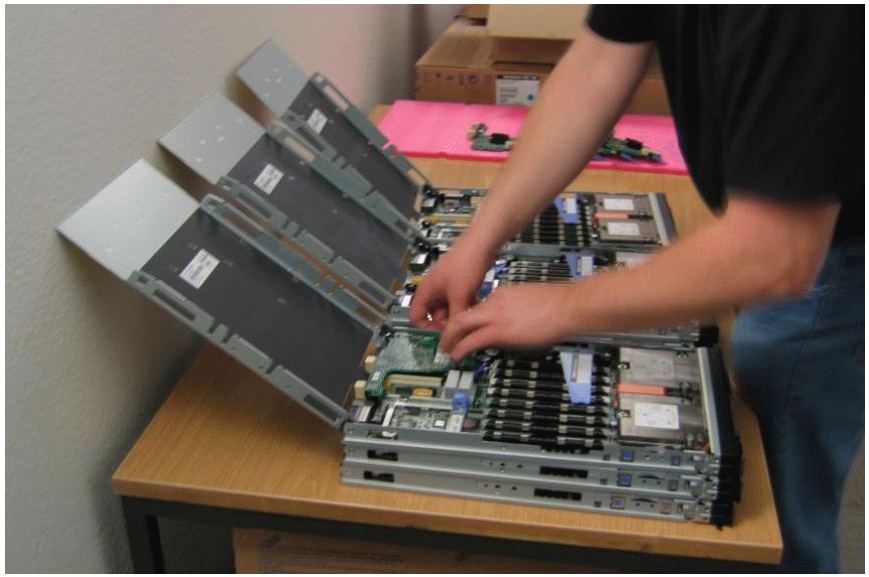
H. Kredel (RUM)

Hardware before Interconnection

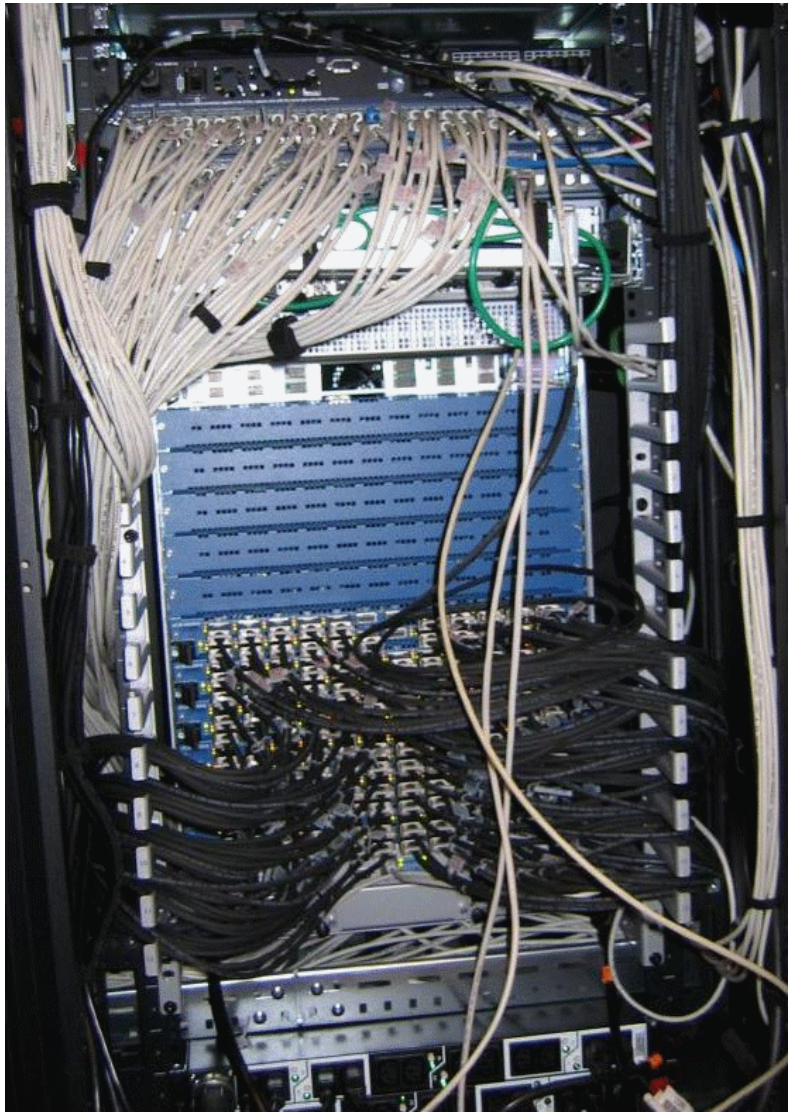
- 10 Blade-Center in Heidelberg and 10 Blade-Center in Mannheim
- Each Blade-Center contains 14 IBM HS21 XM Blades
- Each Blade contains
  - 2 Intel Xeon CPUs, 2.8 GHz (each CPU with 4 Cores)
  - 16 GB Memory
  - 140 GB Hard Drive
  - Gigabit-Ethernet
  - Infiniband Network
- $\Rightarrow$ 1120 Cores in Heidelberg and 1120 Cores in Mannheim

Hardware – Bladecenter





Hardware – Infiniband



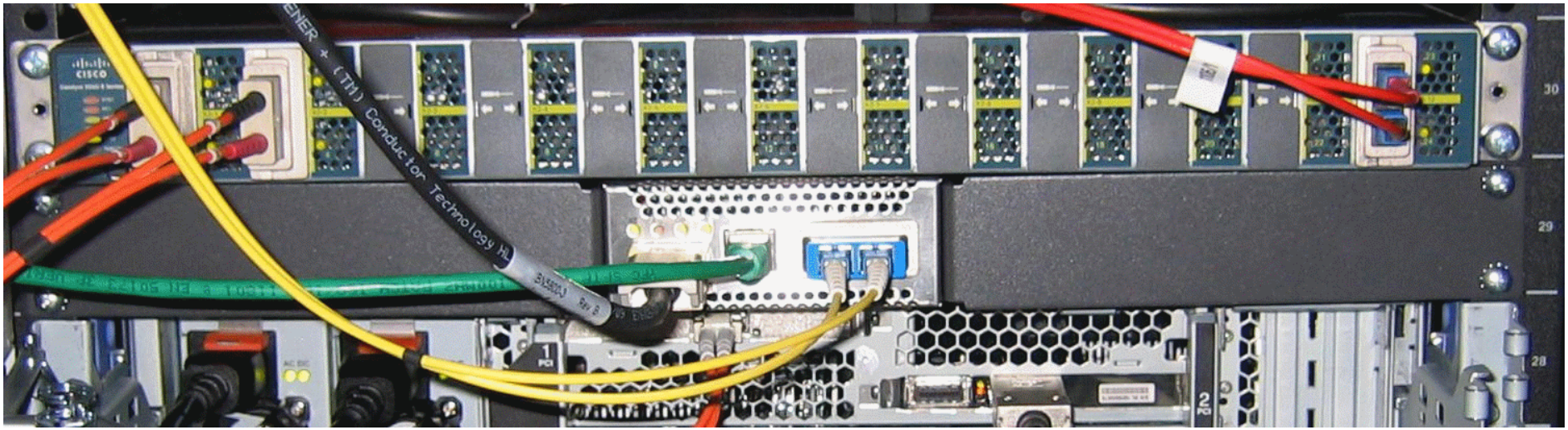
Hardware – Timeline

- January – March 2008: Delivery and assembly
- Operation in 2008:
  - Ethernet and Infiniband are working
  - Batch System is configured
  - NFS-Server for Home directories:
    IWR in Heidelberg, RUM in Mannheim
  - User administration:
    IWR in Heidelberg, RUM in Mannheim
- January 2009: Internal hard drives for the blades
- May – July 2009:
  - bwGRiD Storage System for home directories:
    32 TB, parallel filesystem Lustre
    (one system in Heidelberg and one in Mannheim)
  - URZ takes over user administration in Heidelberg
  - Interconnection of the bwGRiD Clusters

Hardware – bwGRiD Storage System

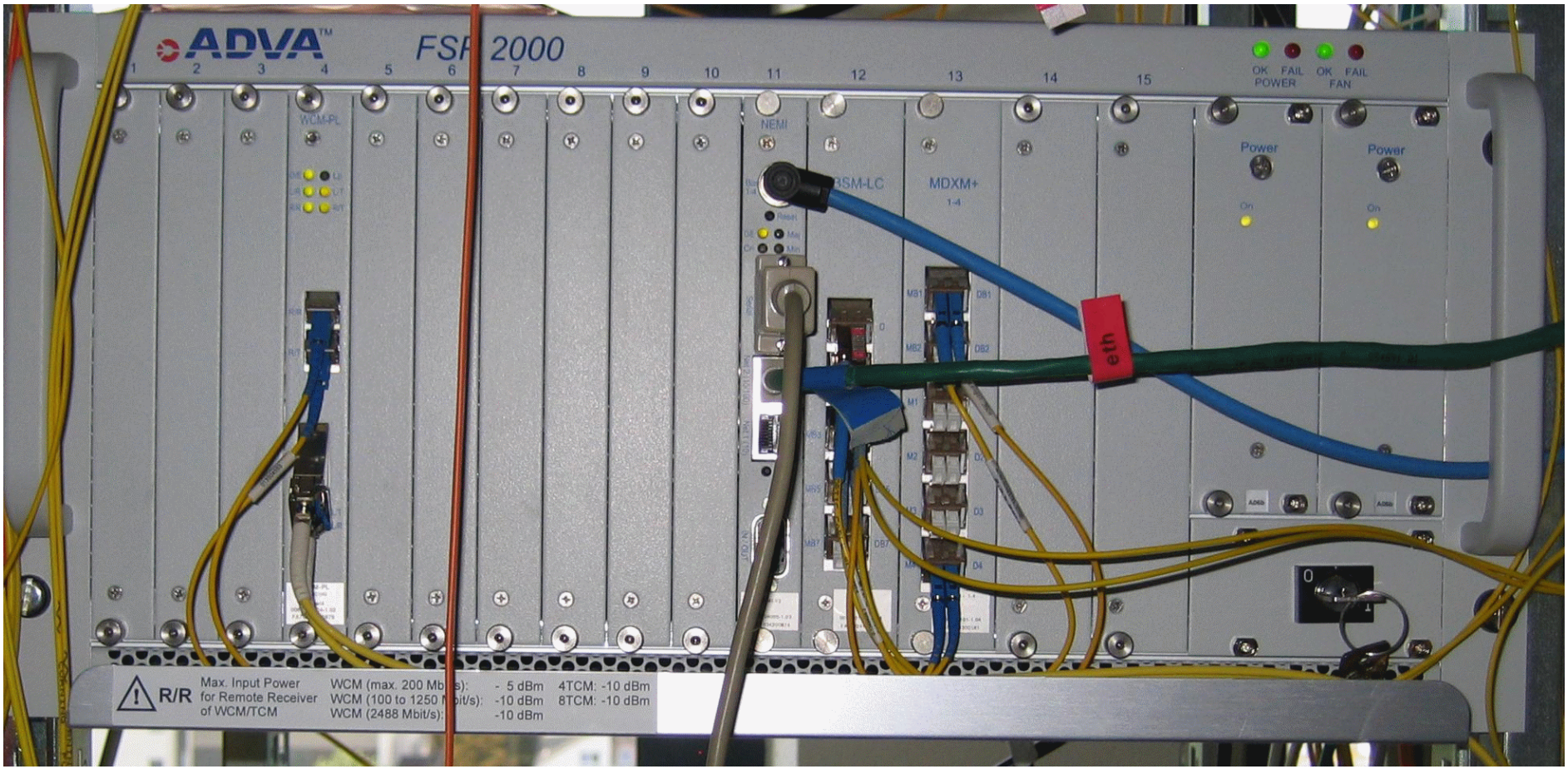
Interconnection of the bwGRiD clusters

- Proposal in 2008
- Acquisition and Assembly until May 2009
- In service since July 2009
- Infiniband over fibre channel: Obsidian Logbow



Interconnection of the bwGRiD clusters

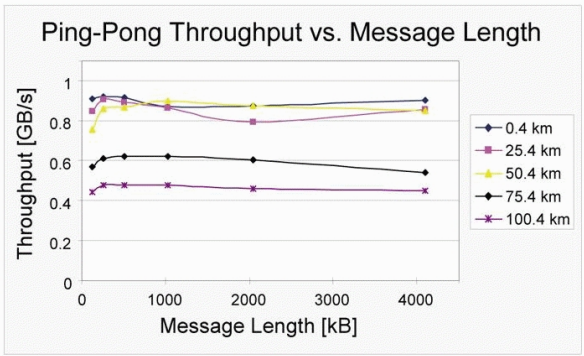
- ADVA: Input to DWDM line

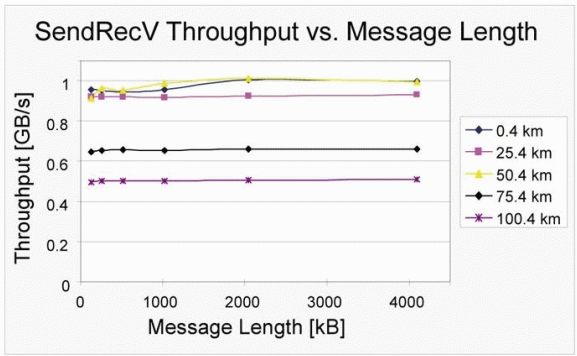


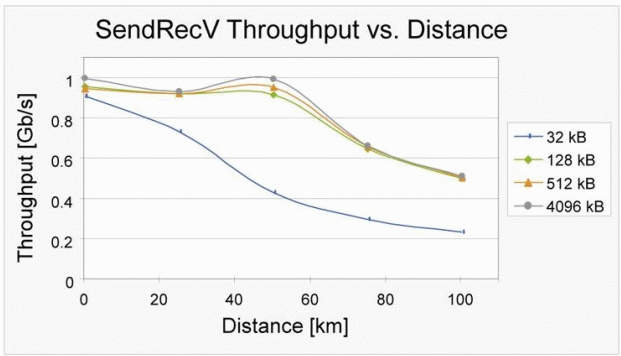
HLRS MPI Performance

- Measurements for different distances
- up to 50-60 km are feasible
- Bandwidth 900-1000 MB/sec
- Latency is not published

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

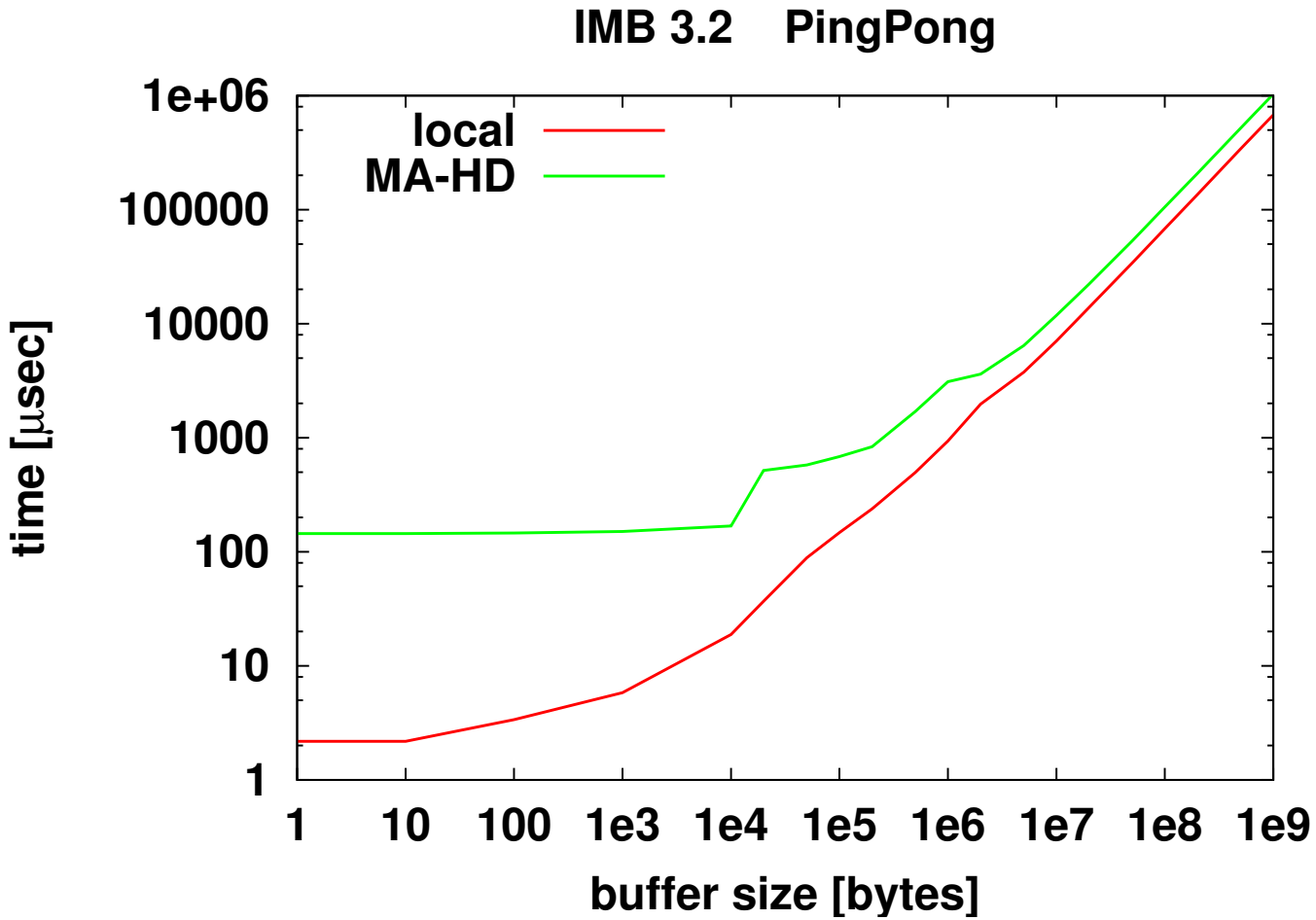






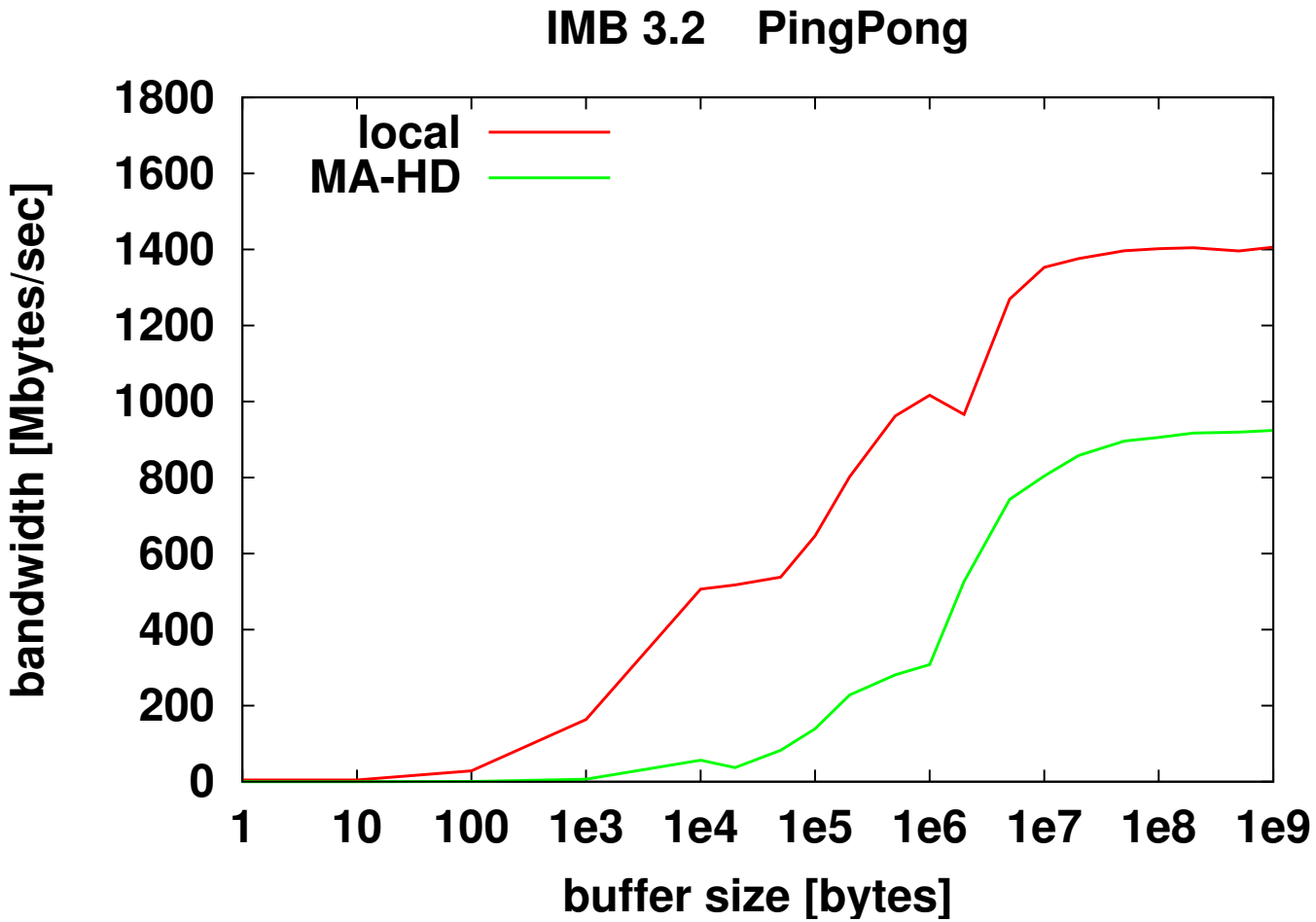
MPI Performance – Latency

Local: ~ 2 μsec
Interconnection: 145 μsec



MPI Performance – Bandwidth

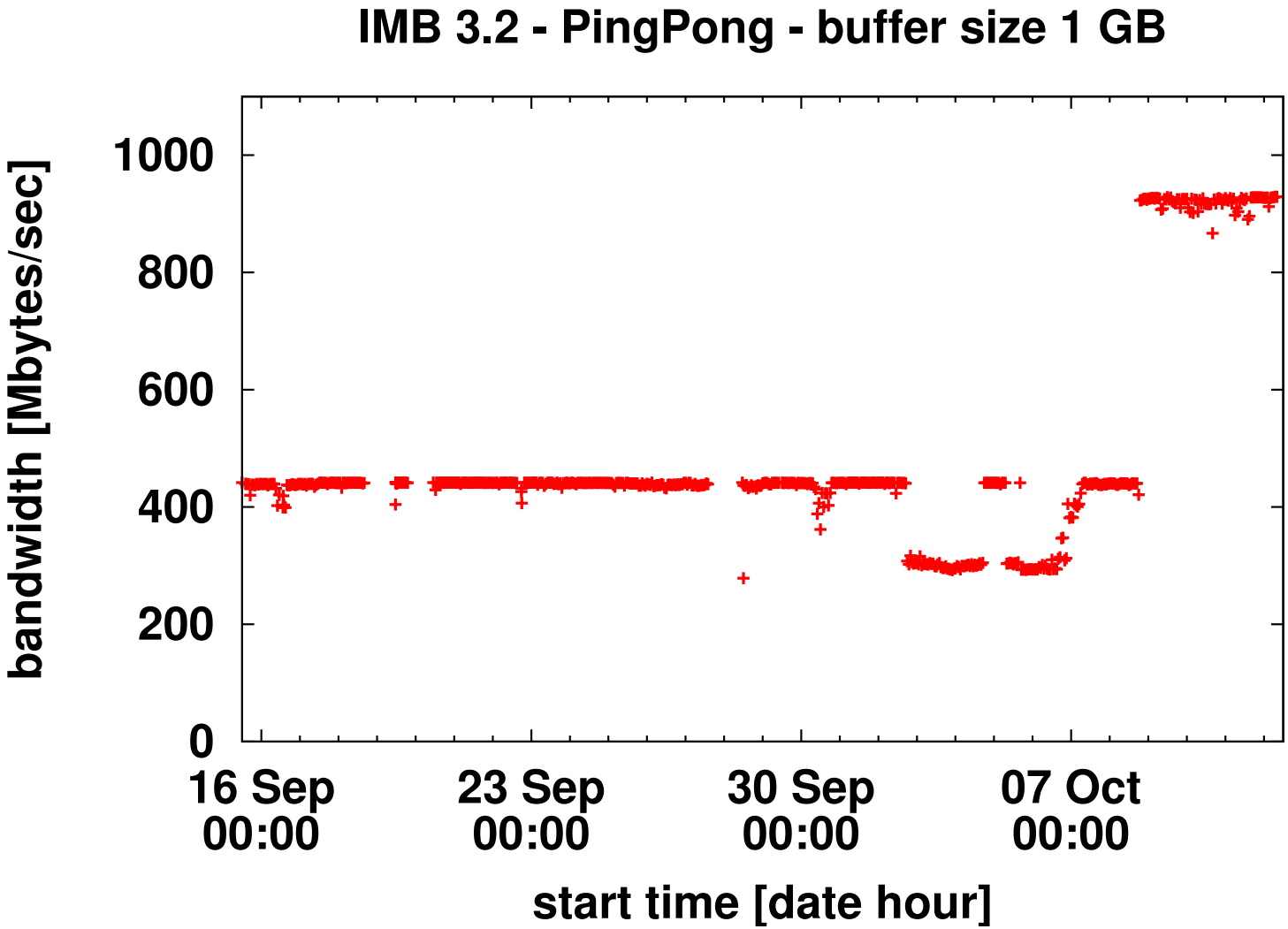
Local: 1400 MB/sec
Interconnection: 930 MB/sec



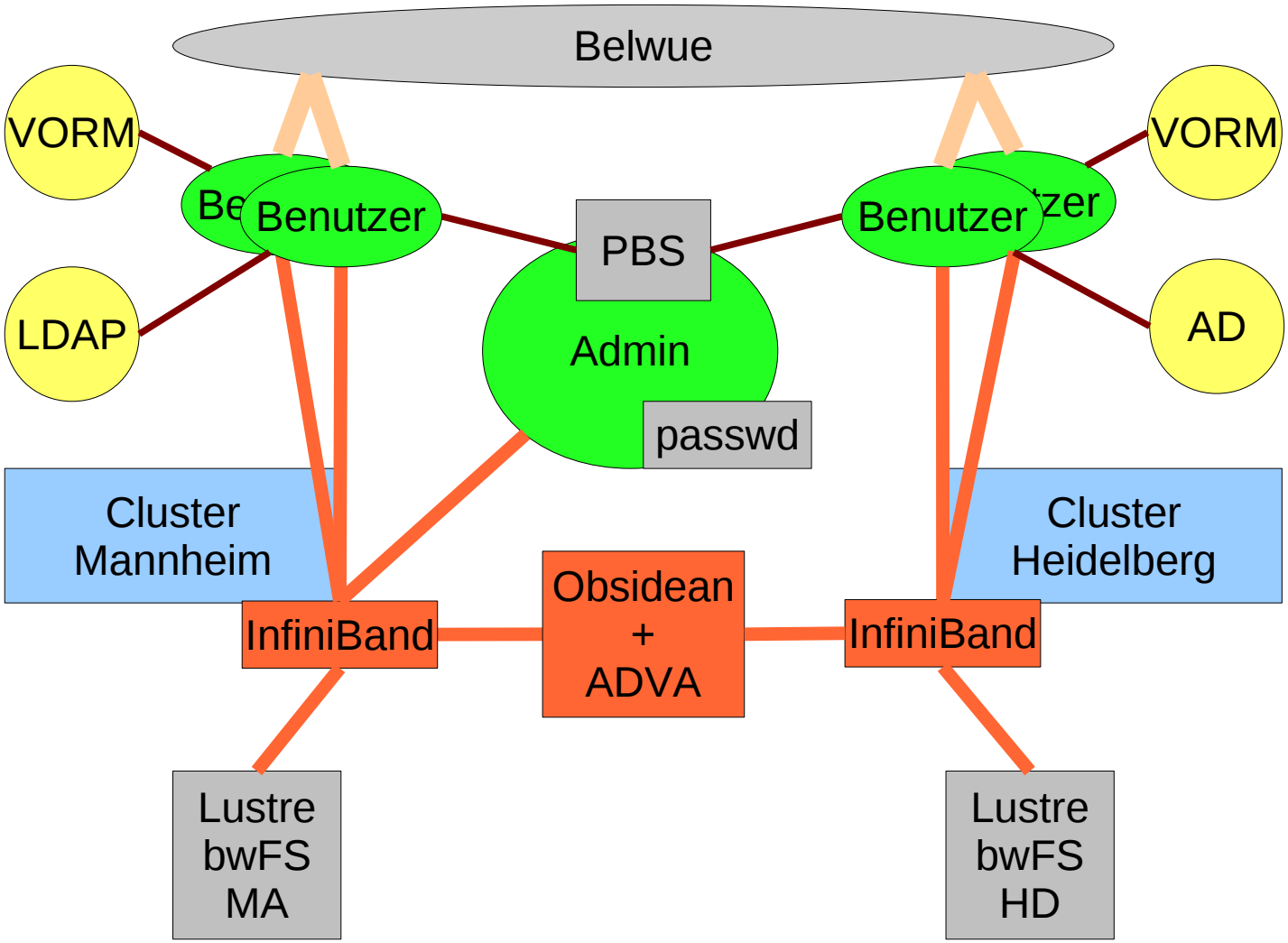
Network Experiences from Interconnection

- Distance MA-HD is 28 km (18 km linear distance)
  $\Rightarrow$ Light needs 116 μsec for this distance
- Latency is high: 145 μsec = Light transit time + 30 μsec
  Local latency only 1.99 μ sec P-t-P (15 μsec coll. comm.)
- Bandwidth is as expected: about 930 MB/sec
  Local bandwidth 1200-1400 MB/sec
- Obsidian needs a license for 40 km
  - Obsidian has buffers for larger distances
  - Activation of buffers with license
  - License for 10 km is not sufficient

MPI Bandwidth – Influence of the Obsidian License



bwGRiD Cluster Mannheim/Heidelberg



Common Cluster Administration

- only one admin server, one PBS
- 2 access nodes for ssh, 10 GE to Belw
- 2 access nodes for gsissh/Globus, 10 GE to Belw
- Cluster-Admin-Tools from HLRS for hardware administration
  - MAC addresses, DHCP table
  - TFTP to boot the kernel
  - NFS for (admin) software and configuration
- 2 bladecenter of the Institute for Theoretical Physics are included in Heidelberg
- both bwGRiD Storage systems are mounted over Infiniband

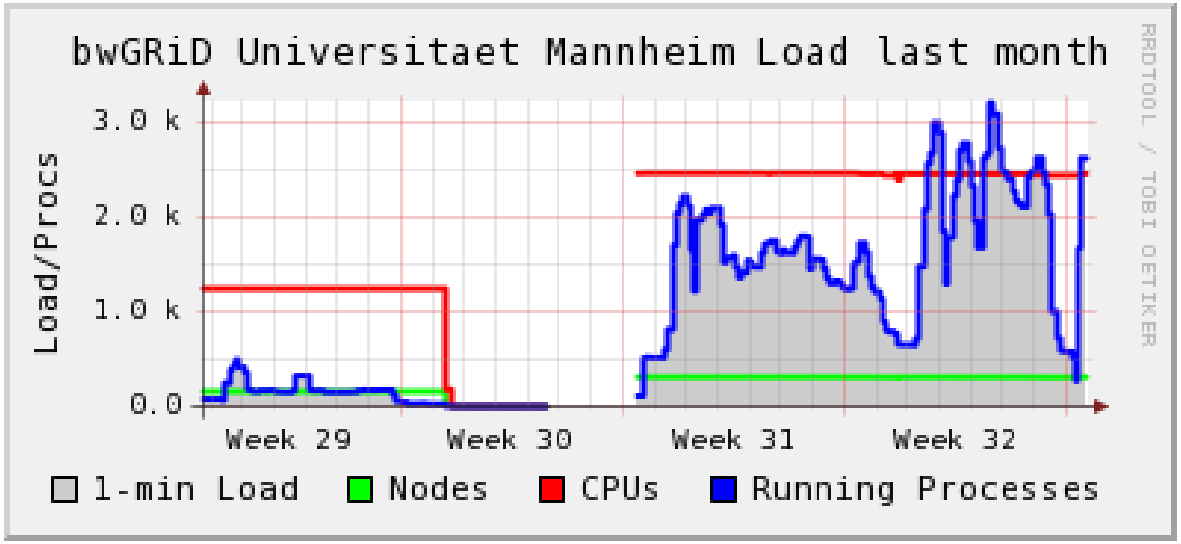
Common User Administration

- Local accounts of bwGRiD users in MA and HD must be different (!)
- Generation of a common passwd and common group files
- Groups get the prefix “ma”, “hd” or “mh” for D-Grid users
- uidNumber +100.000 for MA, +200.000 for HD and +2.000.000 for D-Grid
- Authentication at the access nodes
  - directly using LDAP (MA) and AD (HD)
  - or with D-Grid certificate

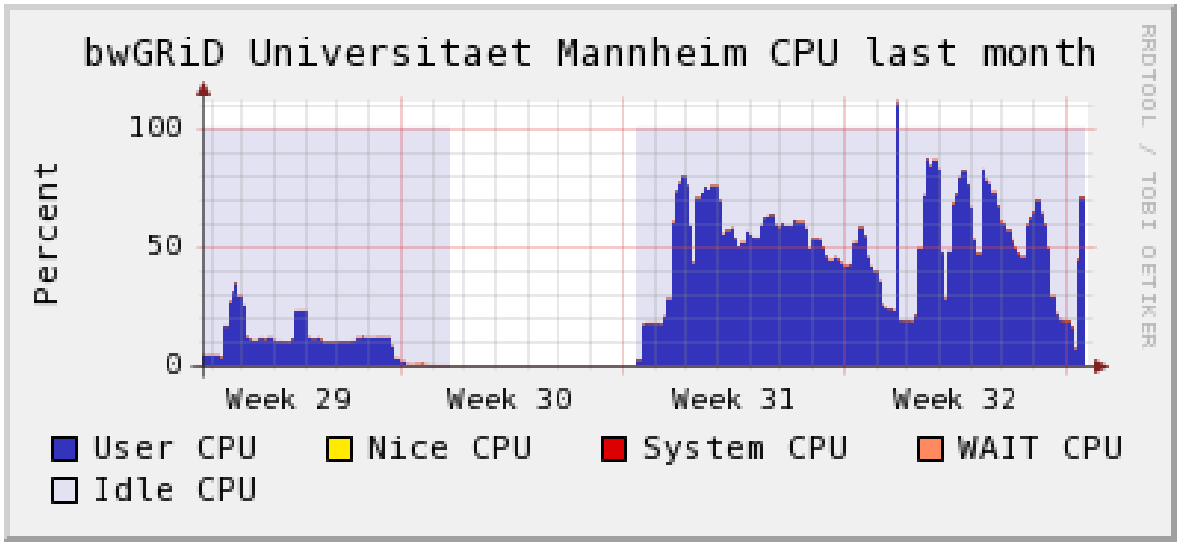
Common Batch System

- Because of high latency: Jobs remain on one site
- Jobs are limited to one cluster, i.e. 140 compute nodes
- PBS Torque with Moab scheduler
- Performance of MPI Jobs with Infiniband communication is not sufficient for 28 km distance (Tests with HPL benchmark)
- Queues: single, normal, itp and Test-Queues

Monitoring Report during activation of the interconnection



Number of processes



Percent CPU Usage

Summary Interconnection

- Network: Obsidian, ADVA and Infiniband are working
- Latency of 145 μsec is very high
- Bandwidth of 930 MB/sec is as expected
- Jobs are limited to one site, because MPI Jobs across the interconnection would slow down
  Main reason: Interconnection is a "shared medium", i.e. all processes use a single line for the whole communication
- Interconnection is useful and stable for a "Single System Cluster" administration
- Better load balance at both sites due to common PBS

Plans for the home directories and the next image

S. Hau (RUM)

Presentation of a Medicine Project

L. Jahnke (Medizinische Fakultt Mannheim)

Discussion of topics for further meetings

Next Meetings (Summer Term 2010)

Dates

- 20. May 2010
- 17. June 2010
- 15. July 2010

Time

- 14:15 – 16:00 ok?
- or 15:15 – 17:00 ?
- or 16:15 – 18:00 ?

Topics

Possible Lectures/Workshops:

- Introduction Batch-System
- Software packages and Parallelization
- Programming in Java, Fortran, C
- Parallelization with MPI/OpenMP
- Grid access and Grid usage

User Contributions:

- ???

Thank you for participating.