A Long-distance InfiniBand Interconnection between two Clusters in Production Use

Sabine Richling, Steffen Hau, Heinz Kredel, Hans-Günther Kruse

IT-Center, University of Heidelberg, Germany
IT-Center, University of Mannheim, Germany

SC'11, State of the Practice, 16. November 2011

Outline

D-Grid and bwGRiD

- bwGRiD Virtual Organization (VO)
  - Community project of the German Grid Initiative D-Grid
  - Project partners are the Universities in Baden-Württemberg
- bwGRiD Resources
  - Compute clusters at 8 locations
  - Central storage unit in Karlsruhe
- bwGRiD Objectives
  - Verifying the functionality and the benefit of Grid concepts for the HPC community in Baden-Württemberg
  - Managing organizational, security, and license issues
  - Development of new cluster and Grid applications

Baden-Württemberg

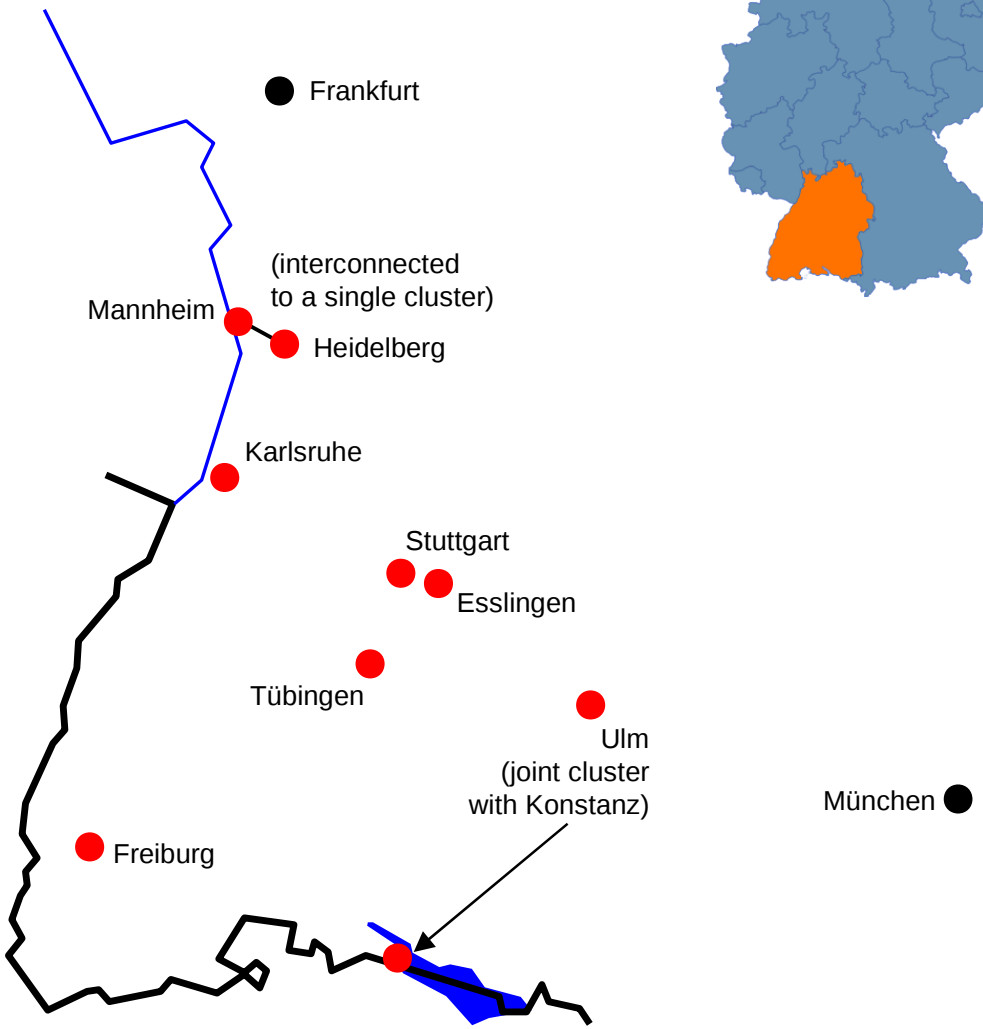
bwGRiD – Resources

Compute Cluster

Site	Nodes
Mannheim	140
Heidelberg	140
Karlsruhe	140
Stuttgart	420
Tübingen	140
Ulm/Konstanz	280
Freiburg	140
Esslingen	180
Total	1580

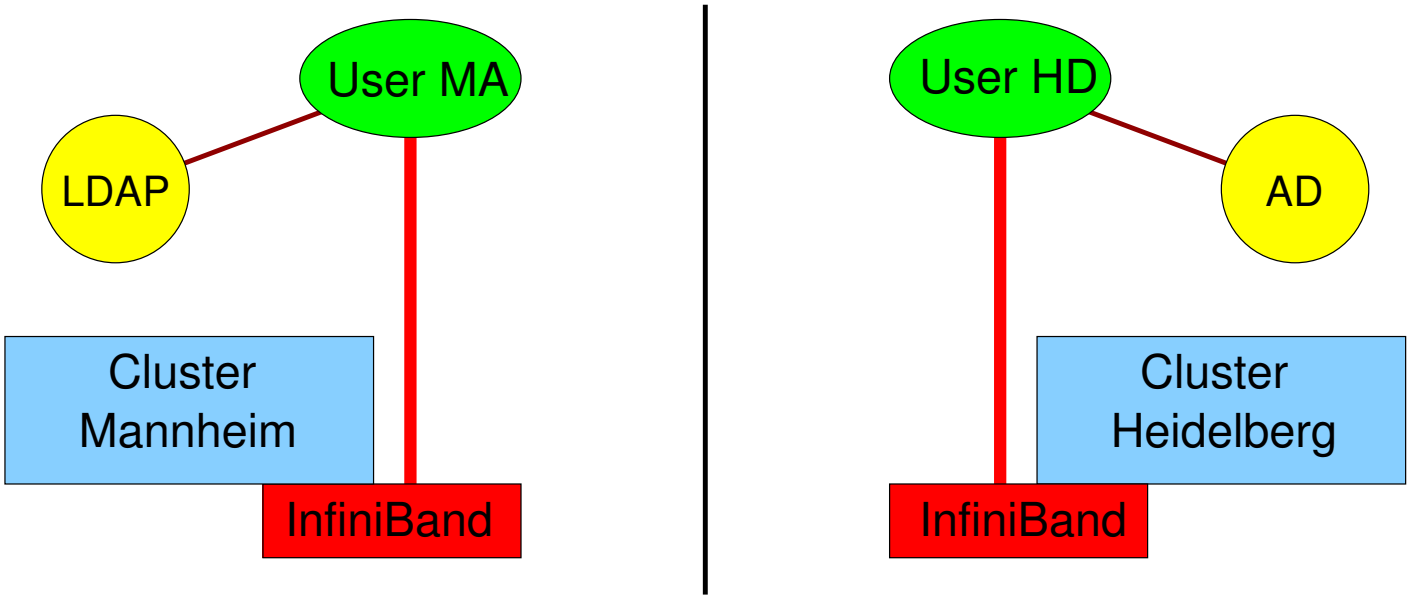
Central Storage

with backup	128 TB
without backup	256 TB
Total	384 TB



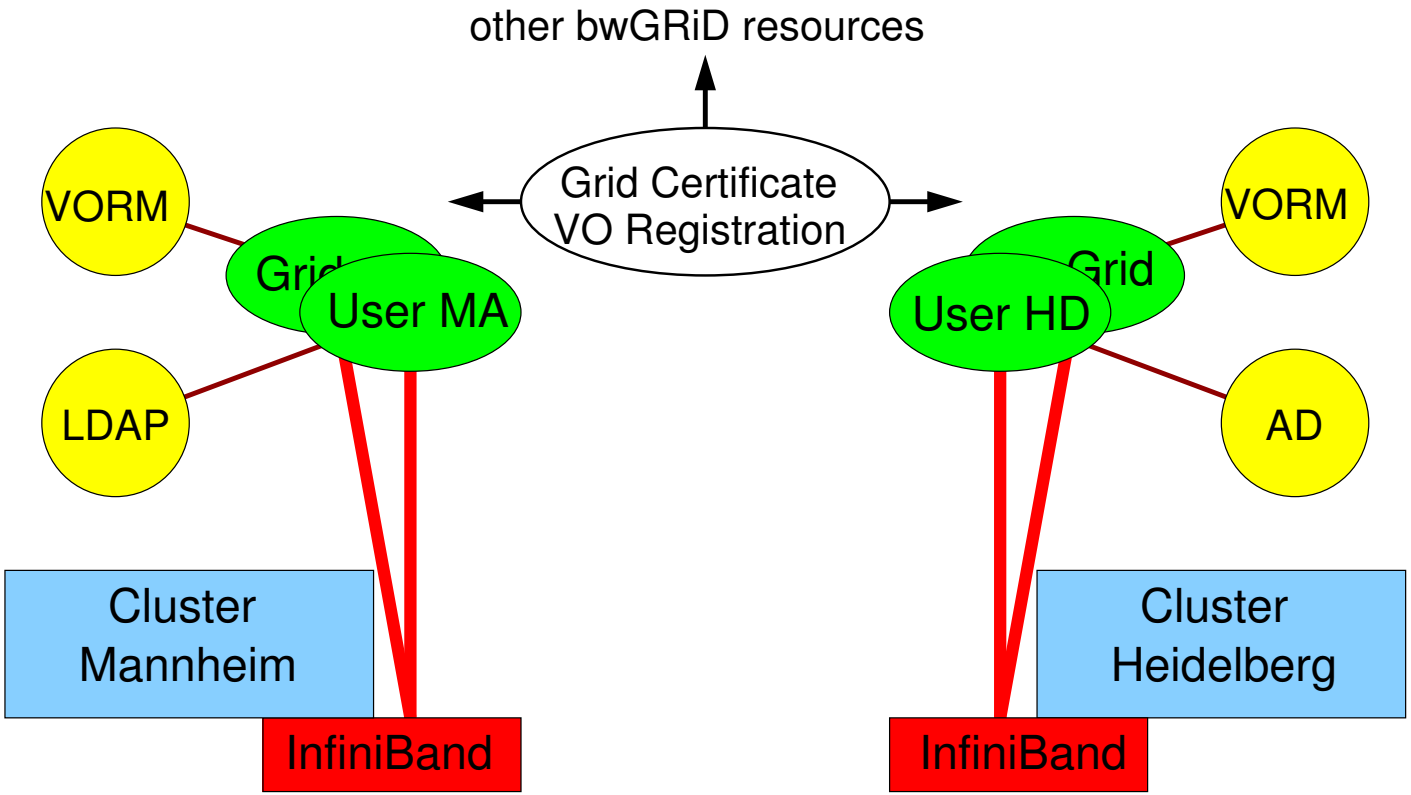
bwGRiD – Situation in MA/HD before interconnection

- Diversity of applications (1–128 nodes per job)
- Many first time HPC users!
- Access with local University Accounts (Authentication via LDAP/AD)



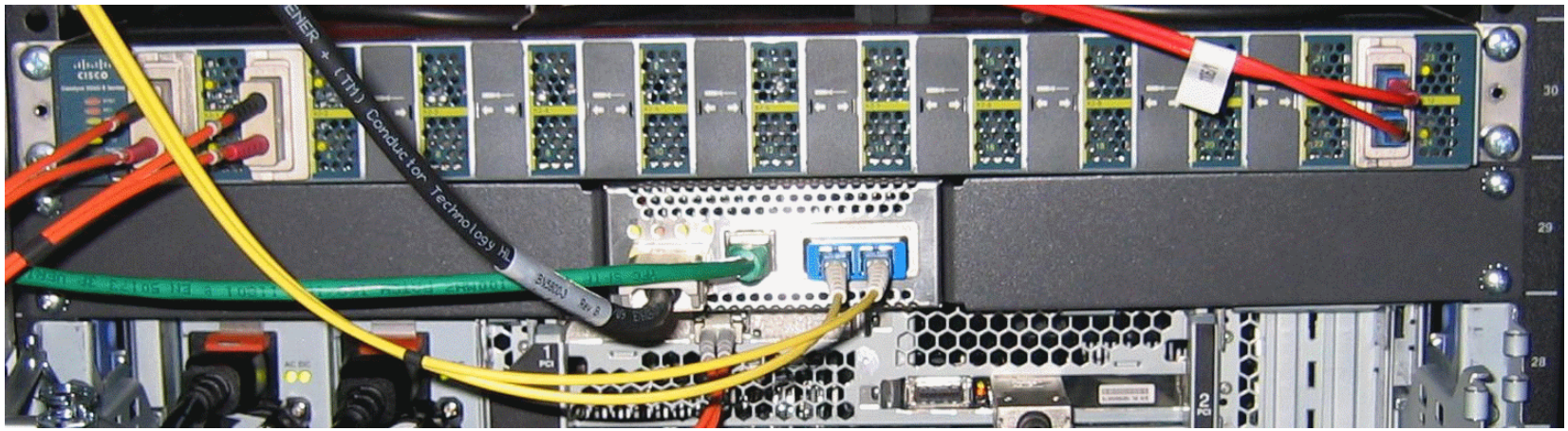
bwGRiD – Situation in MA/HD before interconnection

- Grid certificate allows access to all bwGRiD clusters
- Feasible only for more experienced users



Interconnection of bwGRiD clusters MA/HD

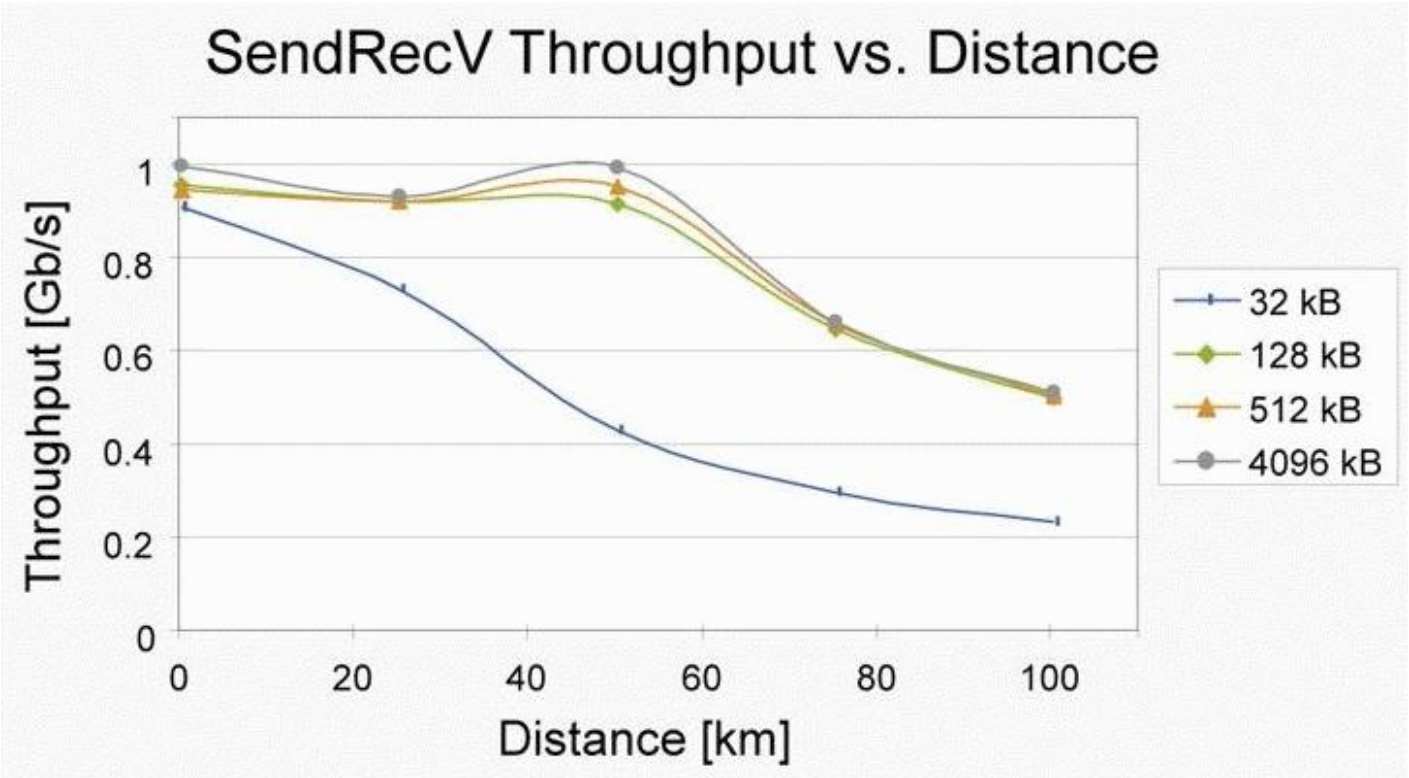
- Proposal in 2008
- Acquisition and Assembly until May 2009
- Running since July 2009
- InfiniBand over Ethernet over fibre optics: Obsidian Longbow adaptor



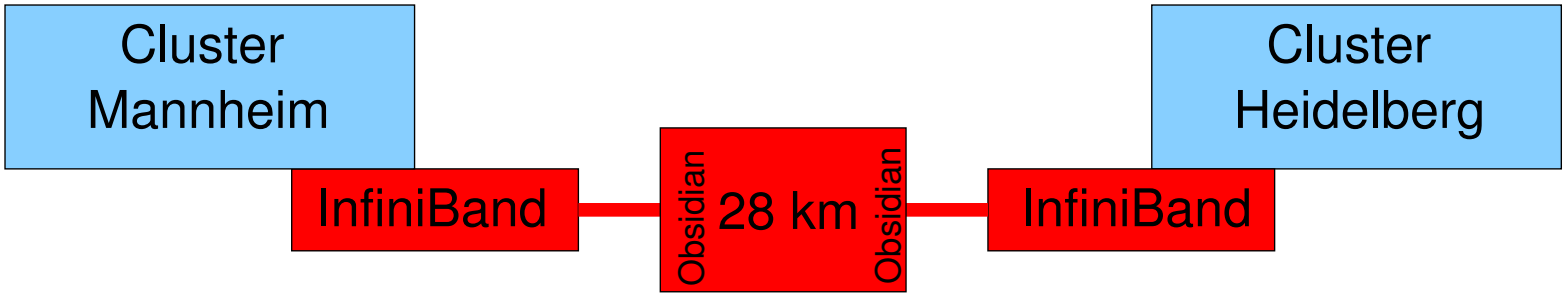
InfiniBand connector (black cable), fibre optic connector (yellow cable)

MPI Performance – Prospects

- Measurements for different distances (HLRS, Stuttgart, Germany)
- Bandwidth 900-1000 MB/sec for up to 50-60 km



MPI Performance – Interconnection MA/HD



Latency is high

145 μsec = 143 μsec light transit time + 2 μsec local latency

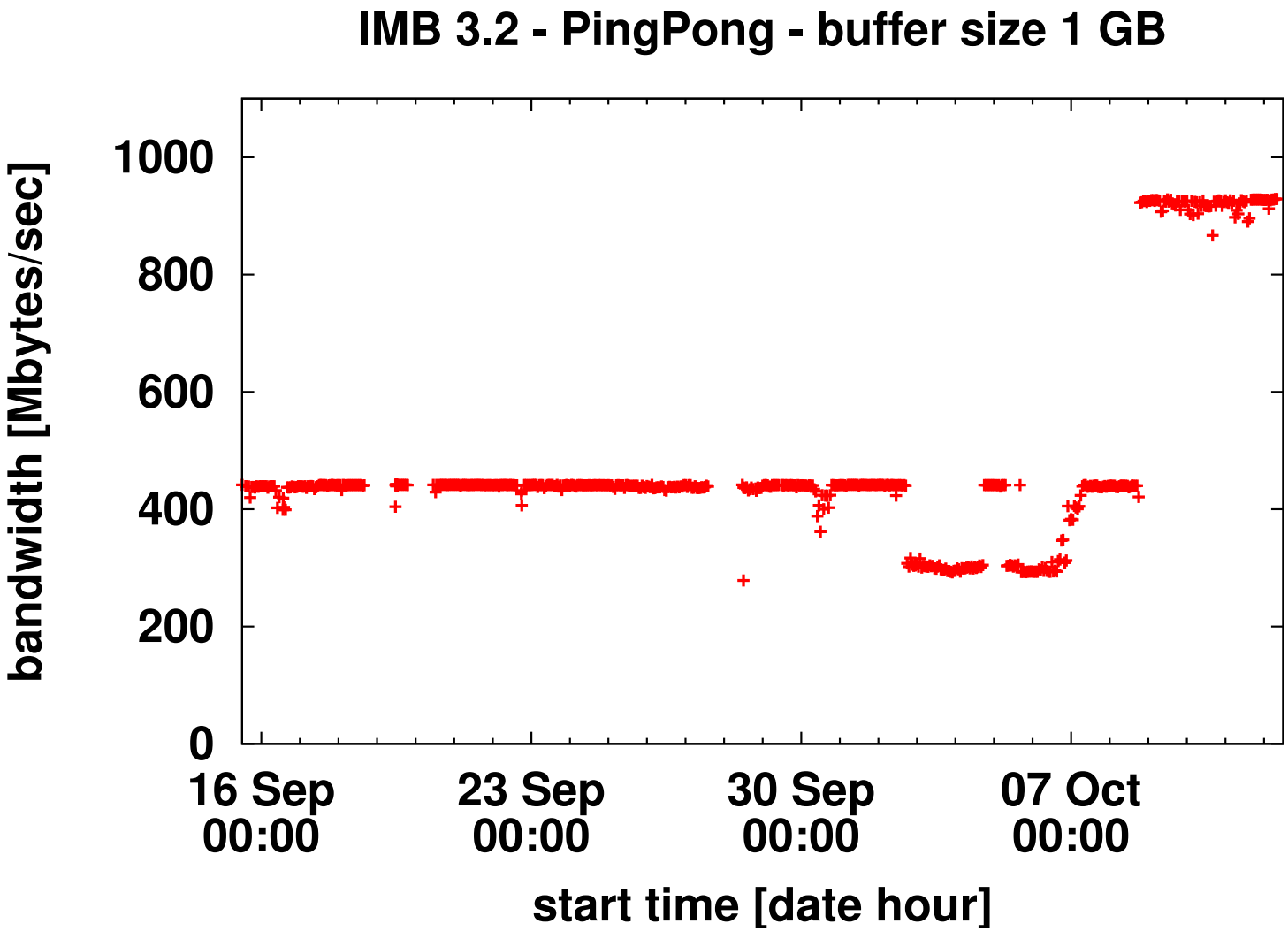
Bandwidth is as expected

about 930 MB/sec (local bandwidth 1200-1400 MB/sec)

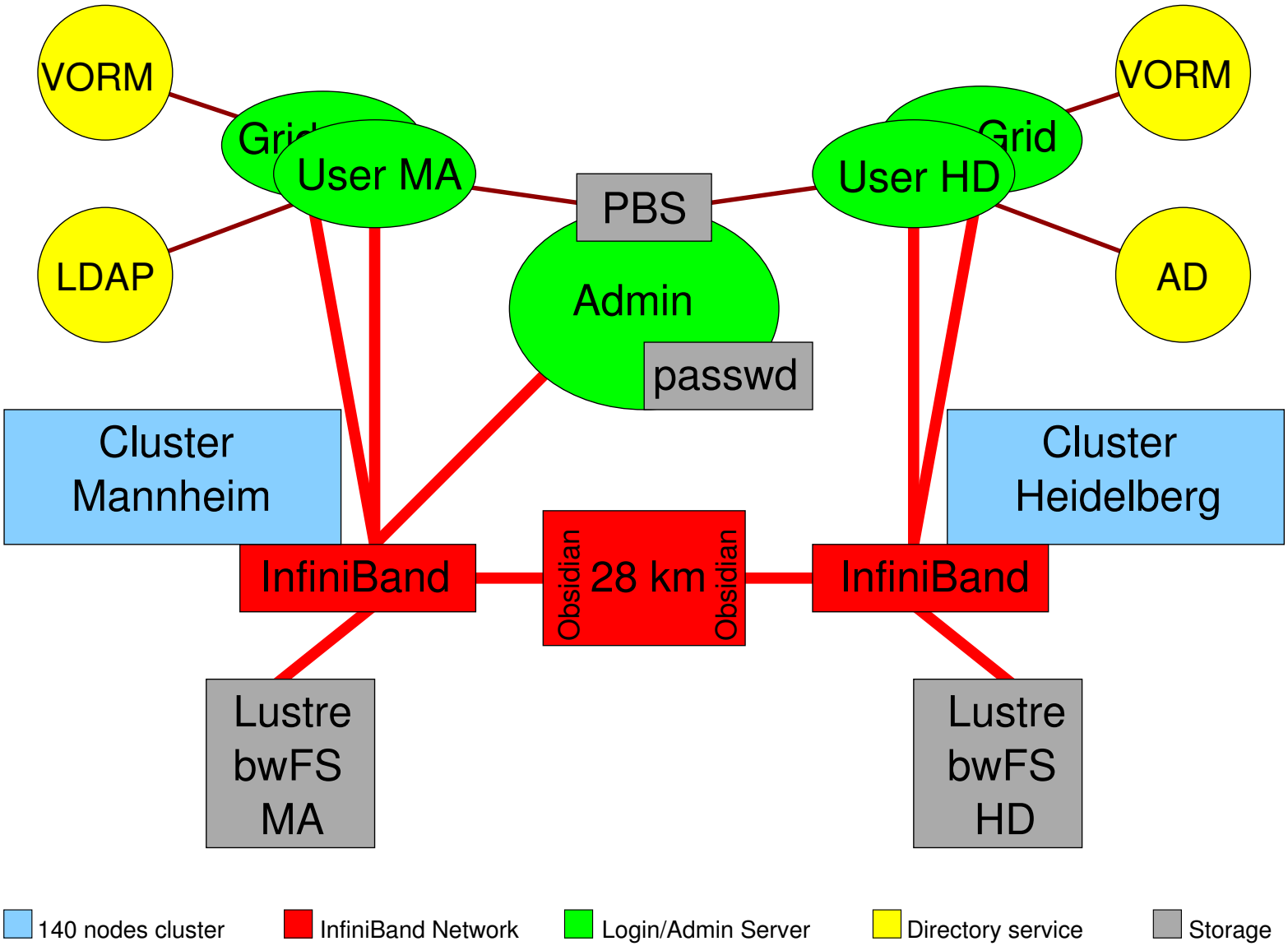
Obsidian needs a license for 40 km

- 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 – Overview



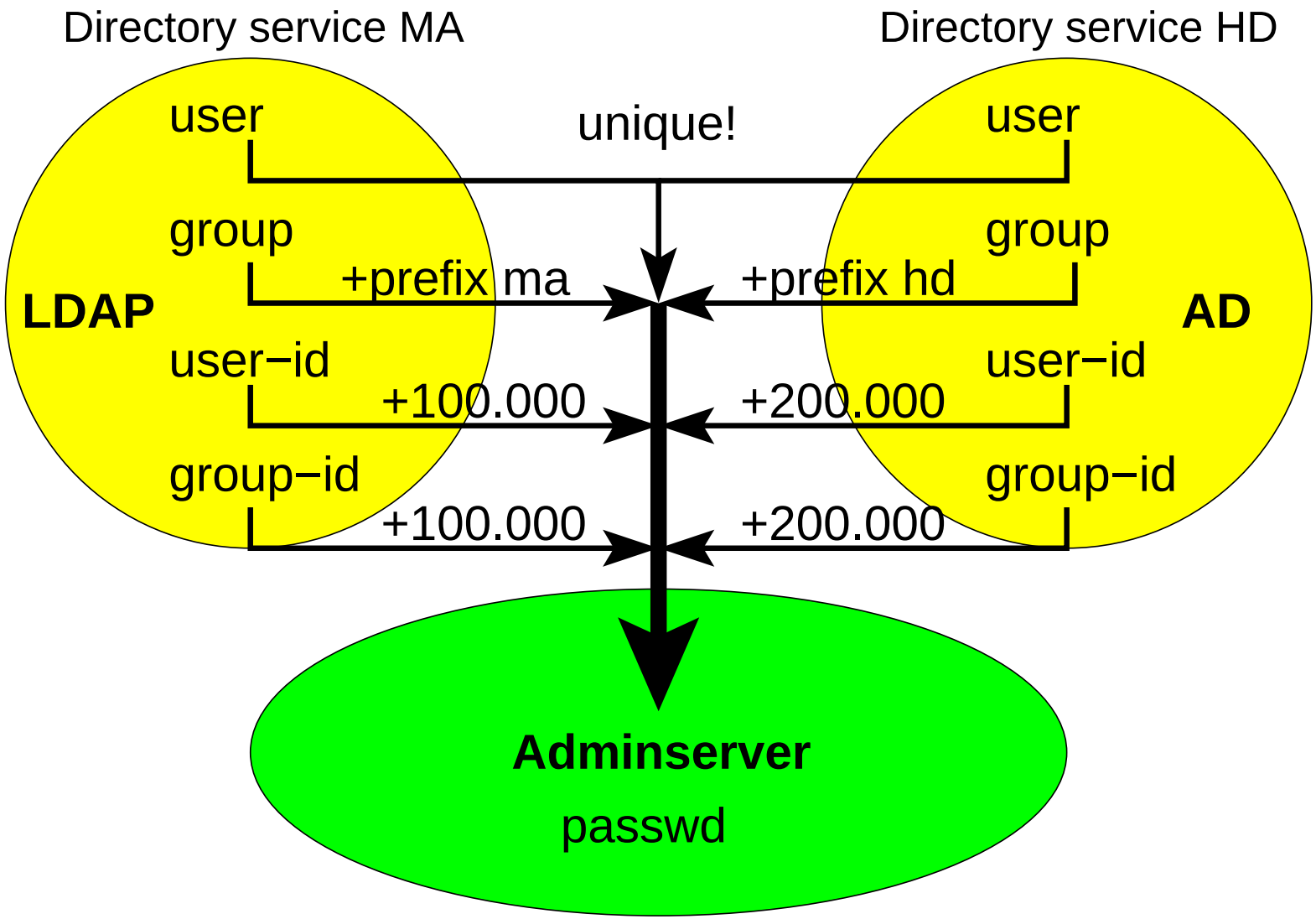
Node Management

- Administration server provides
  - DHCP service for the nodes (MAC-to-IP address configuration file)
  - NFS export for root file system
  - NFS directory for software packages accessible via module utilities
  - queuing and scheduling system
- Node administration
  - adjusted shell scripts originally developed by HLRS
  - IBM management module (command line interface and Web-GUI)

User Management

- Users should have exclusive access to compute nodes
  - user names and user-ids must be unique
  - direct connection to PBS for user authorization via PAM module
- Authentication at the access nodes
  - directly against directory services: LDAP (MA) and AD (HD)
  - or with D-Grid certificate
- Combining information from directory services from both universities
  - Prefix for group names
  - Adding offsets to user-ids and group-ids
  - Activated user names from MA and HD must be different
- Activation process
  - Adding a special attribute for the user in the directory service (for authentication)
  - Updating the user database of the cluster (for authorization)

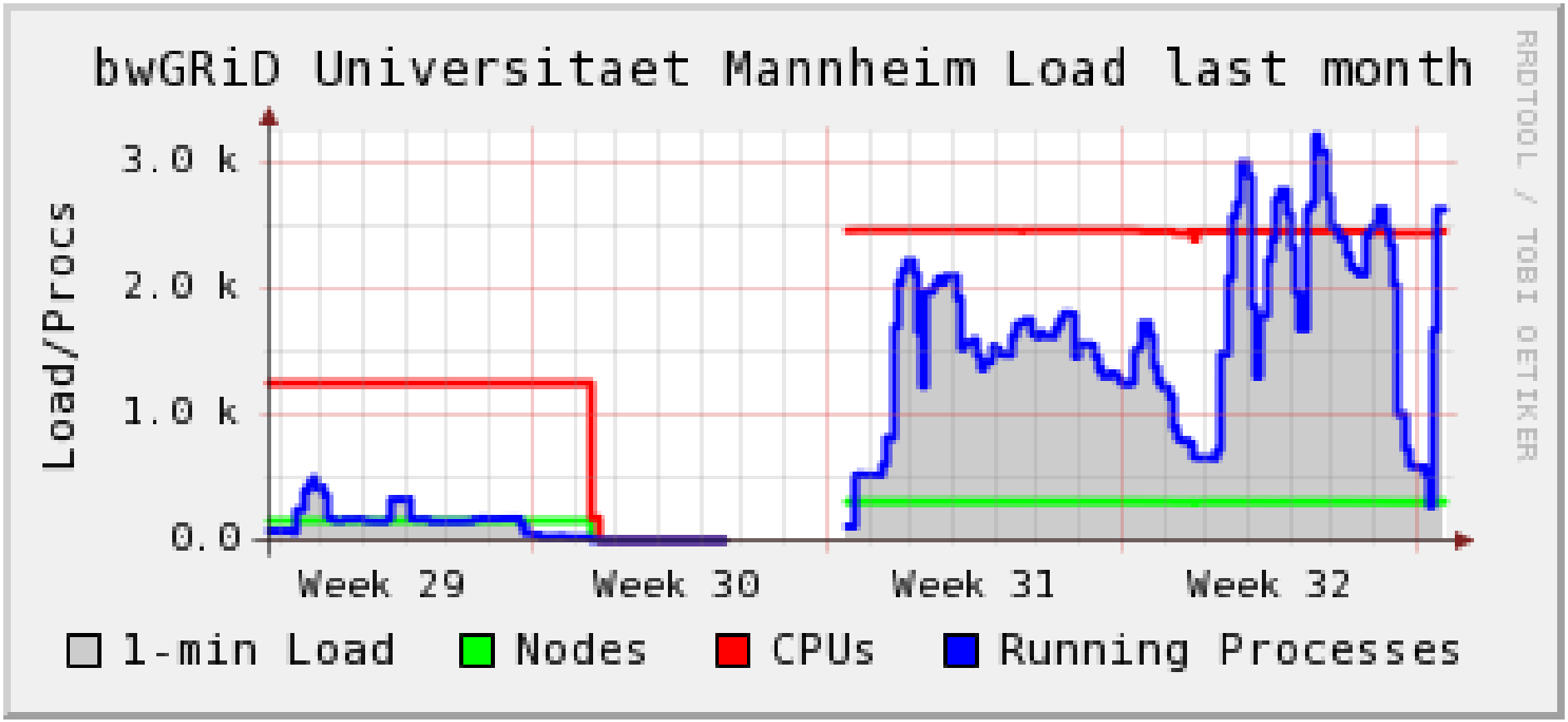
User Management – Generation of configuration files



Job Management

- Interconnection (high latency, limited bandwidth) provides
  - enough bandwidth for I/O operations
  - not sufficient for all kinds of MPI jobs
- Jobs run only on nodes located either in HD or in MA
  (realized with attributes provides by the queuing system)
- Before interconnection
  - In Mannheim: mostly single node jobs $\rightarrow$ free nodes
  - In Heidelberg: many MPI jobs $\rightarrow$ long waiting times
- With interconnection better resource utilization (see Ganglia report)

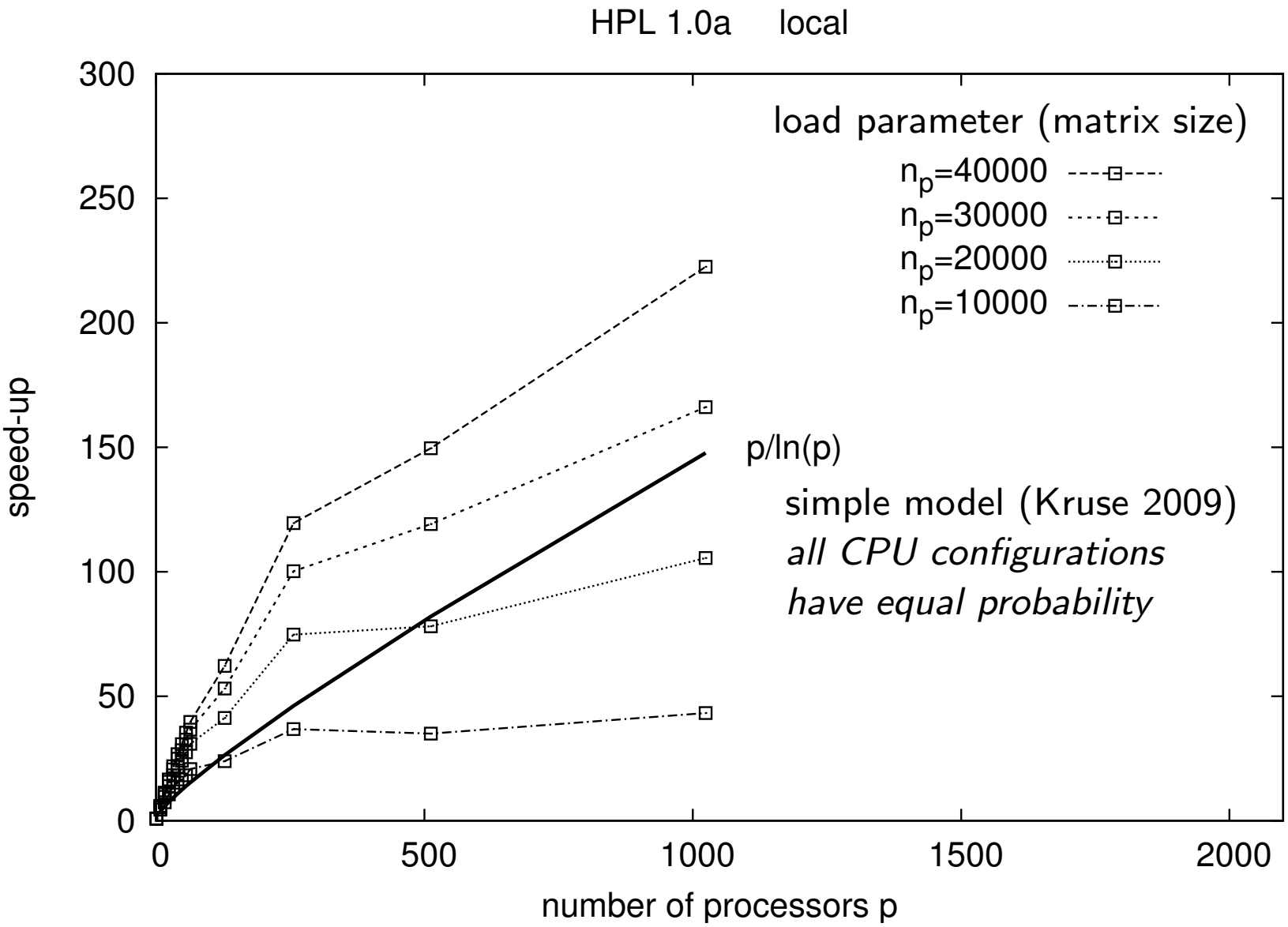
Ganglia Report during activation of the interconnection



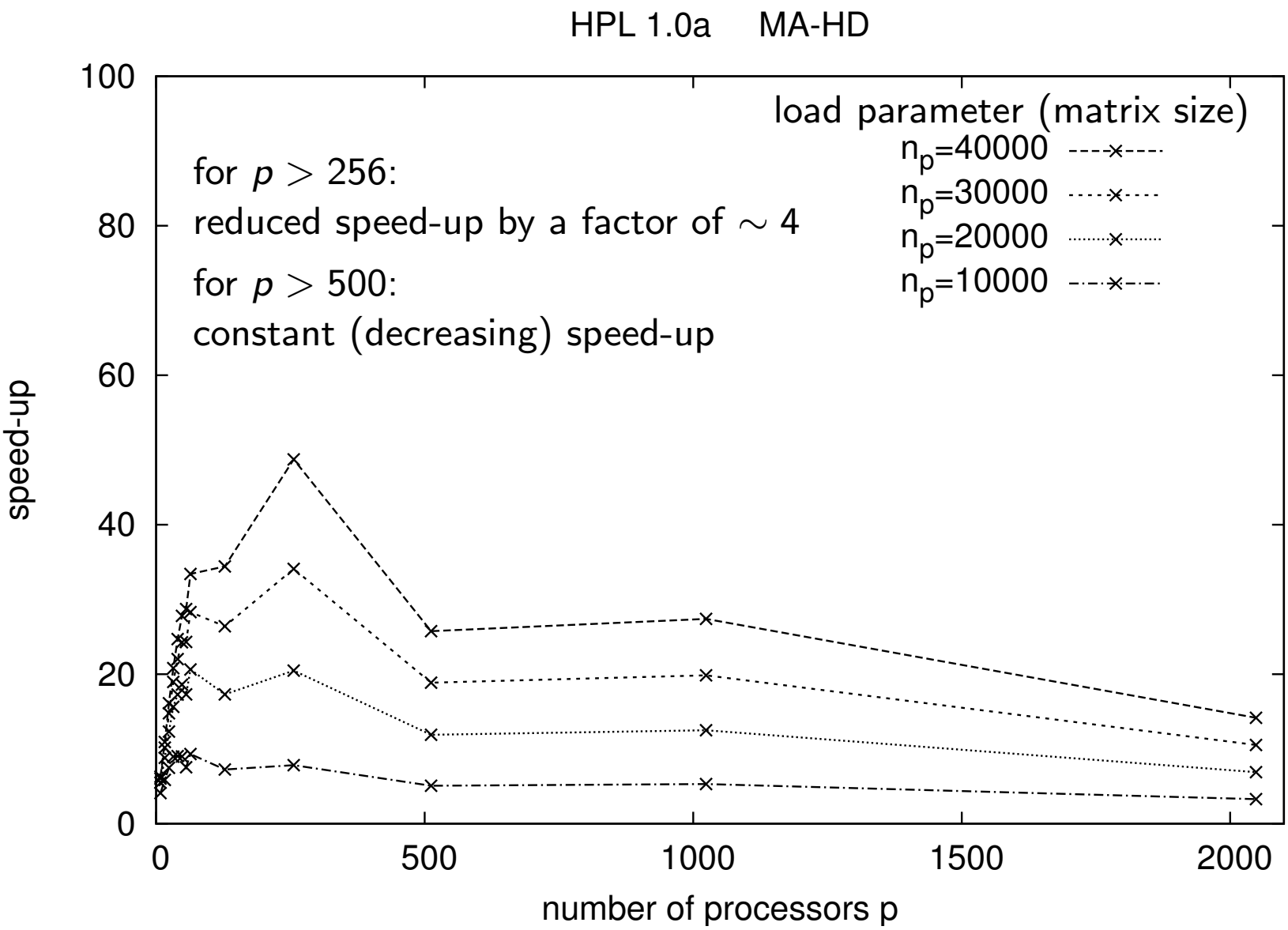
MPI Performance Measurements

- Numerical model
  - High-Performance Linpack (HPL) benchmark
  - OpenMPI
  - Intel MKL
- Model variants
  - Calculations on a single cluster with up to 1024 CPU cores
  - Calculations on the interconnected cluster with up to 2048 CPU cores symmetrically distributed

Results for a single cluster



Results for interconnected cluster



Performance model

Improvement of simple analytical model (Kruse 2009) to analyze the characteristics of the interconnection

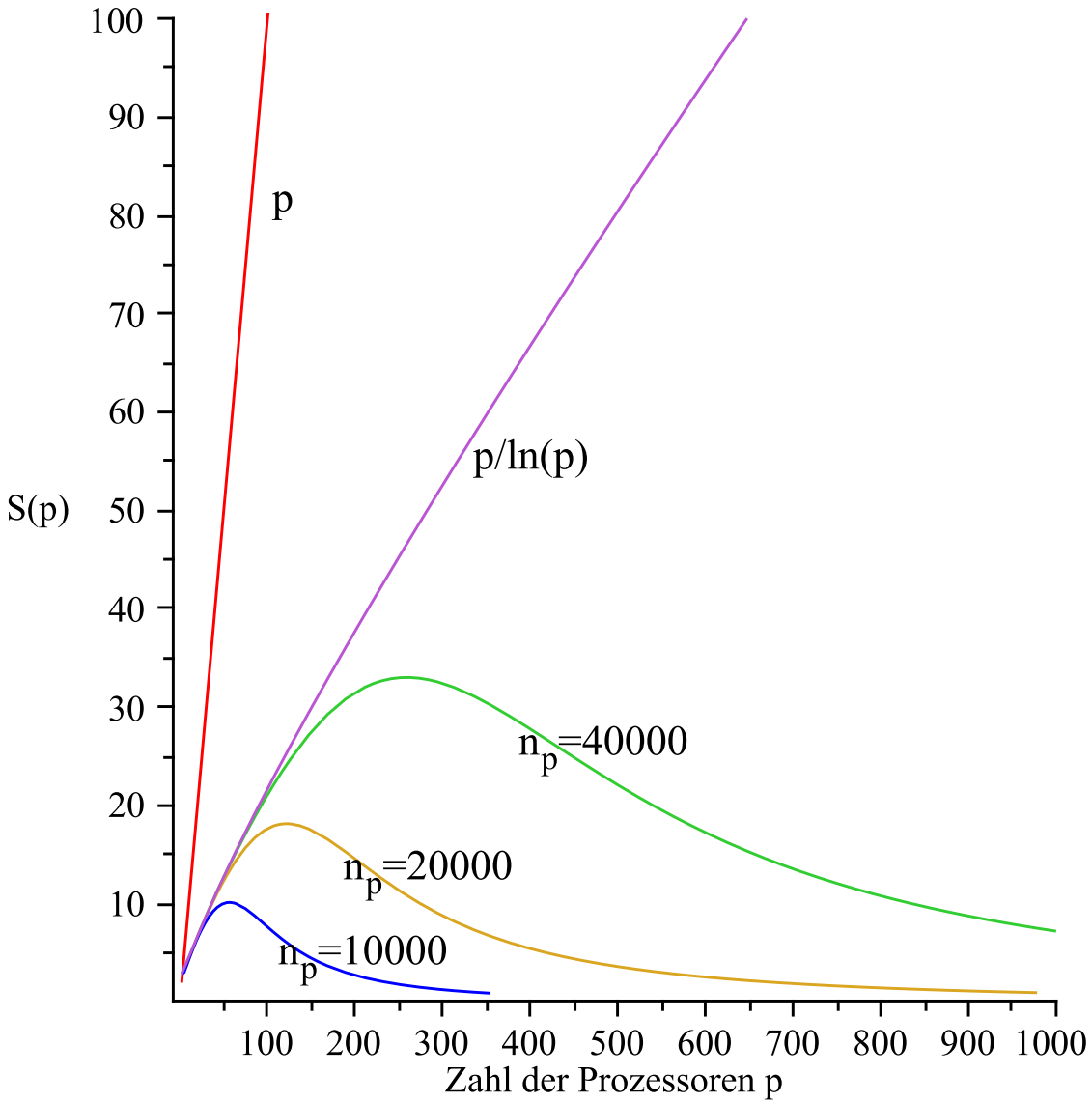
- high latency of 145 μsec
- limited bandwidth of 930 MB/sec (modelled as shared medium)

Result for Speed-up:

$$S(p) \leq \frac{p}{\ln p + \frac{3}{4}\left(\frac{100}{n_p}\right)^3 (1+4p)c(p)}$$

p number of processors
n_p load parameter (matrix size)
$c(p)$ dimensionless function representing the communication topology

Speed-up of the model



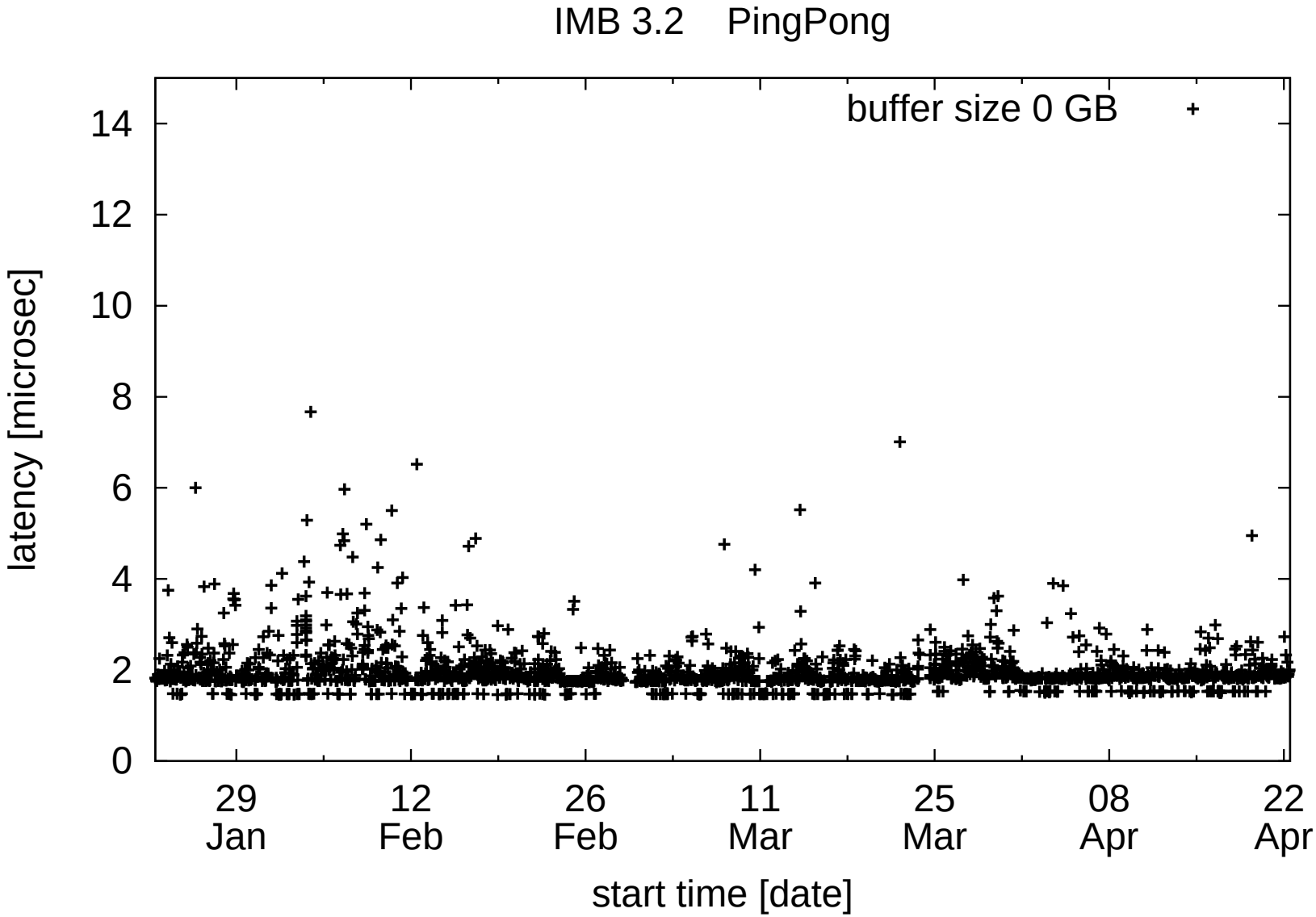
Results:

- Limited bandwidth is the performance bottleneck for shared connection between the clusters
- Double bandwidth:
25 % improvement for
$n_p = 40\,000$
- 100 % improvement with a ten-fold bandwidth

$\Rightarrow$ Jobs run on nodes located either in MA or in HD

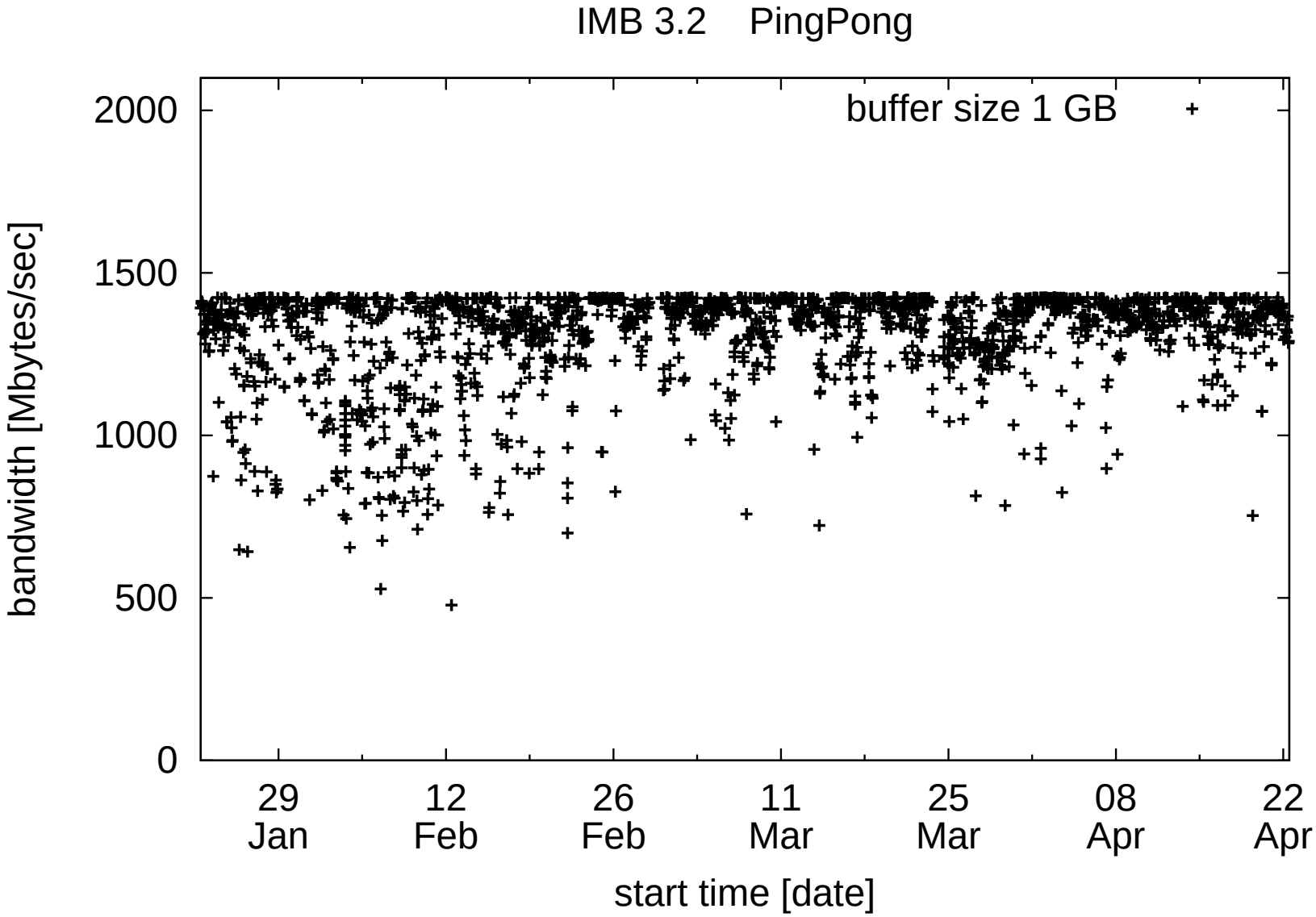
Long-term MPI performance – Latency

between two random nodes in HD or in MA



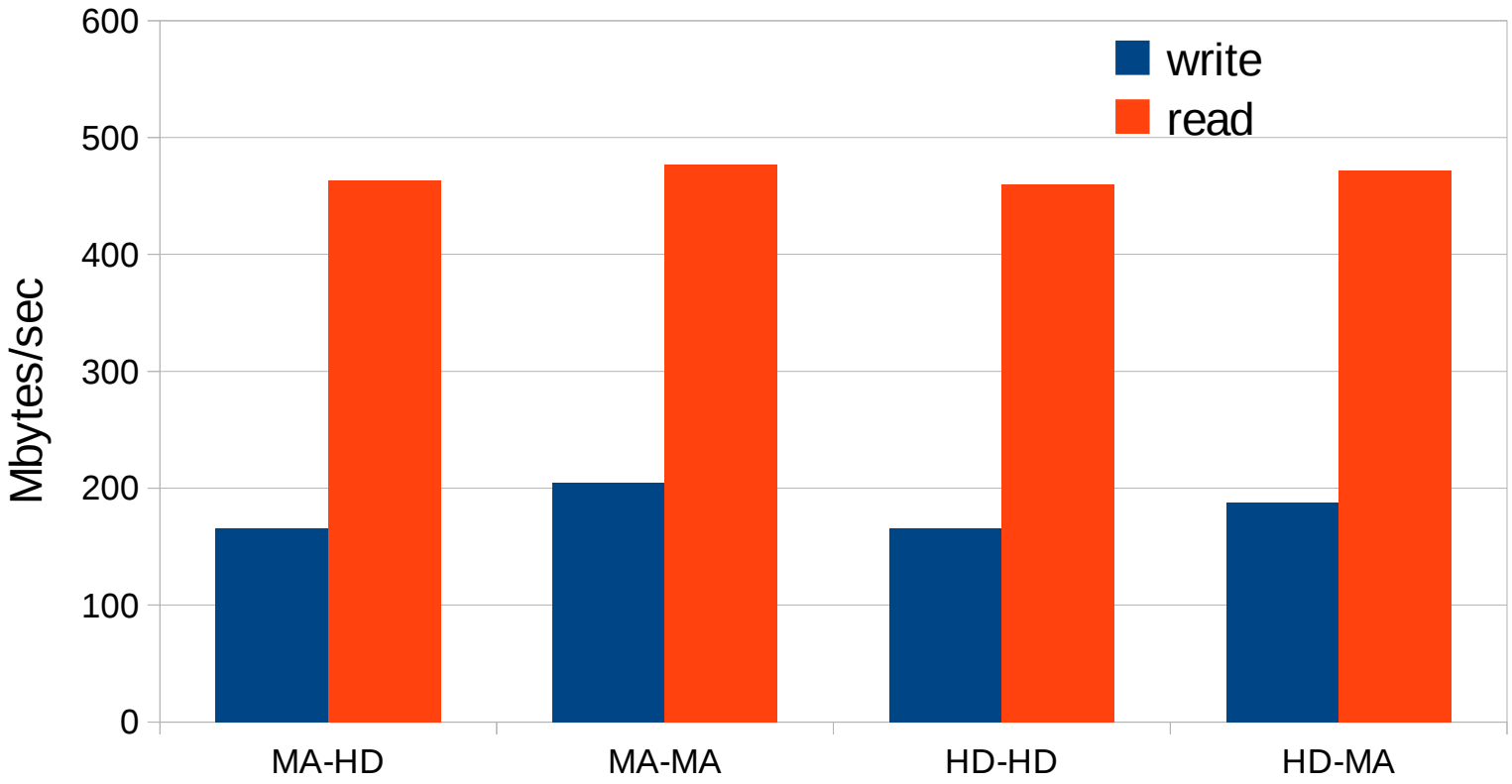
Long-term MPI performance – Bandwidth

between two random nodes in HD or in MA



Storage Access Performance

IOzone benchmark for 32 GB file with records size 4 MB (node – storage)



Summary and Conclusions

- Interconnection network (Obsidian and InfiniBand switches) is stable and works reliable
- Bandwidth of 930 MB/sec is sufficient for Lustre file system access
  - single system administration
  - lower administration costs
  - better load balance
- Setting up a federated authorization is challenging but worthwhile
  - Further reduction of administration costs
  - Lower access barrier for potential users
- Characteristics of the interconnection is not sufficient for all kinds of MPI jobs $\rightarrow$ Jobs remain on one side of the combined cluster
  Possible improvements:
  - Adding more parallel fibre lines (very expensive)
  - Investigation of different job scheduler configurations