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

1 Background
   - D-Grid and bwGRiD
   - bwGRiD MA/HD

2 Interconnection of two bwGRiD clusters

3 Cluster Operation
   - Node Management
   - User Management
   - Job Management

4 Performance
   - MPI Performance
   - Storage Access Performance

5 Summary and Conclusions
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
bwGRiD – Resources

**Compute Cluster**

<table>
<thead>
<tr>
<th>Site</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mannheim</td>
<td>140</td>
</tr>
<tr>
<td>Heidelberg</td>
<td>140</td>
</tr>
<tr>
<td>Karlsruhe</td>
<td>140</td>
</tr>
<tr>
<td>Stuttgart</td>
<td>420</td>
</tr>
<tr>
<td>Tübingen</td>
<td>140</td>
</tr>
<tr>
<td>Ulm/Konstanz</td>
<td>280</td>
</tr>
<tr>
<td>Freiburg</td>
<td>140</td>
</tr>
<tr>
<td>Esslingen</td>
<td>180</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1580</strong></td>
</tr>
</tbody>
</table>

**Central Storage**

<table>
<thead>
<tr>
<th>Type</th>
<th>TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>with backup</td>
<td>128</td>
</tr>
<tr>
<td>without backup</td>
<td>256</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>384</strong></td>
</tr>
</tbody>
</table>

Richling, Hau, Kredel, Kruse (URZ/RUM)
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

Diagram:
- Grid certificate
- VO Registration
- Other bwGRiD resources
- InfiniBand
- Cluster Mannheim
- Cluster Heidelberg

Richling, Hau, Kredel, Kruse (URZ/RUM)
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

![Graph showing SendRecV Throughput vs. Distance](image.png)
Latency is high

$145 \, \mu \text{sec} = 143 \, \mu \text{sec light transit time} + 2 \, \mu \text{sec local latency}$

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

![Graph showing MPI Bandwidth with dates and times]
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

Directory service MA

LDAP

user

user−id

+100.000

group−id

+100.000

+prefix ma

unique!

Directory service HD

AD

user

user−id

+200.000

group−id

+200.000

+prefix hd

Adminserver

passwd
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 → free nodes
- In Heidelberg: many MPI jobs → long waiting times

With interconnection better resource utilization (see Ganglia report)
Ganglia Report during activation of the interconnection
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

\begin{center}
\begin{tikzpicture}
\begin{axis}[
    xlabel={number of processors $p$},
    ylabel={speed-up},
    xmin=0, xmax=2000,
    ymin=0, ymax=300,
    xtick={0,500,1000,1500,2000},
    ytick={0,50,100,150,200,250,300},
    legend pos=north west,
]

% Add the lines for different $n_p$ values
\addplot[style={solid,mark=square}] coordinates {%(40000,200)(30000,150)(20000,100)(10000,50)};
\addplot[style={dashed,mark=square}] coordinates {%(40000,250)(30000,200)(20000,150)(10000,100)};
\addplot[style={dotted,mark=square}] coordinates {%(40000,300)(30000,250)(20000,200)(10000,150)};
\addplot[style={dashdotted,mark=square}] coordinates {%(40000,350)(30000,300)(20000,250)(10000,200)};
\addplot[style={dashed,dotted,mark=square}] coordinates {%(40000,400)(30000,350)(20000,300)(10000,250)};

\legend{$n_p=40000$,$n_p=30000$,$n_p=20000$,$n_p=10000$}
\end{axis}
\end{tikzpicture}
\end{center}

Load parameter (matrix size)

- $n_p=40000$
- $n_p=30000$
- $n_p=20000$
- $n_p=10000$

Simple model (Kruse 2009)

All CPU configurations have equal probability.
Results for interconnected cluster

HPL 1.0a     MA-HD

load parameter (matrix size)

for $p > 256$:
  reduced speed-up by a factor of $\sim 4$

for $p > 500$:
  constant (decreasing) speed-up

Richling, Hau, Kredel, Kruse (URZ/RUM)  Long-distance InfiniBand Connection  Seattle, November 2011
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 $n_p = 40000$
- 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

latency [microsec]
start time [date]
IMB 3.2 PingPong
buffer size 0 GB
Long-term MPI performance – Bandwidth

between two random nodes in HD or in MA

IMB 3.2    PingPong

buffer size 1 GB

Richling, Hau, Kredel, Kruse (URZ/RUM)

Long-distance InfiniBand Connection

Seattle, November 2011
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 → Jobs remain on one side of the combined cluster
  Possible improvements:
  - Adding more parallel fibre lines (very expensive)
  - Investigation of different job scheduler configurations