MEGWARE HPC Cluster am LRZ – eine mehr als 12-jährige Zusammenarbeit

Prof. Dieter Kranzlmüller (LRZ)
LRZ HPC-Systems at the End of the UNIX-Era (Years 2000-2002)

German national supercomputer Hitachi SR800 pseudo vector system with
- 168 SMP nodes
- 8 +1 CPUs per node
- 1376 GB memory
- 5000 GB disk
- 2016 GF peak performance

Bavarian vector computer Fujitsu VPP vector system with
- 52 vector CPUs
- 104 GB memory
- 1214 GB disk
- 114.4 GF peak performance
<table>
<thead>
<tr>
<th>HPC System Type</th>
<th>System Details</th>
<th>Performance</th>
<th>Notes</th>
</tr>
</thead>
</table>
| Bavarian MPP system IBM SP2 | • 77 nodes  
• 16.7 GB memory  
• 334 GB disk  
• 20.7 GF peak performance | | |
| Bavarian vector computer CRAY T90 | • 4 vector CPUs  
• 1.0 GB memory  
• 145 GB disk  
• 7.2 GF peak performance  
• abandonment in 2001 | | |
| Bavarian large shared memory HPC system IBM p690 | • 8 Power 4 CPUs  
• 32 GB memory  
• 936 GB disk  
• 42 GF peak performance | | |
• LRZ home-made Linux Cluster for Munich Universities:
  - 2 dual Pentium II nodes
  - 17 dual Pentium III nodes (9 nodes with Myrinet communication network)
  - 2 quad Pentium III-Xeon nodes
  - 6 Pentium IV nodes
  - 56 GB memory
  - 70 GB disk
  - 62 GF peak performance

• Vendors: FMS, DELL and Synchron
The LRZ HPC Pyramid as HPC Service Concept
2003: Replacement of IBM SP2 by MEGWARE IA32 and IA64 Linux Cluster

- MEGWARE IA32 cluster
  - 105 nodes with Intel 3.06 GHz Pentium4 processor, 2 GB memory
  - Gb Ethernet network
  - 643 GF peak performance
  - #341 in June 2013 Top500 list

- MEGWARE IA64 cluster
  - 17 quad Itanium2 (Madison) nodes with 8 GB memory
  - Myrinet 2000 communication network
  - 354 GF peak performance
  - 1.5 TB disk space (PVFS)
2004: Replacement of Fujitsu VPP by IA64 Linux Cluster and
128-way sgi Altix 3700Bx2

- Sgi Altix 3700Bx2
  - 128 Itanium2 (Madison) processors
  - 512 GB memory
  - NUMALink3 network
  - 819 GF peak performance
  - 10 TB disk space
- MEGWARE IA64 cluster
  - 17 quad Itanium2 (Madison) nodes with 8 GB memory and Myrinet 2000 communication network
  - 67 dual Itanium2 (Madison) nodes with 8 GB memory and Gb Ethernet communication network
  - 1677 GF peak performance
  - 12 TB disk space (Lustre)
The LRZ Linux Cluster in the Year 2005

[Diagram of the LRZ Linux Cluster with details on hardware components and network connections.]
2006: Move from Munich to Garching and Consolidation of HPC Operating Systems and Platforms
2007: Further Extension of the Linux Cluster

- Sgi Altix 4700
  - 256 Itanium Montecito processors
  - 1024 GB memory
  - NUMALink4 network
  - 1638 GF peak performance
  - 6 TB disk space
- x86-64 cluster
  - 232 MEGWARE AMD x86-64 dual core nodes
  - 99 MEGWARE Intel x86-64 quad core nodes
  - 38 MEGWARE AMD dual core quad socket nodes
  - 15 MEGWARE dCache server with 150 TB total disk space
- 15 Sun X4600 dual core eight socket systems
- 182 TB of sgi disk storage for Lustre
### LRZ HPC systems in the Year 2008

<table>
<thead>
<tr>
<th>System</th>
<th>Anzahl Cores</th>
<th>Maximale Rechenleistung (TFlop/s)</th>
<th>Hauptspeicher (TByte)</th>
<th>Platten (TByte)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HLRB II</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGI Altix 4700</td>
<td>9728</td>
<td>62.3</td>
<td>39.1</td>
<td>660</td>
</tr>
<tr>
<td>EM64T/Opteron</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Xeon, Opteron)</td>
<td>2-fach</td>
<td>50</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>4-fach</td>
<td>1188</td>
<td>11.9</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>8-fach</td>
<td>368</td>
<td>3.9</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>16-fach</td>
<td>240</td>
<td>2.7</td>
<td>1.0</td>
</tr>
<tr>
<td>LCG Tier-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-fach</td>
<td>20</td>
<td>0.1</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>4-fach</td>
<td>796</td>
<td>7.8</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>8-fach</td>
<td>544</td>
<td>5.4</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>Linux Cluster</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IA64 Itanium</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-fach</td>
<td>134</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>4-fach</td>
<td>48</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>8-fach</td>
<td>16</td>
<td>0.1</td>
<td>0.032</td>
</tr>
<tr>
<td>SGI-Altix</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128-fach SMP</td>
<td>128</td>
<td>0.8</td>
<td>0.5</td>
<td>182 + 11</td>
</tr>
<tr>
<td>SGI-Altix</td>
<td>256</td>
<td>1.6</td>
<td>1.0</td>
<td>182 + 6</td>
</tr>
<tr>
<td><strong>Teilsumme</strong></td>
<td>582</td>
<td>3.6</td>
<td>2.4</td>
<td>199</td>
</tr>
<tr>
<td><strong>Summe Cluster</strong></td>
<td>3788</td>
<td>35.7</td>
<td>9.8</td>
<td>529</td>
</tr>
</tbody>
</table>
• Most **air-cooled datacenters** are **inefficient**. Cooling needs as much energy as IT equipment and both are thrown-away.

• Provocative: datacenter is a huge **“heater with integrated logic.”**

• **PUE of new LRZ data center ~ 1.5**
LRZ Activities to enhance the Power and Cooling Effectiveness of its Data Centre #1

- Use Total Cost of Ownership (TCO) as an important evaluation criteria in procurements
  - Invest and maintenance
  - Power bill (incl. cooling)
  - Total power cooling of components for the calculation of total IT operation costs

- Use of virtualization techniques (VMware)

- Improve PUE
LRZ Activities to further enhance the Cooling Effectiveness of its Data Centre #2

Implementation of a cold and hot aisle containment which is compatible with the argon fire extinguishing concept

Use of additional cold air ducts at power intensive racks (10 kW)

Installation of a room neutral and direct liquid cooled rack solution for very high power densities > 15 kW per rack
Indirect Liquid Cooled Rack Solutions

- Room neutral
- Better cooling efficiencies due to reduced air throw distances
- Optimal cold/hot aisle confinement

Rear Door Heat Exchanger

Closed Racks with Integrated Heat Exchangers
Air versus Water Cooling

Air cooling is the de-facto standard

But:

<table>
<thead>
<tr>
<th></th>
<th>Air</th>
<th>Water</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal Conductivity</td>
<td>0.026 W/(m*K)</td>
<td>0.56 W/(m*K)</td>
<td>21.5 x</td>
</tr>
<tr>
<td>Thermal Capacity</td>
<td>1.00 J/(g*K)</td>
<td>4.18 J/(g*K)</td>
<td>4.18 x</td>
</tr>
</tbody>
</table>

Water as coolant allows higher inlet temperatures (free cooling!)
Water enables better heat reuse
Air Cooling versus Direct Liquid Cooling

![Bar chart showing CPU Package Temperature for different cooling methods]

- **Air, 23°C**: 51
- **Water, 30°C**: 37
- **Water, 40°C**: 48
- **Water, 50°C**: 56

Cooling still sufficient with >40°C inlet temperature! Enables year-round free cooling (no chillers!)
2009-2011: Construction of New Building with Warm Water Cooling Loops & Procurement of Direct Warm Water Cooled HPC Systems

- Heat flux > 90% to water; very low chilled water requirement
- Power advantage over air-cooled node:
  - Warm water cooled ~10%
    (cold water cooled ~15%)
  - due to lower $T_{\text{components}}$ and no fans
- Typical operating conditions: $T_{\text{air}} = 25 - 30^\circ$ C, $T_{\text{water}} = 18 - 45^\circ$ C
2011: Delivery and Installation of CooLMUC

- The world's first AMD-based direct water-cooled cluster with
  - 178 nodes (2x8 core AMD Magny Cour 2.0 GHz CPUs and 16 GByte RAM per node)
  - IB QDR network
  - Thorough power monitoring for compute & cooling hardware
  - Completely closed racks (no dependence on room air conditioning)
  - Reuse of waste-heat for cooling through a SorTech adsorption chiller
Weitere Details zu CooLMUC → Vortrag von Herrn Wilde
MEGWARE HPC Cluster am LRZ – eine mehr als 12-jährige Zusammenarbeit

● Fazit

- MEGWARE geht auf Kundenwünsche ein und ist in der Lage auch sehr innovative HPC-Lösungen anzubieten
- LRZ ist mit den HPC-Lösungen von MEGWARE und dem MEGWARE-Support sehr zufrieden
  - Gute partnerschaftliche Arbeitsatmosphäre
  - Schnelle Reaktionszeiten
  - Hohe HPC-Expertise