User Support for Extreme Scale Computing at the LRZ

Dieter Kranzlmüller

Munich Network Management Team
Ludwig-Maximilians-Universität München (LMU) &
Leibniz Supercomputing Centre (LRZ)
of the Bavarian Academy of Sciences and Humanities
MNM Team
Munich Network Management Team
Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

With 156 employees + 38 extra staff for more than 90,000 students and for more than 30,000 employees including 8,500 scientists.
Computer Centre for all Munich Universities

IT Service Provider:
• Munich Scientific Network (MWN)
• Web servers
• e-Learning
• E-Mail
• Groupware
• Special equipment:
  • Virtual Reality Laboratory
  • Video Conference
  • Scanners for slides and large documents
  • Large scale plotters

IT Competence Centre:
• Hotline and support
• Consulting (security, networking, scientific computing, ...)
• Courses (text editing, image processing, UNIX, Linux, HPC, ...)
The Munich Scientific Network (MWN)
Regional Computer Centre for all Bavarian Universities

Computer Centre for all Munich Universities
5-sided projection room +
large-scale high-resolution powerwall
- National Supercomputing Centre
- Regional Computer Centre for all Bavarian Universities
- Computer Centre for all Munich Universities
Gauss Centre for Supercomputing (GCS)

- Combination of the 3 German national supercomputing centers:
  - John von Neumann Institute for Computing (NIC), Jülich
  - High Performance Computing Center Stuttgart (HLRS)
  - Leibniz Supercomputing Centre (LRZ), Garching n. Munich

- Founded on 13. April 2007

- Hosting member of PRACE (Partnership for Advanced Computing in Europe)
- European Supercomputing Centre

- National Supercomputing Centre

- Regional Computer Centre for all Bavarian Universities

- Computer Centre for all Munich Universities
SuperMUC @ LRZ

Video: SuperMUC rendered on SuperMUC by LRZ

http://youtu.be/OlAS6iiqWrQ
## Top 500 Supercomputer List (June 2012)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer/Year Vendor</th>
<th>Cores</th>
<th>$R_{\text{max}}$</th>
<th>$R_{\text{peak}}$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/NNSA/LLNL United States</td>
<td>Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom / 2011 IBM</td>
<td>1572864</td>
<td>16324.75</td>
<td>20132.66</td>
<td>7890.0</td>
</tr>
<tr>
<td>2</td>
<td>RIKEN Advanced Institute for Computational Science (AICS) Japan</td>
<td>K computer, SPARC64 Vllfx 2.0GHz, Tofu interconnect / 2011 Fujitsu</td>
<td>705024</td>
<td>10510.00</td>
<td>11280.38</td>
<td>12659.9</td>
</tr>
<tr>
<td>3</td>
<td>DOE/SC/Argonne National Laboratory United States</td>
<td>Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM</td>
<td>786432</td>
<td>8162.38</td>
<td>10066.33</td>
<td>3945.0</td>
</tr>
<tr>
<td>4</td>
<td>Leibniz Rechenzentrum Germany</td>
<td>SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR / 2012 IBM</td>
<td>147456</td>
<td>2897.00</td>
<td>3185.05</td>
<td>3422.7</td>
</tr>
<tr>
<td>5</td>
<td>National Supercomputing Center in Tianjin China</td>
<td>Tianhe-1A - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT</td>
<td>185368</td>
<td>2566.00</td>
<td>4701.00</td>
<td>4040.0</td>
</tr>
<tr>
<td>6</td>
<td>DOE/SC/Oak Ridge National Laboratory United States</td>
<td>Jaguar - Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090 / 2009 Cray Inc.</td>
<td>298592</td>
<td>1941.00</td>
<td>2627.61</td>
<td>5142.0</td>
</tr>
<tr>
<td>7</td>
<td>CINECA Italy</td>
<td>Fermi - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM</td>
<td>163840</td>
<td>1725.49</td>
<td>2097.15</td>
<td>821.9</td>
</tr>
<tr>
<td>8</td>
<td>Forschungszentrum Juelich (FZJ) Germany</td>
<td>JuQUEEN - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM</td>
<td>131072</td>
<td>1380.39</td>
<td>1677.72</td>
<td>657.5</td>
</tr>
<tr>
<td>9</td>
<td>CEA/TGCC-GENCI France</td>
<td>Curie thin nodes - Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR / 2012 Bull</td>
<td>77184</td>
<td>1359.00</td>
<td>1667.17</td>
<td>2251.0</td>
</tr>
<tr>
<td>10</td>
<td>National Supercomputing Centre in Shenzhen (NSCS) China</td>
<td>Nebulace - Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning</td>
<td>120640</td>
<td>1271.00</td>
<td>2984.30</td>
<td>2580.0</td>
</tr>
</tbody>
</table>

www.top500.org
LRZ Supercomputers

next to come (2014): SuperMUC Phase II 6.4 PFlop/s
SuperMUC and its predecessors
SuperMUC and its predecessors
SuperMUC and its predecessors
LRZ Building Extension

Picture: Horst-Dieter Steinhöfer

Figure: Herzog+Partner für StBAM2 (staatl. Hochbauamt München 2)

Picture: Ernst A. Graf
Increasing numbers

<table>
<thead>
<tr>
<th>Date</th>
<th>System</th>
<th>Flop/s</th>
<th>Cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>HLRB-I</td>
<td>2 Tflop/s</td>
<td>1512</td>
</tr>
<tr>
<td>2006</td>
<td>HLRB-II</td>
<td>62 Tflop/s</td>
<td>9728</td>
</tr>
<tr>
<td>2012</td>
<td>SuperMUC</td>
<td>3200 Tflop/s</td>
<td>155656</td>
</tr>
<tr>
<td>2014</td>
<td>SuperMUC Phase II</td>
<td>3.2 + 3.2 Pflop/s</td>
<td>229960</td>
</tr>
</tbody>
</table>
SuperMUC Architecture

Snapshots/Replika
1.5 PB
(separate fire section)

$HOME
1.5 PB / 10 GB/s

18 Thin node islands
(each >8000 cores)

1 Fat node island
(8200 cores) ➔ SuperMIG

10 PB
200 GB/s

GPFS for
$WORK
$SCRATCH

D. Kranzlmüller
Extreme Scaling on SuperMUC
LRZ Application Mix

- Computational Fluid Dynamics: Optimisation of turbines and wings, noise reduction, air conditioning in trains
- Fusion: Plasma in a future fusion reactor (ITER)
- Astrophysics: Origin and evolution of stars and galaxies
- Solid State Physics: Superconductivity, surface properties
- Geophysics: Earthquake scenarios
- Material Science: Semiconductors
- Chemistry: Catalytic reactions
- Medicine and Medical Engineering: Blood flow, aneurysms, air conditioning of operating theatres
- Biophysics: Properties of viruses, genome analysis
- Climate research: Currents in oceans
First SuperMUC Extreme Scale Workshop

Participants:
- 15 international projects

Prerequisites:
- Successful run on 4 islands (32768 cores)

Participating Groups (Software packages):
- LAMMPS, VERTEX, GADGET, WaLBerla, BQCD, Gromacs, APES, SeisSol, CIAO

Successful results (> 64000 Cores):
- Invited to participate in PARCO Conference (Sept. 2013) including a publication of their approach
**LRZ Extreme Scale Workshop**

- **Regular SuperMUC operation**
  - 4 Islands maximum
  - Batch scheduling system

- **Entire SuperMUC reserved 2,5 days for challenge:**
  - 0,5 Days for testing
  - 2 Days for executing
  - 16 (of 19) Islands available

- **Consumed computing time for all groups:**
  - 1 hour of runtime = 130,000 CPU hours
  - 1 year in total
### Results (Sustained TFlop/s on 128000 cores)

<table>
<thead>
<tr>
<th>Name</th>
<th>MPI</th>
<th># cores</th>
<th>Description</th>
<th>TFlop/s/island</th>
<th>TFlop/s max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linpack</td>
<td>IBM</td>
<td>128000</td>
<td>TOP500</td>
<td>161</td>
<td>2560</td>
</tr>
<tr>
<td>Vertex</td>
<td>IBM</td>
<td>128000</td>
<td>Plasma Physics</td>
<td>15</td>
<td>245</td>
</tr>
<tr>
<td>GROMACS</td>
<td>IBM, Intel</td>
<td>64000</td>
<td>Molecular Modelling</td>
<td>40</td>
<td>110</td>
</tr>
<tr>
<td>Seissol</td>
<td>IBM</td>
<td>64000</td>
<td>Geophysics</td>
<td>31</td>
<td>95</td>
</tr>
<tr>
<td>waLBerla</td>
<td>IBM</td>
<td>128000</td>
<td>Lattice Boltzmann</td>
<td>5.6</td>
<td>90</td>
</tr>
<tr>
<td>LAMMPS</td>
<td>IBM</td>
<td>128000</td>
<td>Molecular Modelling</td>
<td>5.6</td>
<td>90</td>
</tr>
<tr>
<td>APES</td>
<td>IBM</td>
<td>64000</td>
<td>CFD</td>
<td>6</td>
<td>47</td>
</tr>
<tr>
<td>BQCD</td>
<td>Intel</td>
<td>128000</td>
<td>Quantum Physics</td>
<td>10</td>
<td>27</td>
</tr>
</tbody>
</table>
5 Software packages were running on max 16 islands:
- LAMMPS
- VERTEX
- GADGET
- WaLBerla
- BQCD

VERTEX reached 245 TFlop/s on 16 islands (A. Marek)
Results

- 4 Software packages were running on max 8 islands:
  - Gromacs
  - APES
  - SeisSol
  - CIAO

- GROMACS reached 201 TFlop/s on 8 islands (C. Kutzner)
Lessons learned

- Hybrid (MPI+OpenMP) on SuperMUC still slower than pure MPI (e.g. GROMACS), but applications scale to larger core counts (e.g. VERTEX)
- Core pinning needs a lot of experience by the programmer
- Parallel IO still remains a challenge for many applications, both with regard to stability and speed.
- Several stability issues with GPFS were observed for very large jobs due to writing thousands of files in a single directory. This will be improved in the upcoming versions of the application codes.
LRZ Extreme Scale Suite (LESS)

- Platform and architecture agnostic framework for automatic compilation and submission of validation and test jobs
- Framework source originates from DEISA benchmark suite
- Extended in EU Project Scalalife to Gromacs, Dalton and Discrete
- Implemented in XML and perl
- System architecture is described as XML file
- Many system architectures available
  - Hardware: SGI Altix, UV, ice, IBM dataplex, cell, CRAY, generic x86
  - Compilers: icc, gcc, xlc
  - Batch systems: PBS, SLURM, Loadleveller, ...
- Software Packages: BQCD, GROAMCS, Lammps, Gadget, APES,CIAO, SeisSol, GPI, pbdMPI, doRedis, Blender
Next Steps

- LRZ Extreme Scale Benchmark Suite (LESS) will be available in two versions: public and internal
- All teams will have the opportunity to run performance benchmarks after upcoming SuperMUC maintenances
- Next workshop will be June/July 2014
- Initiation of the LRZ Partnership Initiative piCS
Astrophysics: world's largest simulations of supersonic, compressible turbulence with a numerical grid resolution of $4096^3$ points.

Slices through the three-dimensional gas density (top panels) and vorticity (bottom panels) for fully developed, highly compressible, supersonic turbulence, generated by solenoidal driving (left-hand column) and compressive driving (right-hand column), and a grid resolution of $4096^3$ cells.

Federrath C MNRAS 2013;mnras.stt1644
SeisSol – Earthquake Simulation at Petascale

Key Features

- ADER-DG: high approximation order in space and time
- Adaptive tetrahedral meshes for highly complex geometries
- Dynamic rupture simulation coupled to seismic wave propagation

Code Generation for Matrix Kernels:

- Optimal code generation in an offline pre-compile phase
- Generation of vector instruction when the auto-vectorizer fails
- Selection of the optimal kernel (sparse or dense) for every matrix and numerical order

Check-out detailed poster @SC’13!

Several sparsity patterns of a 5th-order discretization
D. Kranzlmüller

Extreme Scaling on SuperMUC
~1 PF Sustained Performance on SuperMUC

Strong Scaling Benchmark Run:
- 20,000 elements per core (recursively generated mesh on cube domain)
- 6th order (1.5 trillion unknowns) using 30% of SuperMUC's memory
- 0.98 PF, more than 30% of peak performance

Weak Scaling Study:
- SCEC LOH.1, 7,252,482 cells
- 6th order (3.6 billion unknowns)
- 2.25 TF on 256 cores (40.6% peak)
- 153 TF on 32K cores (21.6% peak)
- 4.5x speedup in time to solution due to kernel generation

Check-out detailed poster @SC’13!
SuperMUC @ LRZ
Power Consumption at LRZ
“Global carbon-dioxide (CO2) emissions from fossil-fuel combustion reached a record high of 31.6 gigatonnes (Gt) in 2011”, an increase of 1.0 Gt or 3.2% on 2010, the previous record year.”

International Energy Agency
Financial Consequences (Power Bill)

€ / KWh for industrial customers

Source: energy.eu (effective April 2011)
Data Centres are “Heaters with integrated logic”

Torsten Bloth, IBM Lab Services - © IBM Corporation
Cooling SuperMUC
IBM System x iDataPlex
Direct Water Cooled Rack
Cooling Concept – Dedicated Free Cooling

Loop 1

Fot Nodes Storage

Thin Nodes

Computerroom air cooling max. 600 kW <10% of new capacity (Loop 1)

D. Kranzlmüller

Extreme Scaling on SuperMUC
Cooling Infrastructure

Photos: StBAM2 (stata. Hochbauamt München 2)
Cooling Infrastructure (Roof)
SuperMUC HPL Energy Consumption

Energy efficiency
(single number in GFlops/Watt)

9,380E-01 (PDU, 10 minutes resolution, whole run)
9,359E-01 (PDU, 1 minutes resolution, whole run, without cooling)
9,359E-01 (PDU, 1 minutes resolution, whole run, cooling included)
8,871E-01 (machine room measurement, whole run)
7,296E-01 (infrastructure measurement, whole run)
Extreme Scaling on SuperMUC

Dieter Kranzlmüller
kranzlmueller@lrz.de