Extreme Scale Computing at LRZ

Dieter Kranzlmüller

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

- Institute Building
- Visualization Centre
- Lecture Halls
- Cuboid containing computing systems, 72 x 36 x 36 meters

Networks Grid computing
Cloud Computing High Performance Computing
Virtualization IT Security
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

Virtual Reality & Visualization Centre (LRZ)
Examples from the V2C

Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities

- 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)

National HPC-Strategy

- Europäische Höchstleistungsrechenzentren (Tier 0)
- Nationale Höchstleistungsrechenzentren (Tier 1)
- Thematische HPC-Zentren, Zentren mit regionalen Aufgaben (Tier 2)
- HPC-Server (Tier 3)

Gauss Centre for Supercomputing (GCS) (Garching, Stuttgart, Jülich)

Aachen, Berlin, DKRZ, Dresden, DWD, Karlsruhe, Hannover, MPG/RZG, udgl.

Hochschule/Institut
PRACE Research Infrastructure Created

- Establishment of the legal framework
  - PRACE AISBL created with seat in Brussels in April (Association Internationale Sans But Lucratif)
  - 20 members representing 20 European countries
  - Inauguration in Barcelona on June 9

- Funding secured for 2010 - 2015
  - 400 Million € from France, Germany, Italy, Spain
    Provided as Tier-0 services on TCO basis
  - Funding decision for 100 Million € in The Netherlands expected soon
  - 70+ Million € from EC FP7 for preparatory and implementation
    Grants INFSO-RI-211528 and 261557
    Complemented by ~ 60 Million € from PRACE members

PRACE Tier-0 Systems

- **Curie @ GENCI**: Bull Cluster, 1.7 PFlop/s
- **FERMI @ CINECA**: IBM BG/Q, 2.1 PFlop/s
- **Hermit @ HLRS**: Cray XE6, 1 PFlop/s
- **JUQUEEN @ FZJ**: IBM Blue Gene/Q, 5.9 PFlop/s
- **MareNostrum @ BSC**: IBM System X iDataPlex, 1 PFlop/s
- **SuperMUC @ LRZ**: IBM System X iDataPlex, 3.2 PFlop/s
PRACE Tier-0 Access

- Single pan-European Peer Review
- [http://www.prace-project.eu/Call-Announcements?lang=en](http://www.prace-project.eu/Call-Announcements?lang=en)
- Early Access Call in May 2010
  - 68 proposals asked for 1870 Million Core hours
  - 10 projects granted with 328 Million Core hours
  - Principal Investigators from D (5), UK (2) NL (1), I (1), PT (1)
  - Involves researchers from 31 institutions in 12 countries
- Further calls being scheduled (every 6 months)
  - Call open February > Access September same year
  - Call open September > Access March next year
- Example from 8th Regular Call closed on 15 October 2013
  - Spatially adaptive radiation-hydrodynamical simulations of reionization
  - Project leader: Dr Andreas Pawlik, Max Planck Society, GERMANY
  - Research field: Universe Sciences
  - Resource Awarded: 33,800,000 core hours on SuperMUC, Germany;

Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

- 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/OIAsSiguWrQ

Top 500 Supercomputer List (June 2012)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer/Year Vendor</th>
<th>Cores</th>
<th>Pmax</th>
<th>Ppeak</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/NSF/UNIV. United States</td>
<td>Sequoia - SunBlueGenG, Power BQC 16C 3.6 GHz, Oak Ridge / 2011</td>
<td>15723164</td>
<td>3232.25</td>
<td>2013.09</td>
<td>3840.0</td>
</tr>
<tr>
<td>2</td>
<td>DoE/Argonne National Laboratory United States</td>
<td>K Computer: SPARKS With 2.80GHz, Tsubame Supercomputer / 2011</td>
<td>2755624</td>
<td>1215.00</td>
<td>1198.00</td>
<td>12568.0</td>
</tr>
<tr>
<td>3</td>
<td>DOE/Argonne National Laboratory United States</td>
<td>Mira - BlueGene/Q, Power BQC 16C 2.7GHz, Oak Ridge / 2011</td>
<td>7865220</td>
<td>1162.35</td>
<td>1006.67</td>
<td>3485.0</td>
</tr>
<tr>
<td>4</td>
<td>Lohas Forschcentrum Germany</td>
<td>SuperMUC - BlueGene/PS3, 1637069, E.ON / 2012</td>
<td>147456</td>
<td>2087.00</td>
<td>3085.05</td>
<td>3423.7</td>
</tr>
<tr>
<td>5</td>
<td>National Supercomputing Center in Beijing, China</td>
<td>Tianhe-1A - MLX, 1 HPC, Xeon E5-2670 2.7GHz, 1000.00</td>
<td>180996</td>
<td>2586.00</td>
<td>2761.00</td>
<td>4045.0</td>
</tr>
<tr>
<td>6</td>
<td>DOE/Sage Ridge National Laboratory United States</td>
<td>Jaguar - Cray X3, Opteron 6574 8C 2.0GHz, 2000.00, Oak Ridge / 2004</td>
<td>3995952</td>
<td>3541.00</td>
<td>2567.01</td>
<td>5142.0</td>
</tr>
<tr>
<td>7</td>
<td>CINECA Italy</td>
<td>Pegaso - BlueGene/PS3, Power BQC 16C 2.7GHz, 1000.00, Oak Ridge / 2012</td>
<td>6304090</td>
<td>1723.49</td>
<td>2007.15</td>
<td>821.9</td>
</tr>
<tr>
<td>8</td>
<td>Forschungszentrum Jülich (FZJ) Germany</td>
<td>JUREK - BlueGene/PS3, Power BQC 16C 2.7GHz, Oak Ridge / 2012</td>
<td>1310772</td>
<td>1300.30</td>
<td>1777.72</td>
<td>877.0</td>
</tr>
<tr>
<td>9</td>
<td>CNR/ICT-Green IT Italy</td>
<td>Milena - IBM PowerPC 475, HP Superdome, 1000.00</td>
<td>77188</td>
<td>1389.00</td>
<td>1607.17</td>
<td>2521.0</td>
</tr>
<tr>
<td>10</td>
<td>National Supercomputing Centre in Beijing, China</td>
<td>Nebulae - Cray XC30 8C 2.0GHz, Xeon E5-2670, 1000.00, Oak Ridge / 2010</td>
<td>1239808</td>
<td>1271.00</td>
<td>2944.30</td>
<td>2860.0</td>
</tr>
</tbody>
</table>
LRZ Supercomputers

SuperMUC Phase 1 + 2

Phase 1
3.2 PFLOP/s

SuperMUC

Phase 2
3.2 PFLOP/s

next to come (2014): SuperMUC Phase II
6.4 PFlop/s

10-fold every 3.5 years
Double every 12 Months

D. Kranzlmüller

Energy Efficiency and Extreme Scaling
SuperMUC and its predecessors

- D. Kranzlmüller

Energy Efficiency and Extreme Scaling
SuperMUC and its predecessors

LRZ Building Extension

Picture: Horst-Dieter Steinhofer

Figure: Herzog+Partner für SBBAM2 (städt. Hochbauamt München 2)
### 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

- **$HOME**
  - 1.5 PB / 10 GB/s
  - Snapshots/Replicas
  - 1.5 PB (separate fire section)

- **$WORK**
  - GPFS for SuperMUC

- **$SCRATCH**
  - 10 PB
  - 200 GB/s

- Compute nodes
  - 18 Thin node islands (each >8000 cores)
  - 1 Fat node island (8200 cores) → SuperMIG

- Spine Infiniband switch

- Internet

- Active and Backup − 30 PB

- Disaster Recovery Site

- pruned tree (4:1)

- Storage, etc.

- 10 PB
  - 200 GB/s

- I/O nodes

- Login Support nodes

- Parallel Storage
IBM System x iDataPlex
Direct Water Cooled Rack

iDataPlex DWC Rack
w/ water cooled nodes
(front view)

iDataPlex DWC Rack
w/ water cooled nodes
(rear view of water manifolds)

IBM iDataPlex dx360 M4

Torsten Bloth, IBM Lab Services - © IBM Corporation
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
1st LRZ Extreme Scale Workshop

- July 2013:
  - 1st LRZ 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

---

1st 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)
Lessons learned – Technical Perspective

- 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.

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.

- 2nd LRZ Extreme Scaling Workshop ➔ 2-5 June 2014
  - Full system production runs on 18 islands with sustained Pflop/s (4h SeisSol, 7h Gadget)
  - 4 existing + 6 additional full system applications
  - High I/O bandwidth in user space possible (66 GB/s of 200 GB/s max)
  - Important goal: minimize energy*runtime (3-15 W/core)

- Initiation of the LRZ Partnership Initiative πCS
Partnership Initiative
Computational Sciences πCS

- **Individualized services** for selected scientific groups – flagship role
  - Dedicated point-of-contact
  - Individual support and guidance and targeted training & education
  - Planning dependability for use case specific optimized IT infrastructures
  - Early access to latest IT infrastructure (hard- and software) developments and specification of future requirements
  - Access to IT competence network and expertise at Computer Science and Mathematics departments

- **Partner contribution**
  - Embedding IT experts in user groups
  - Joint research projects (including funding)
  - Scientific partnership – joint publications

- **LRZ benefits**
  - Understanding the (current and future) needs and requirements of the respective scientific domain
  - Developing future services for all user groups

Goals for LRZ:

- Thematic focusing – **Environmental Computing**
- Strengthening science through innovative, high performance IT technologies and modern IT infrastructures and IT services
- Interdisciplinary integration (technical and personnel) of scientists and (international) research groups
- Novel requirements and research results at the interface of scientific computing and computer-based sciences
- Increased prospects for attracting research funding through established IT expertise as contribution to application projects
- Outreach and exploitation
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
Extreme Scale Computing at the Leibniz Supercomputing Centre (LRZ)

Dieter Kranzlmüller
kranzlmueller@lrz.de