The new SuperMUC petascale system and applications

Dieter Kranzlmüller

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

With approx. 250 employees
for more than 100,000 students and
for more than 30,000 employees
including 8,500 scientists

• European Supercomputing Centre
• National Supercomputing Centre
• Regional Computer Centre for all Bavarian Universities
• Computer Centre for all Munich Universities

Photo: Ernst Graf
LRZ as IT Competence Centre:
Operating Cutting-edge IT Infrastructure

High Speed Networking:
Munich Scientific Network

High Performance Computing:
SuperMUC, LinuxCluster

Big Data:
Bavarian State Library Digital Archive

Virtual Reality & Visualisation:
V2C (CAVE, Powerwall)
LRZ as IT Competence Centre: Providing Comprehensive IT Services for Science

- **Big Data:** Bavarian State Library Digital Archive
- **High Speed Networking:** Munich Scientific Network
- **High Performance Computing:** SuperMUC, LinuxCluster
- **Virtual Reality & Visualisation:** V2C (CAVE, Powerwall)
- **Service Desk**
- **Application Software Support**
- **Backup & Archive**
- **Authentication & Authorization**
- **IT Security**
- **Personal Consulting**
- **Tailored Solutions**
- **Trainings & Workshops**
- **Application Software Support**
- **Backup & Archive**
- **Attendance & Authorization**
- **IT Security**

Dieter Kranzlmüller

HCMUT, Vietnam
Video: SuperMUC rendered on SuperMUC by LRZ

http://youtu.be/OiAS6iqqWrQ
<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer/Year Vendor</th>
<th>Cores</th>
<th>R&lt;sub&gt;max&lt;/sub&gt;</th>
<th>R&lt;sub&gt;peak&lt;/sub&gt;</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</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 VIII fx 2.0GHz, Tofu interconnect / 2011</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</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</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</td>
<td>186368</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 2.20GHz, Cray Gemini Interconnect, NVIDIA 2090 / 2009</td>
<td>298592</td>
<td>1941.00</td>
<td>2627.01</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</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</td>
<td>131072</td>
<td>1380.39</td>
<td>1677.72</td>
<td>657.5</td>
</tr>
<tr>
<td>9</td>
<td>CEA/TC/GCC-GENCI France</td>
<td>Curie thin nodes - Bullx B510, Xeon E5-2680 8C 2.70GHz, Infiniband QDR / 2012</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>Nebulæ - Dawning TC3000 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010</td>
<td>120640</td>
<td>1271.00</td>
<td>2984.30</td>
<td>2580.0</td>
</tr>
</tbody>
</table>

www.top500.org
LRZ Supercomputers – Chronology

1 Peta...
10-fold every 3.5 years
Double every 13 Months

1 Tera...

SuperMUC
IBM series X iDataplex
+ Linux-Cluster

HLRB2: SGI Altix 4700
+ Linux-Cluster

HLRB1: Hitachi SR8000
+ Linux-Cluster

Fujitsu
VPP/52

IBM SP2
Cray T90

KSR

Cray Y-MP8

Cray Y-MP2


SuperMUC Phase II
LRZ Application Mix

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

Dieter Kranzlmüller

HCMUT, Vietnam
Phylogenetic Tree Computation

Sequencing → Alignment → Phylogenetic Tree

Alexandros Stamatakis, H-ITS
# of possible trees for 150 species

- $4226517247809112252219618802377042809718932383449$
- $8822942857479880831434032178759024536798491951168$
- $3076494692867414802738570221298292428457687814873$
- $4552121861861600804474608426626044448936698500560$
- $2468116186441264227425440726676614927906540649360$
- $2976397461917469326750931190889241406694054603576$
- $66015625$

- $\approx 4.22 \times 10^{301}$
1KITE Project  
(1,000 Insect Transcriptome Evolution)

- Alexandros Stamatakis  
  Scientific Computing Group,  
  Heidelberg Institute for Theoretical Studies (HITS) /  
  Exelixis Lab

- „Big Data“ and High Performance Computing
- Novel software and applications needed
- Reading the data: only 1 minute (instead of 15 minutes)
- 1000 Processors: 17 hours (instead of 10 days)
- Load balancing

1KITE Dataflow
Neotropical Rainforests are hyperdiverse ecosystems.
Since Humboldt and Bonpland, we know about the high animal and plant richness.
New study now finds that unicellular eukaryotes are even more diverse.
Particularly the parasitic Apicomplexa dominate these forests.
Their presence might drive the diversity of macro-organisms.

Data Analysis (TU Kaiserslautern, HITS, KIT)

- More than 130 million DNA sequences were analysed
- Most of them belong to yet unknown microbial species
- Thus, a thorough method was necessary for classifying those sequences
- The method takes the evolutionary history of known species into account
- But this comes at the cost of increased computational needs
- Approximately 1 million computation hours on SuperMUC were necessary


SuperMUC and its predecessors
LRZ Building Extension

Picture: Horst-Dieter Steinhöfer

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

Picture: Ernst A. Graf
SuperMUC Architecture

Snapshots/Replika
1.5 PB
(separate fire section)

$HOME
1.5 PB / 10 GB/s

NAS
80 Gbit/s

Internet

Achive and Backup ~

Desaster Recovery Site

18 Thin node islands
(each >8000 cores)

1 Fat node island
(8200 cores) \(\rightarrow\) SuperMIG

SB-EP
16 cores/node
2 GB/core

Compute nodes

non blocking

WM-EX
40 cores/node
6.4 GB/core

Compute nodes

non blocking

GPFS for
$WORK

$SCRATCH

10 PB
200 GB/s

Parallel Storage

I/O
nodes

Pruned tree
(4:1)

Storage, etc.
Infiniband switches

Dieter Kranzlmüller

HCMUT, Vietnam
Power Consumption at LRZ

Stromverbrauch in MWh

Cooling SuperMUC
Usage of Intel Xeon E5 2697v3 processors

- Direct liquid cooling
  - 10% power advantage over air cooled system
  - 25% power advantage due to chiller-less cooling

Energy-aware scheduling
- 6% power advantage
- ~40% power advantage
- Annual savings: ~2 Mio. € for SuperMUC Phase 1 and 2

Photos: Torsten Bloth, Lenovo
## Increasing Numbers of Cores

<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>2015</td>
<td>SuperMUC Phase II</td>
<td>3.2 + 3.2 Pflop/s</td>
<td>229960</td>
</tr>
</tbody>
</table>
### Results:

<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>31</td>
<td>95</td>
</tr>
<tr>
<td>Seissol</td>
<td>IBM</td>
<td>64000</td>
<td>Geophysics</td>
<td>40</td>
<td>110</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>

- Sustained TFlop/s on 64000/128000 cores
Dr. Christian Pelties, Department of Earth and Environmental Sciences (LMU)
Prof. Michael Bader, Department of Informatics (TUM)

1,42 Petaflop/s on 147,456 Cores of SuperMUC
(44,5 % of Peak Performance)

http://www.uni-muenchen.de/informationen_fuer/presse/presseinformationen/2014/pelties_seisol.html

Picture: Alex Breuer (TUM) / Christian Pelties (LMU)
LRZ benefits
- Understanding the (current and future) needs and requirements of the respective scientific domain
- Developing future services for all user groups
- Thematic focusing: Environmental Computing

EU Project Series DRIHM*
- Flash Project estimates for 1990-2006
- > 29 billion euros in damages produced by floods
- > 4,500 total number of casualties
Professor Peter V. Coveney

- holds a chair in Physical Chemistry
- is an Honorary Professor in Computer Science at University College London (UCL)
- is Professor Adjunct at Yale University School of Medicine (USA).

- is Director of the Centre for Computational Science (CCS) and of the Computational Life and Medical Sciences Network (CLMS) at UCL.

- https://www.ucl.ac.uk/chemistry/people/peter-coveney

- leads CompBioMed, A Centre of Excellence in Computational Biomedicine

- http://www.compbio.med.eu
Goal: **advance the role of computationally based modelling and simulation within biomedicine.**

Three related user communities:
- academic,
- industrial and
- clinical researchers

All wish to build, develop and extend such capabilities in line with the increasing power of high performance computers.

Three distinct exemplar research areas:
- cardiovascular,
- molecularly-based and
- neuro-musculoskeletal medicine.
How to Utilize SuperMUC as a Tool?

- Target question:
  
  Can we use the genomic data from an individual candidate and predict whether a standard drug for the treatment of breast cancer will help or not?

- Goal:

  A demonstration of feasibility with the power of high performance computing

- Key questions:
  
  - Provide an answer to the question above
  - Determine how to use IT-Infrastructures for this question
  - Detect insufficiencies of using IT-Infrastructures for this question
  - Derive a workflow for utilizing HPC in daily operation
Molecular Simulation

- Running on all cores of SuperMUC Phase1+2
- Docking simulation of potentials drugs for breast cancer
- 37 hours total run time
- 241,672 cores
- 8.900.000 CPU hours
- 5 Terabytes of data produced

EU CoE CompBioMed
http://www.compbioimed.eu
EU Projects COMPAT and MAPPER
http://www.compat-project.eu
What’s next? A View into the Future

Until today:
- HLRB-II (pre-SuperMUC): Top 500 06/2007: 56.5 Tflop/s
- SuperMUC Phase 1: Top 500 06/2012: 2897 Tflop/s

Coming up:
- SuperMUC NG (Next Generation) – Procurement on-going
## Consulting the Top 500 List - www.top500.org

### Projected Performance Development

![Projected Performance Graph](Image)

### Accelerator/CP Family

<table>
<thead>
<tr>
<th>Accelerator/CP Family</th>
<th>Count</th>
<th>System Share (%)</th>
<th>Rmax (GFlops)</th>
<th>Rpeak (GFlops)</th>
<th>Cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nvidia Kepler</td>
<td>50</td>
<td>10</td>
<td>59,004,619</td>
<td>92,655,119</td>
<td>1,668,690</td>
</tr>
<tr>
<td>Intel Xeon Phi</td>
<td>21</td>
<td>4.2</td>
<td>55,066,905</td>
<td>86,361,180</td>
<td>4,756,732</td>
</tr>
<tr>
<td>Nvidia Fermi</td>
<td>8</td>
<td>1.6</td>
<td>7,309,880</td>
<td>14,735,848</td>
<td>572,740</td>
</tr>
<tr>
<td>Hybrid</td>
<td>3</td>
<td>0.6</td>
<td>4,621,240</td>
<td>7,933,520</td>
<td>415,960</td>
</tr>
<tr>
<td>Nvidia Pascal</td>
<td>2</td>
<td>0.4</td>
<td>13,086,000</td>
<td>20,884,480</td>
<td>267,232</td>
</tr>
<tr>
<td>ATI Radeon</td>
<td>1</td>
<td>0.2</td>
<td>532,600</td>
<td>1,098,000</td>
<td>38,400</td>
</tr>
<tr>
<td>PEZY-SC</td>
<td>1</td>
<td>0.2</td>
<td>1,001,010</td>
<td>1,533,460</td>
<td>1,313,280</td>
</tr>
</tbody>
</table>

### Rankings

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>National Supercomputing Center in Wuxi, China</td>
</tr>
<tr>
<td>2</td>
<td>National Super Computer Center in Guangzhou, China</td>
</tr>
<tr>
<td>3</td>
<td>DOE/SC/Laborate United States</td>
</tr>
<tr>
<td>4</td>
<td>DOE/NN United States</td>
</tr>
<tr>
<td>5</td>
<td>DOE/SC/Laborate United States</td>
</tr>
<tr>
<td>6</td>
<td>Joint Cer Perform Japan</td>
</tr>
<tr>
<td>7</td>
<td>RIKEN A Compute Japan</td>
</tr>
<tr>
<td>8</td>
<td>Swiss National supercomputing Centre (SCS), Switzerland</td>
</tr>
<tr>
<td>9</td>
<td>DOE/SC/Argonne National Laboratory United States</td>
</tr>
<tr>
<td>10</td>
<td>DOE/NSA/LANL/SNL United States</td>
</tr>
</tbody>
</table>

---

**Notes:**
- **Mizuho** - Cray XC40, Xede, 2698v3 16C 2.3GHz, Aris interconnect, NVIDIA Ti Cray Inc.
- **Mira** - BlueGene/Q, Powery 16C 1.6GHz, Custom IBM
- **Trinity** - Cray XC40, Xede, 2698v3 16C 2.3GHz, Aris interconnect Cray Inc.
What’s next? A View into the Future

Until today:
- HLRB-II (pre-SuperMUC): Top 500 06/2007: 56.5 Tflop/s
- SuperMUC Phase 1: Top 500 06/2012: 2897 Tflop/s

Coming up:
- SuperMUC NG (Next Generation) – Procurement on-going

Selection criteria:
- LRZ application mix (compute, memory, bandwidth characteristics)
  - Number of cores
  - Memory per core
  - Interconnect
- Accelerators (Manycore, GPGPU, ...)
- Virtualization (Docker, Cloud, ...)
- Workflow engines, HTC applications, ...
- Power consumption (in total, over time, ...)
Slide courtesy Herbert Huber
Conclusions

- Excellent research needs excellent tools
- Supercomputers provide the highest possible computational performance, interconnectivity and memory capacity

- The complexity of (super-)computers (such as SuperMUC NG) is steadily increasing (not only on the extreme scale)
- Demand of domain science drives computer science research to new frontiers

- Users need the possibility to execute (and optimize) their codes on the full size machines
- The LRZ Partnership Initiative Computational Science (piCS) tries to improve user support

1. **Choose focus topics to serve as lighthouse**
   - National agreement within GCS: LRZ focuses on Environment (& Energy)

2. **Choose user communities**
   - Already active at LRZ?
   - Not active at LRZ?

3. **Invite them for introductory piCS Workshops**
   - Show faces & tour
   - Discussion on joint topics, requirements, interests, ...

4. **Establish links between communities and specific points-of-contact**
   - Whom to talk to, if there are questions?
   - When to talk to them? In general, as early as possible
   - Maybe, place people into the research groups (weekly, for a certain period)

5. **Run joint lectures (e.g. hydrometeorology and computer science)**

6. **Apply for joint projects**

7. **Use **HPC Machines** efficiently to do science**
The new SuperMUC petascale system and applications

Contributions from: A. Bode, A. Stamatakis, L. Czech, A. Frank, M. Brehm, H. Huber, M. Bader, F. Jamitzky, A. Parodi, ...