LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN — MNM TEAM — lrz

πCS

Partnership Initiative Computational Science – Using HPC efficiently

Dieter Kranzlmüller

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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN — lrz

From Michael Resch's Talk

- Prof. Resch:
  - „Education and Training HPC and the challenges we see in HPC"
  - „It is a question of algorithms, of methods, of how you use the system"
  - „For a successful simulation, you need to know computer science, mathematics, and the respective field of application"
  - „Focus at HLRS on engineering"
  - „Training is necessary!"

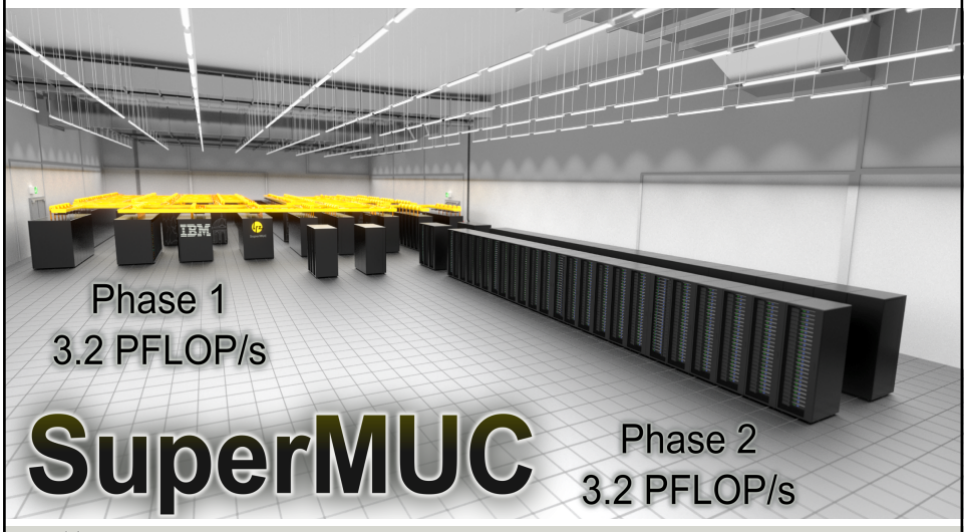
Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

With approx. 230 employees
for more than 100.000 students and
for more than 30.000 employees
including 8.500 scientists

- European Supercomputing Centre
- National Supercomputing Centre ➔ GCS
- Regional Computer Centre for all Bavarian Universities
- Computer Centre for all Munich Universities

Photo: Ernst Graf

SuperMUC Phase 1 + 2



LRZ Building Extension

Picture: Horst-Dieter Steinhöfer

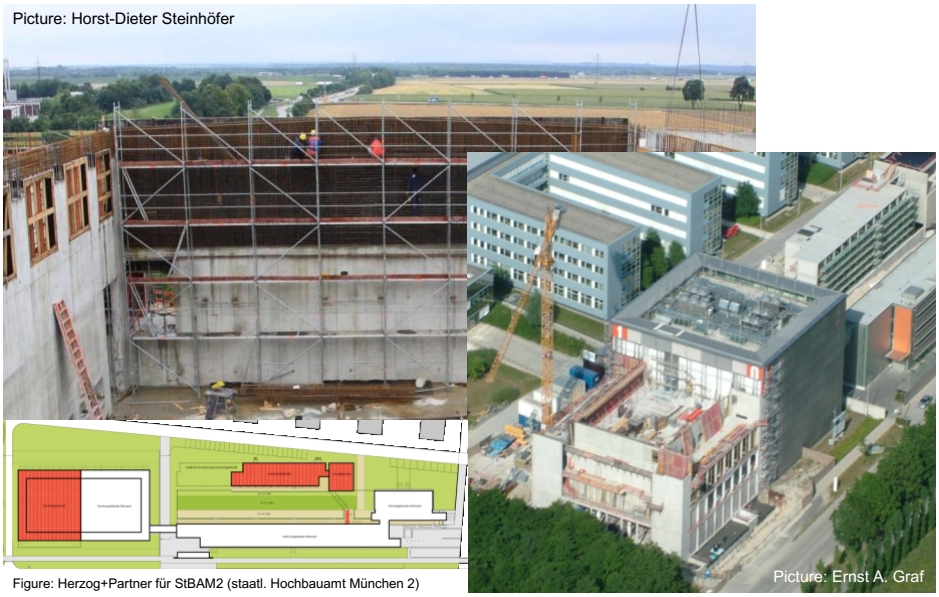
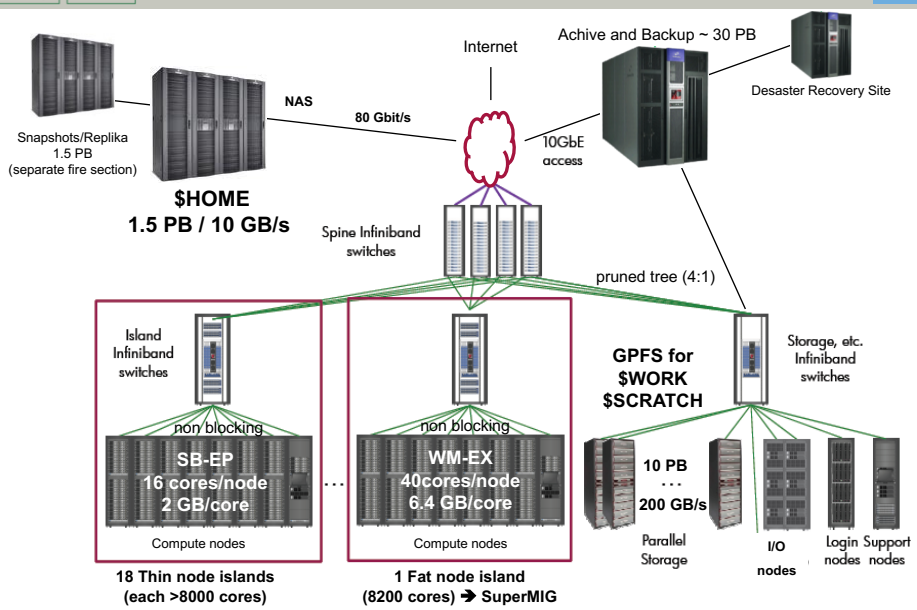


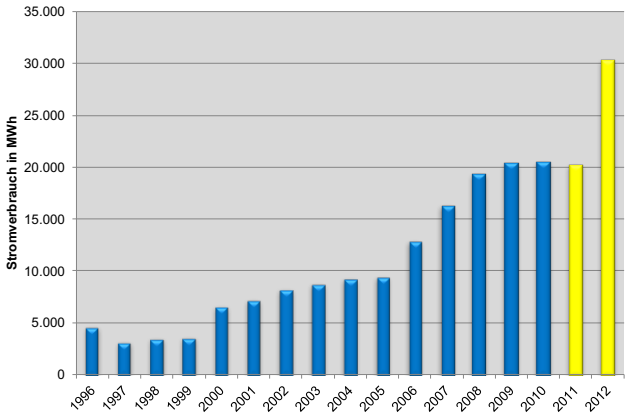
Figure: Herzog+Partner für StBAM2 (staatl. Hochbauamt München 2)

Picture: Ernst A. Graf

SuperMUC Architecture



Power Consumption at LRZ



Cooling SuperMUC

SuperMUC Phase 2 @ LRZ

Photos: Torsten Bloth, Lenovo

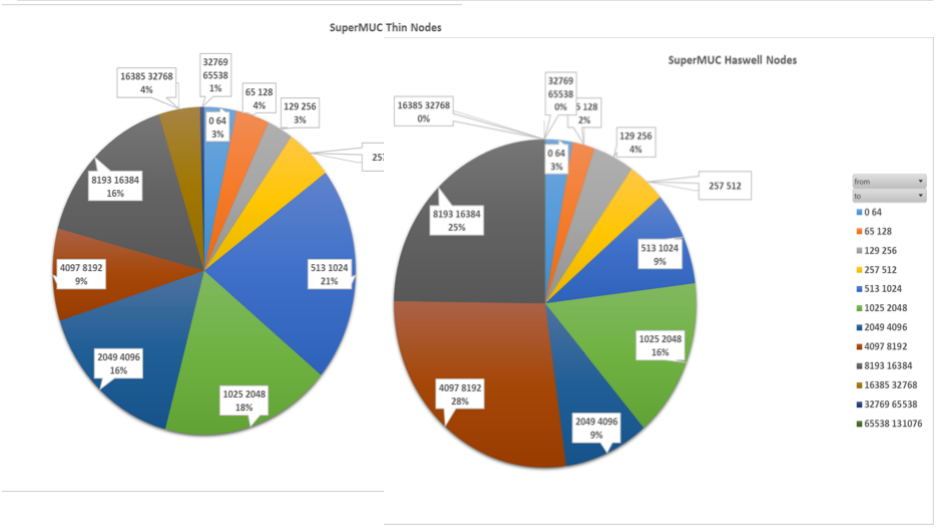
High Energy Efficiency

- ✓ 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
  - Total annual savings of ~2 Mio. € for SuperMUC Phase 1 and 2

Increasing numbers

Date	System	Flop/s	Cores
2000	HLRB-I	2 Tflop/s	1512
2006	HLRB-II	62 Tflop/s	9728
2012	SuperMUC	3200 Tflop/s	155656
2015	SuperMUC Phase II	3.2 + 3.2 Pflop/s	229960

SuperMUC Jobsize 2015 (in Cores)



Challenges on Extreme Scale Systems

- Size: number of cores > 100.000
- Complexity/Heterogeneity
- Reliability/Resilience
- Energy consumption as part of Total Cost of Ownership (TCO)
  - Execute codes with optimal power consumption (or within a certain power band) ➔ Frequency scaling
  - Optimize for energy-to-solution ➔ Allow more codes within given budget
  - Improved performance ➔ (in most cases) improved energy-to-solution

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)

Name	MPI	# cores	Description	TFlop/s/island	TFlop/s max
Linpack	IBM	128000	TOP500	161	2560
Vertex	IBM	128000	Plasma Physics	15	245
GROMACS	IBM, Intel	64000	Molecular Modelling	40	110
Seissol	IBM	64000	Geophysics	31	95
waLBerla	IBM	128000	Lattice Boltzmann	5.6	90
LAMMPS	IBM	128000	Molecular Modelling	5.6	90
APES	IBM	64000	CFD	6	47
BQCD	Intel	128000	Quantum Physics	10	27

Extreme Scaling Continued

- Lessons learned ➔ Stability and scalability
- 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)
- 3rd Extreme Scale-Out with new SuperMUC Phase 2

Extreme Scale-Out SuperMUC Phase 2

- 12 May – 12 June 2015 (30 days)
- Selected Group of Early Users

- Nightly Operation: general queue max 3 islands
- Daytime Operation: special queue max 6 islands (full system)

- Total available: 63,432,000 core hours
- Total used: 43,758,430 core hours (Utilisation: 68.98%)

Lessons learned (2015):

- Preparation is everything
- Finding Heisenbugs is difficult
- MPI is at its limits
- Hybrid (MPI+OpenMP) is the way to go
- I/O libraries getting even more important

4th Extreme Scale Workshop 2016

- 4 Day Workshop (29 February – 3 March 2016)
- 13 Projects:

	Application	Field	Institution	PI
1	INDEXA	CDF	TU München	M. Kronbichler
2	MPAS	Climate Science	KIT	D. Heinzeller
3	Inhouse	Material Science	TU Dresden	F. Ortmann
4	HemeLB	Life Science	UC London	P. Coveney
5	KPM	Chemistry	FAU Erlangen	M. Kreutzer
6	SWIFT	Cosmology	U Durham	M. Schaller
7	LISO	CFD	TU Darmstadt	S. Kraheberger
8	ILDBC	Lattice Boltzmann	FAU Erlangen	M. Wittmann
9	Walberla	Lattice Boltzmann	FAU Erlangen	Ch. Godenschwager
10	GASPI	Framework	ITWM Kaiserslautern	M. Kühn
11	GADGET	Cosmology	LMU München	K. Dolag
12	VERTEX	Astrophysics	MPI for Astrophysics	T. Melson
13	PSC	Plasma	LMU München	K. Bamberg

- 147,456 cores in 9216 Nodes
- 14.1 Mio CPUh
- Max Time per Job 6h
- Daily and nightly operation mode

COMPAT ➔ CompBioMed (Peter Coveney, UCL)

- Molecular Modelling Simulation running on all cores of SuperMUC Phase 1+2
- Docking simulation of potentials drugs for breast cancer
- Goal: A demonstration of feasibility with the power of high performance computing
- 37 hours total run time
- 241,672 cores
- 8.900.000 CPU hours
- Tools developed in EU Projects MAPPER and COMPAT: http://www.compat-project.eu/
- Lessons learned ➔ CompBioMed, a Centre of Excellence in Computational Biomedicine http://www.compbiomed.eu

D. Kranzlmüller SSIP 2017 Workshop 20

Summary from the 4th LRZ Extreme Scale Workshop

	Lesson learned
1	Although Phase 1 is a well running system for some time now (since 2011) still some quirks and problems in the system have been found and have been fixed.
2	The main focus was on the performance optimization of the user codes and lead to great results.
3	One code was using Phase1+Phase2 for the first time. (for 37h all available 241,672 cores)
4	Applications from JUQUEEN and Piz Daint show how the general purpose architecture of SuperMUC compares to specialized architectures like GPUs or BlueGene.
5	Application codes now reach the Pflop/s range.

Extreme Scaling - Conclusions

- The number of compute cores, the complexity (and heterogeneity) is steadily increasing – introducing new challenges/issues
- Users need to possibility to reliably execute (and optimize) their codes on the full size machines with more than 100.000 cores
- The Extreme Scaling Workshop Series @ LRZ offers a number of incentives for users ➔ Next Workshop Spring 2017
- The lessons learned from the Extreme Scaling Workshop are very valuable for the operation of the center
  - Improve performance of applications and energy consumption during operations
  - Improve reliability/stability of hard- and software environment under extreme conditions
  - Learn how to use the infrastructure and prepare processes for operations

Partnership Initiative Computational Sciences πCS

- **Individualized services** for selected scientific groups – **flagship role**
  - Dedicated point-of-contact
  - Individual support and guidance and targeted training &
  - Planning dependability for use case specific optimi… …uctures
  - Early access to latest IT infrastructure (hard- … developments and specification of future requirement…
  - Access to IT competence network … at CS and Math departments
- **Partner contribution**
  - Embedding IT exp… …groups
  - Joint resear… …including funding)
  - Scienti… …ip – equal footing – joint publications
- **LRZ …**
  - Un… …tanding the (current and future) needs and requirements of the respective scientific domain
  - Developing future services for all user groups
  - Thematic focusing: **Environmental Computing**

The LRZ Partnership Initiative Computational Science (piCS)
http://www.sciencedirect.com/science/article/pii/S1877050914003433

piCS Cookbook

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 efficiently


Partnership Initiative Computational Science –
Using HPC efficiently
Dieter Kranzlmüller
kranzlmueller@lrz.de