

The LRZ Extreme Scaling Workshops – How to push the barrier by pushing the users?

Dieter Kranzlmüller

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



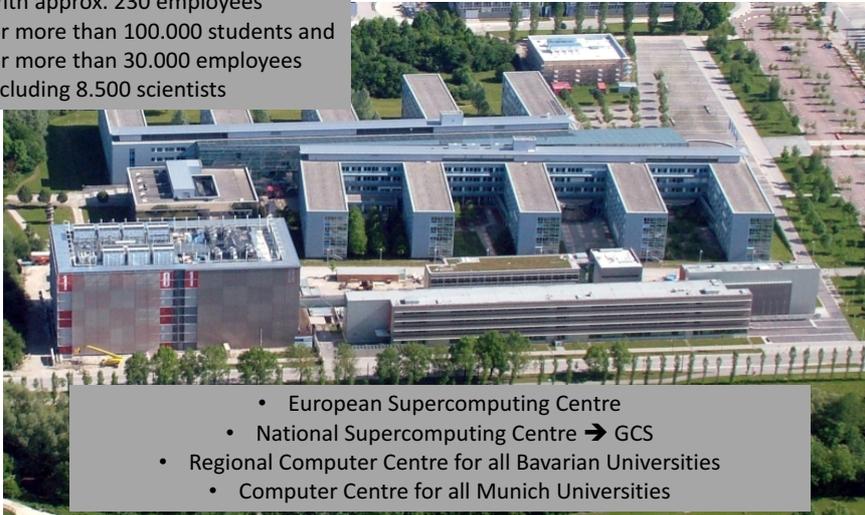
Thanks to **Ferdinand Jamitzky** @ Application Support Group LRZ

High(est) Performance Computing in Germany

- 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)

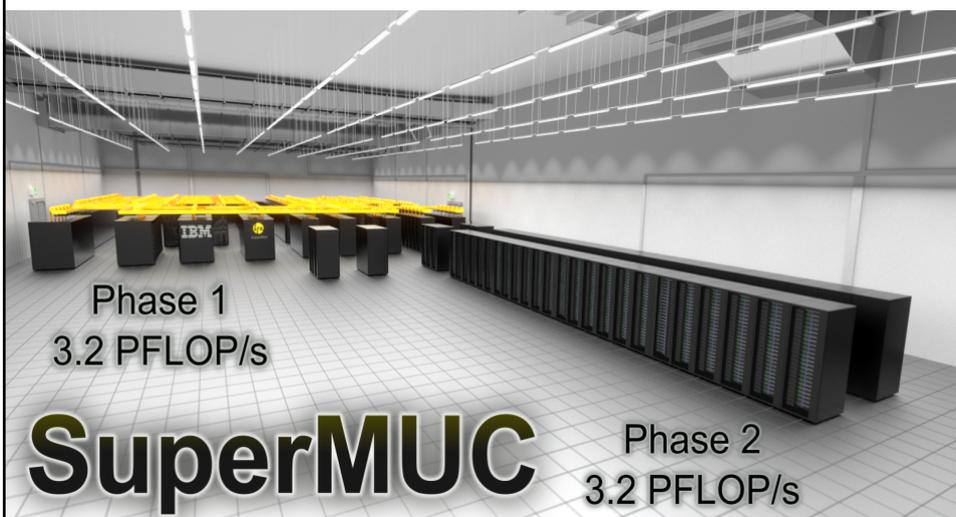


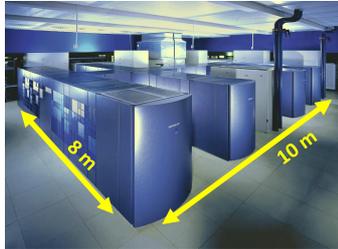
With approx. 230 employees
for more than 100.000 students and
for more than 30.000 employees
including 8.500 scientists



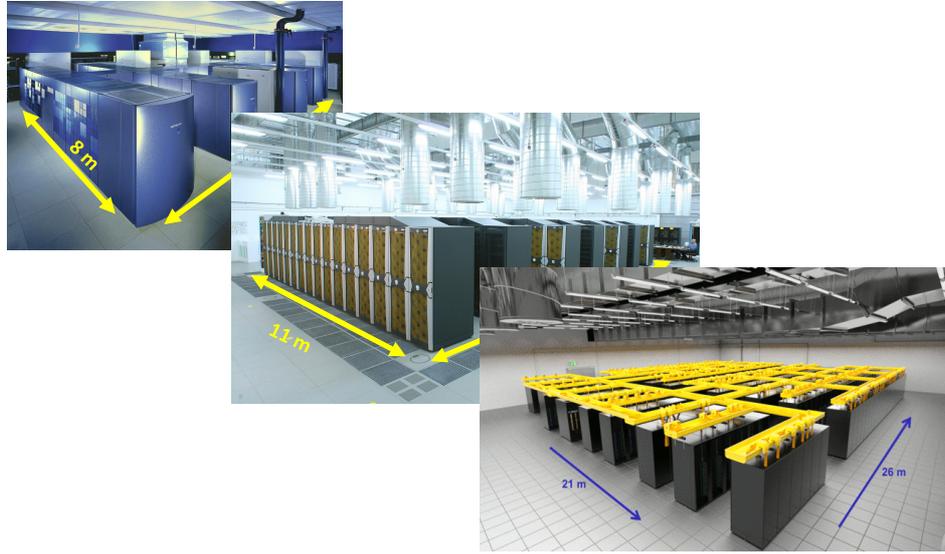
- European Supercomputing Centre
- National Supercomputing Centre → GCS
- Regional Computer Centre for all Bavarian Universities
- Computer Centre for all Munich Universities

Photo: Ernst Graf





SuperMUC and its predecessors



LRZ Building Extension

Picture: Horst-Dieter Steinhöfer

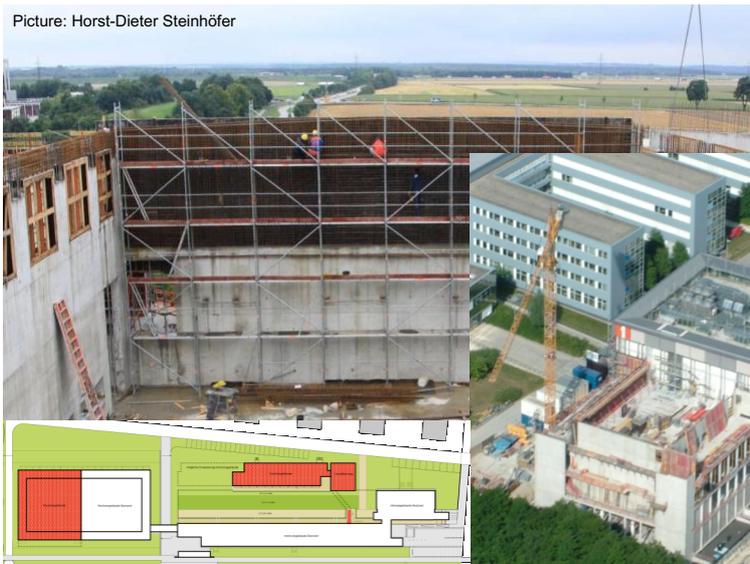
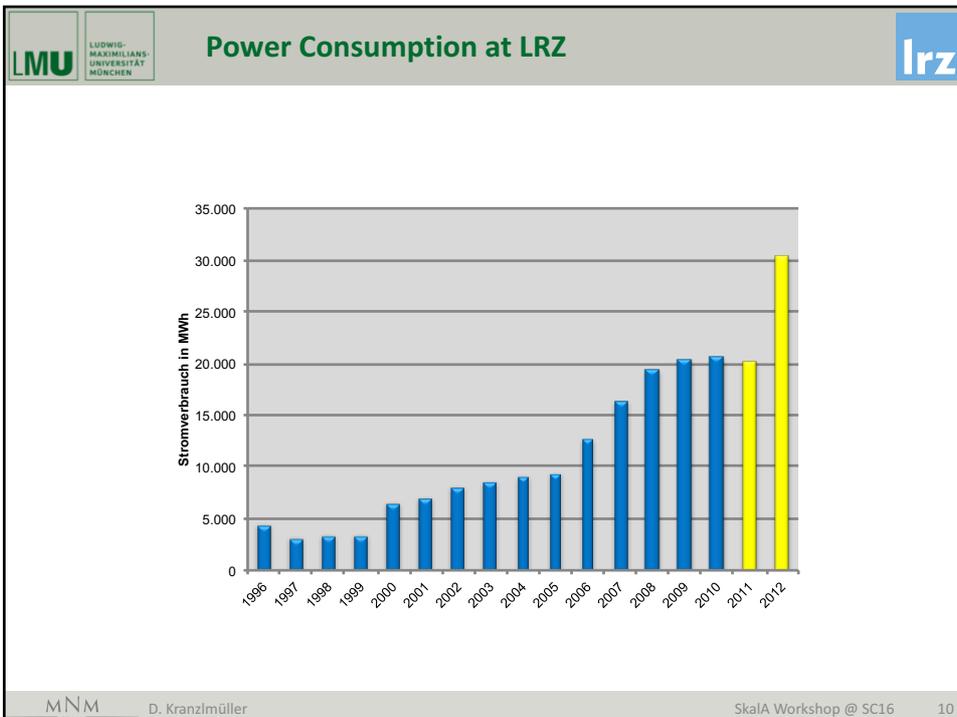
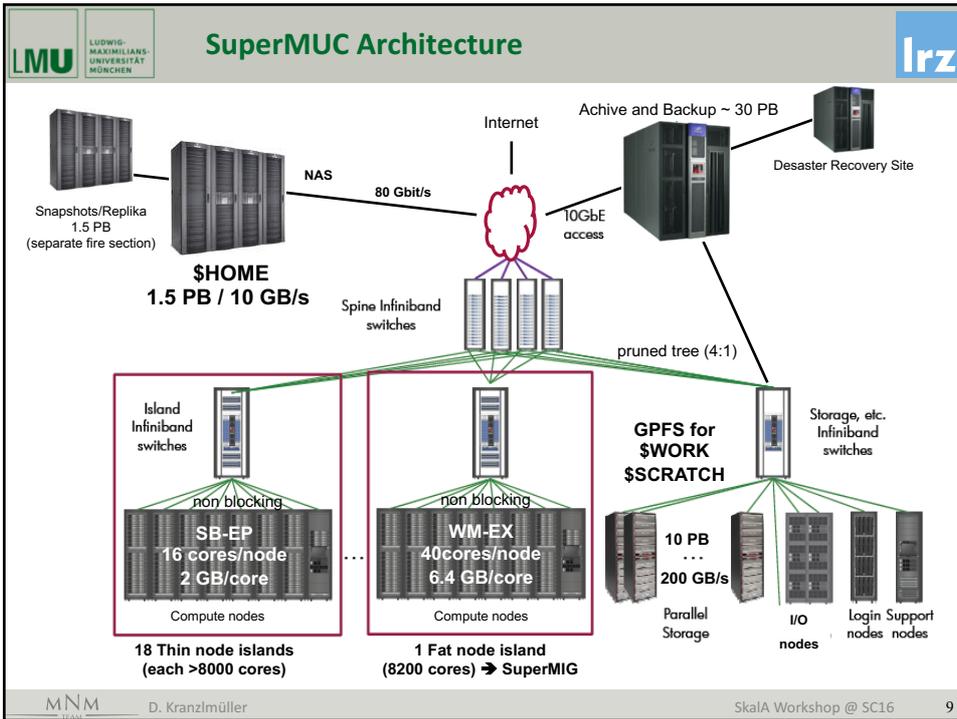


Figure: Herzog+Partner für StBAM2 (staatl. Hochbauamt München 2)

Picture: Ernst A. Graf



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Cooling SuperMUC

lrz

MNM D. Kranzmüller SkalA Workshop @ SC16 11

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

SuperMUC Phase 2 @ LRZ

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

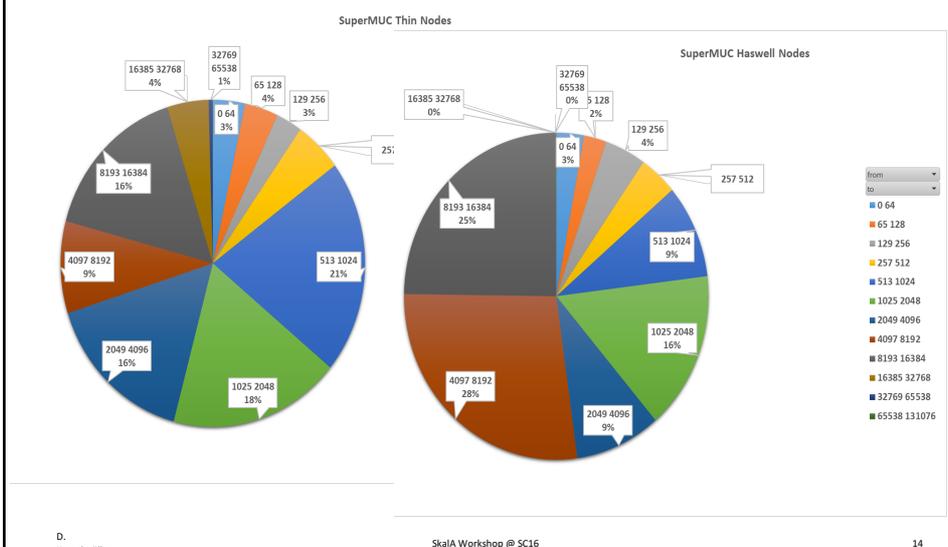
MNM D. Kranzmüller Slide: Herbert Huber SkalA Workshop @ SC16 12

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)



- 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

- 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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN **1st LRZ Extreme Scale Workshop** lrz

- 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

MNM D. Kranzmüller Skala Workshop @ SC16 17

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN **Results (Sustained TFlop/s on 128000 cores)** lrz

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

MNM D. Kranzmüller Skala Workshop @ SC16 18

- 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

- 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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN **4th Extreme Scale Workshop 2016** lrz

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

	Application	Field	Institution	PI
1	INDEXA	CFD	TU München	M. Kronbichler
2	MPAS	Climate Science	KIT	D. Heinzeller
3	Inhouse	Material Science	TU Dresden	F. Ortman
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

MNM

VERTEX: Simulation Code for Supernova Explosions (plasma + neutrino dynamics)

A. Marek and Team (Max Planck-Institute for Astrophysics, Garching)

Finalists:

- INDEXA
- PSC
- waLBerla
- VERTEX

Leibniz Extreme Scaling Award
Extreme Scale Workshop 2016@LRZ

Quelle
Kein St
Für Hi

Motivate your users!

MNM

SandyBridge core (node) performance:

- scale measured performance with run-time on SandyBridge node (320 GFlop/s/node) => 35 GFlop/s/node ~ 11 % of peak performance

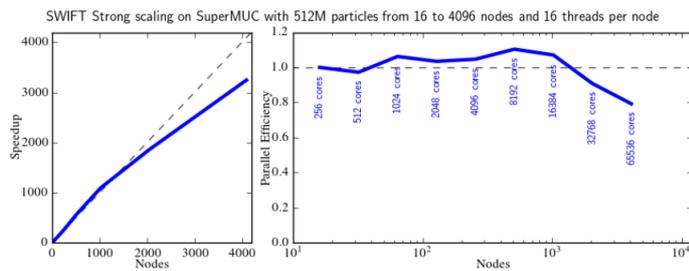
Performance on 147k cores of SuperMuc:

- scale SandyBridge performance with the measured parallel efficiency of weak scaling runs from 1 node to 9216 nodes

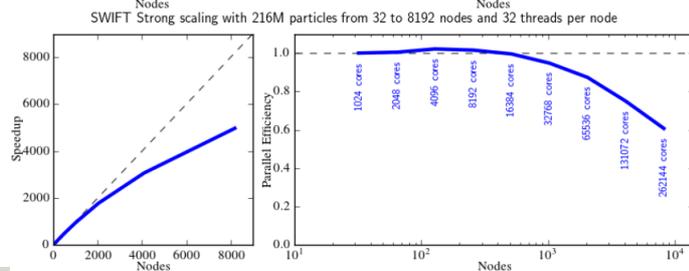
=> ~ 0.4 PFlop/s on LRZ's SuperMUC

~ 9 % of peak performance

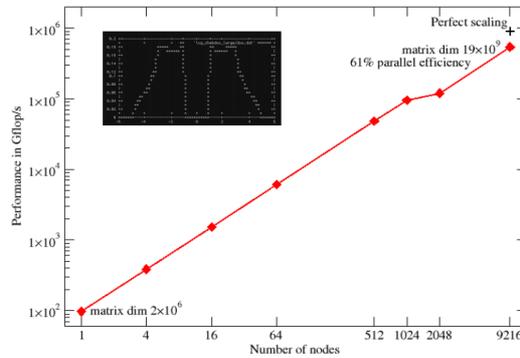
SuperMUC



JUQUEEN



- Kernel Polynomial Method
- Approximate the complete eigenvalue spectrum of a large sparse matrix
- 542 Tflop/s on Phase1
- 19% of LINPACK performance
- Drop from 2->4 islands, not yet understood
- Equal Performance on 9216 SuperMUC nodes and 4096 Piz Daint Nodes
- 50% better on Piz Daint because of GPU



GPI (M. Kühn, Fraunhofer ITWM, Kaiserslautern)

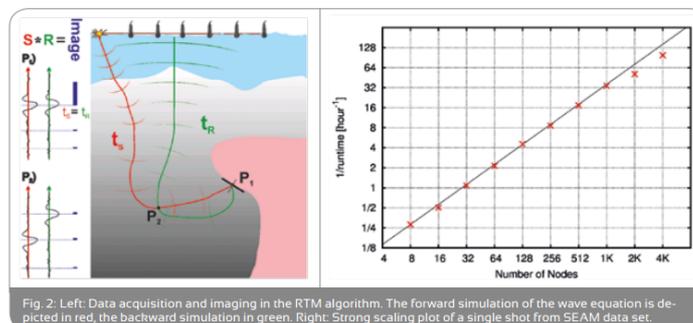


Fig. 2: Left: Data acquisition and imaging in the RTM algorithm. The forward simulation of the wave equation is depicted in red, the backward simulation in green. Right: Strong scaling plot of a single shot from SEAM data set.

Reverse Time Migration with GPI-2 (M. Kühn, Fraunhofer ITWM, Kaiserslautern)

Our benchmark calculates a single shot at 15Hz of the well established synthetic SEAM benchmark [2]. The velocity is modeled as Tilted Transverse Isotropic (TTI), the simulation domain has 800x915x1291 voxels and the wave equation is discretized with an 8th order stencil in space and 2nd order in time.



LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN



The WALBERLA Framework

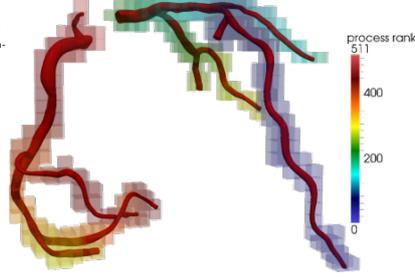
Flows in Complex Geometries with LBM

C. Godenschwager & F. Schornbaum, Chair for System Simulation, FAU Erlangen-Nürnberg

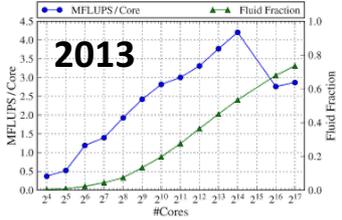
Observations during the workshop

- SuperMUC Phase 1 more stable than in 2013
- Fat node island is invaluable for pre-/postprocessing tasks

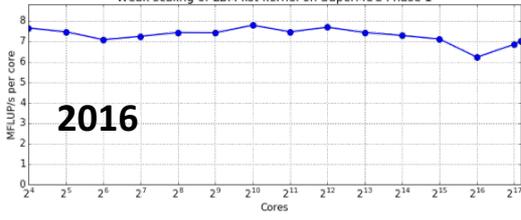
Weak scaling experiments with updated code base



process rank
511
400
200
0



2013



2016





LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN



Material Science, TU Dresden (F. Ortmann)

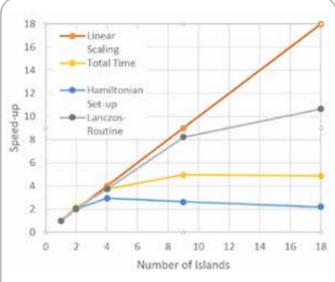


Fig. 1: Strong scaling plot on SuperMUC. Reference point is a single island with 8192 cores. Sample size is 4,558,159,872.

Linear-Scaling Transport Approach for Innovative Electronic Materials (F. Ortmann, TU Dresden)

We demonstrate here that the dominating part of the code which is the Lanczos routine for matrix-vector multiplication scales very well beyond 32,768 cores on SuperMUC. For this intensive part, we measure a speed-up of 8.2 on 73,728 cores (9 islands) compared to a single island (grey line in figure 1). This corresponds to 91% efficiency.



- 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>



CompBioMed

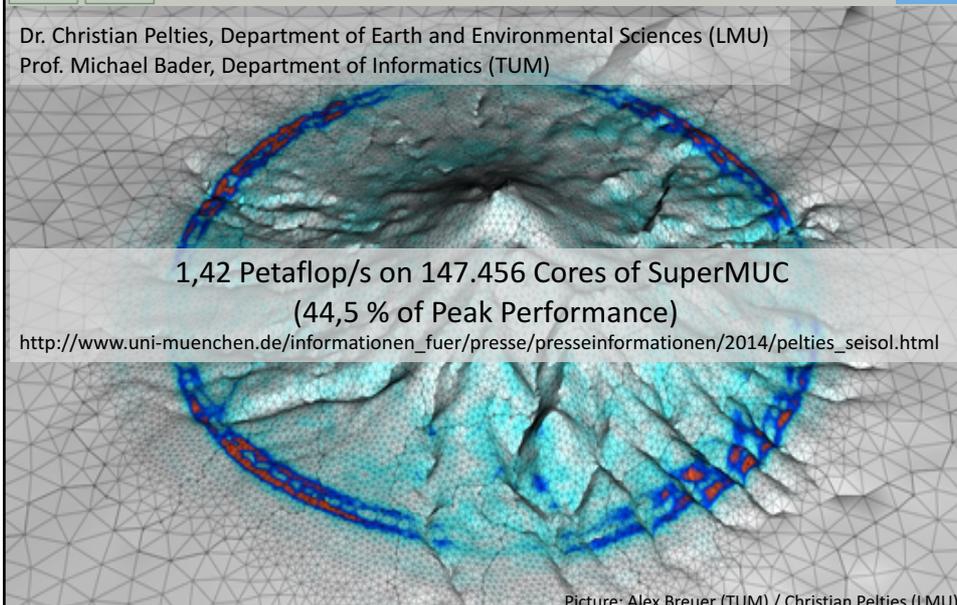
	Lesson learned
1	Although Phase 1 is a well running system for some time now (since 2011) still some quirks and problems in the system have been found and have been fixed.
2	The main focus was on the performance optimization of the user codes and lead to great results.
3	One code was using Phase1+Phase2 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.

- 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

- **Individualized services** for selected scientific groups – **flagship role**
 - Dedicated point-of-contact
 - Individual support and guidance and targeted training & development
 - Planning dependability for use case specific optimization of IT infrastructures
 - Early access to latest IT infrastructure (hard- and software) and developments and specification of future requirements
 - Access to IT competence network and expertise at CS and Math departments
- **Partner contribution**
 - Embedding IT expertise in scientific groups
 - Joint research projects (including funding)
 - Scientific publications – equal footing – joint publications
- **LRZ contribution**
 - Understanding the (current and future) needs and requirements of the respective scientific domain
 - Developing future services for all user groups
 - Thematic focusing: **Environmental Computing**

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN **SeisSol - Numerical Simulation of Seismic Wave Phenomena** **lrz**

Dr. Christian Pelties, Department of Earth and Environmental Sciences (LMU)
 Prof. Michael Bader, Department of Informatics (TUM)



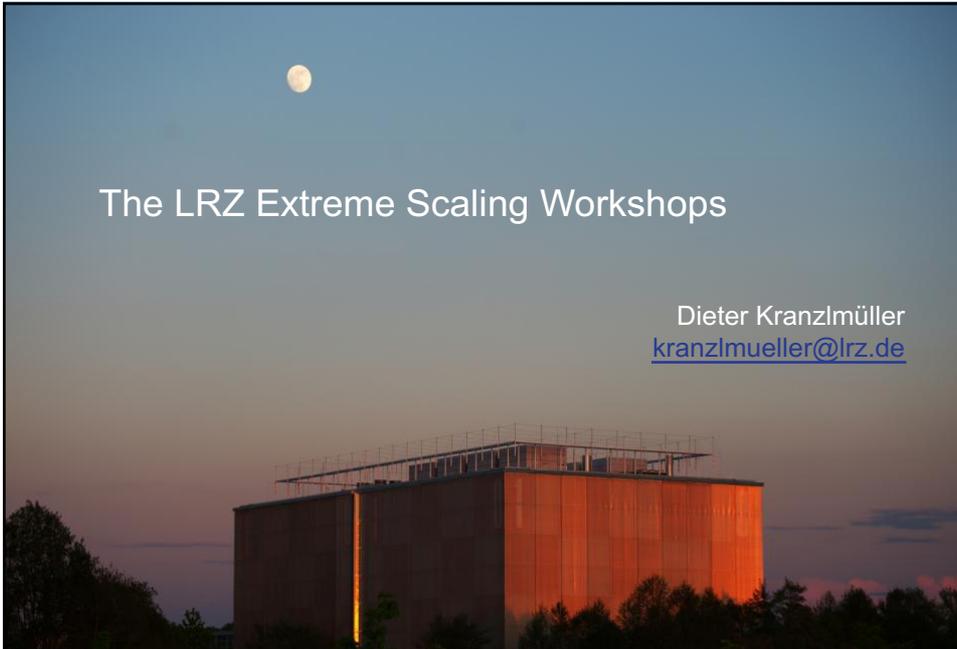
1,42 Petaflop/s on 147.456 Cores of SuperMUC
 (44,5 % of Peak Performance)
http://www.uni-muenchen.de/informationen_fuer/presse/presseinformationen/2014/pelties_seisol.html

Picture: Alex Breuer (TUM) / Christian Pelties (LMU)

MNM D. Kranzmüller SkaIA Workshop @ SC16 33

The LRZ Extreme Scaling Workshops

Dieter Kranzmüller
kranzmueller@lrz.de



lrz  **MCSG** **bgce**  **KONWIHR** **GCS** **GA**  **Gauss Alliance** **PRACE** **prospect-hpc**  **ETP 4 HPC**