

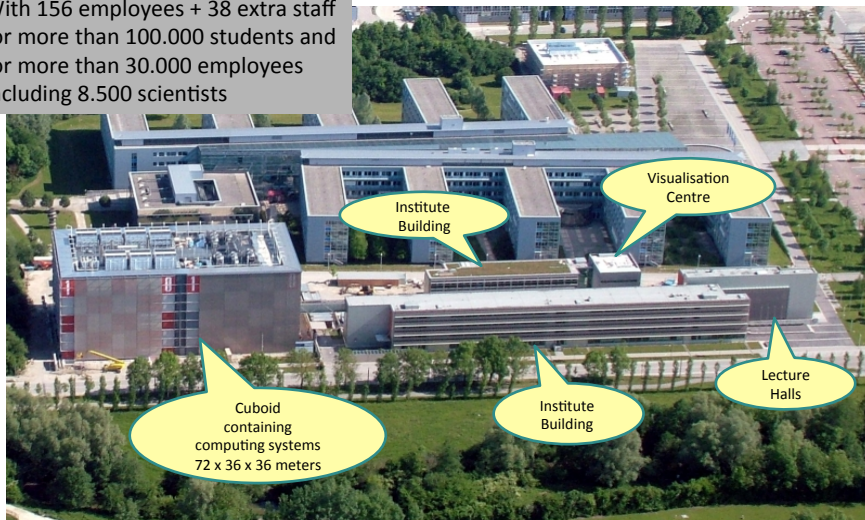
# Extreme Scaling vs. Energy Efficiency – A Challenge for Computer Science

Dieter Kranzlmüller

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



With 156 employees + 38 extra staff for more than 100.000 students and for more than 30.000 employees including 8.500 scientists



LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

## Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities

**■ 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, ...)

D. Kranzmüller

ACOMP 2015 3

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

## The Munich Scientific Network (MWN)

<span style="color: yellow;">●</span> Hochschule	<span style="color: orange;">●</span> Studienarchiv/Erziehung	<span style="color: blue;">●</span> 10-100 Mbps	<span style="color: red;">●</span> 1-100 Mbps	<span style="color: green;">●</span> 10-100 Mbps	<span style="color: purple;">●</span> 1-100 Mbps
<span style="color: yellow;">○</span> Fachbereich	<span style="color: orange;">○</span> sonstige Einrichtungen	<span style="color: blue;">○</span> 10-100 Mbps	<span style="color: red;">○</span> 1-100 Mbps	<span style="color: green;">○</span> 10-100 Mbps	<span style="color: purple;">○</span> 1-100 Mbps
<span style="color: yellow;">◇</span> Fachbereich	<span style="color: orange;">◇</span> sonstige Einrichtungen	<span style="color: blue;">◇</span> 10-100 Mbps	<span style="color: red;">◇</span> 1-100 Mbps	<span style="color: green;">◇</span> 10-100 Mbps	<span style="color: purple;">◇</span> 1-100 Mbps

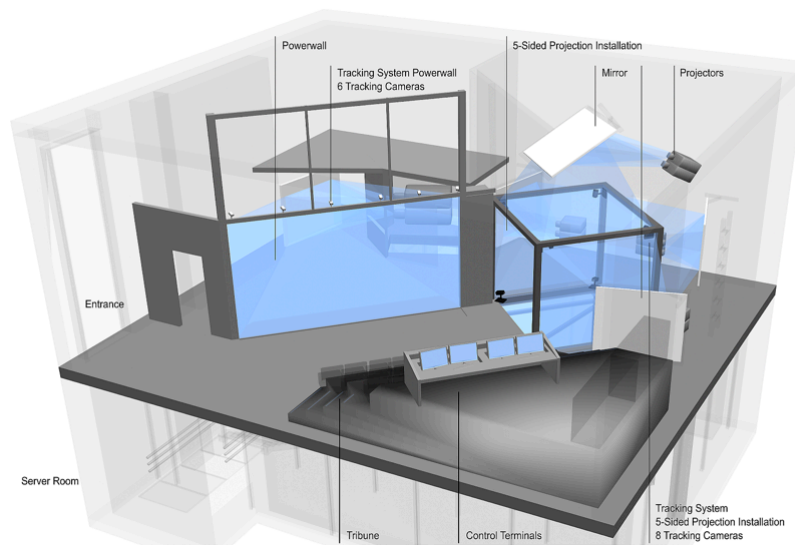
Stand: 31.12.2010/Kantor

D. Kranzmüller

ACOMP 2015 4

■ Regional Computer  
Centre for all  
Bavarian Universities

■ Computer Centre for all  
Munich Universities



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

## Examples from the V2C

lrz

MNM D. Kranzmüller

ACOMP 2015 7

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

## Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities

lrz

- National Supercomputing Centre
- Regional Computer Centre for all Bavarian Universities
- Computer Centre for all Munich Universities

MNM D. Kranzmüller

ACOMP 2015 8

- 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)



- 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 INFOSO-RI-211528 and 261557 Complemented by ~ 60 Million € from PRACE members



**LMU** LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN **PRACE Tier-0 Systems** **lrz**

- **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





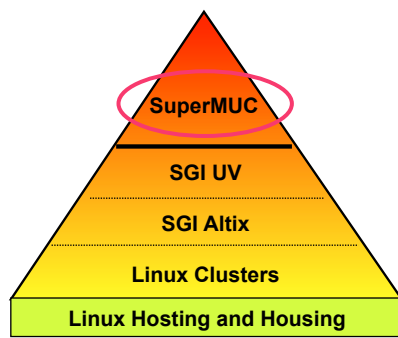




MNM D. Kranzmüller ACOMP 2015 11

**LMU** LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN **Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities** **lrz**

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



MNM D. Kranzmüller ACOMP 2015 12

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

## SuperMUC @ LRZ



Video: SuperMUC rendered on SuperMUC by LRZ  
<http://youtu.be/OIAS6iiqWrQ>

D. Kranzlmüller

ACOMP 2015 13

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

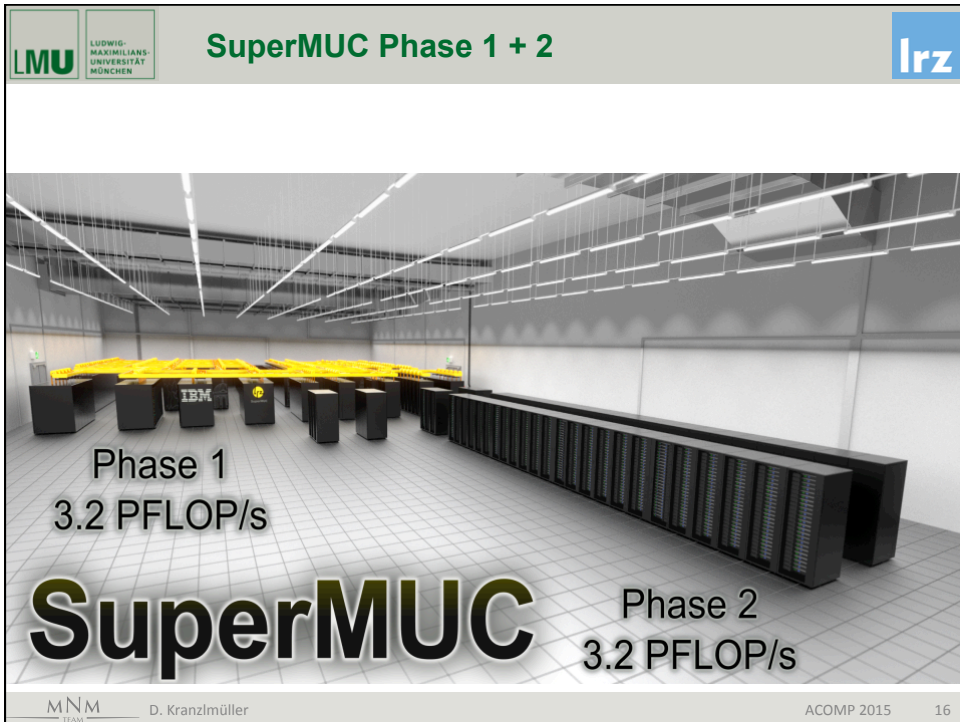
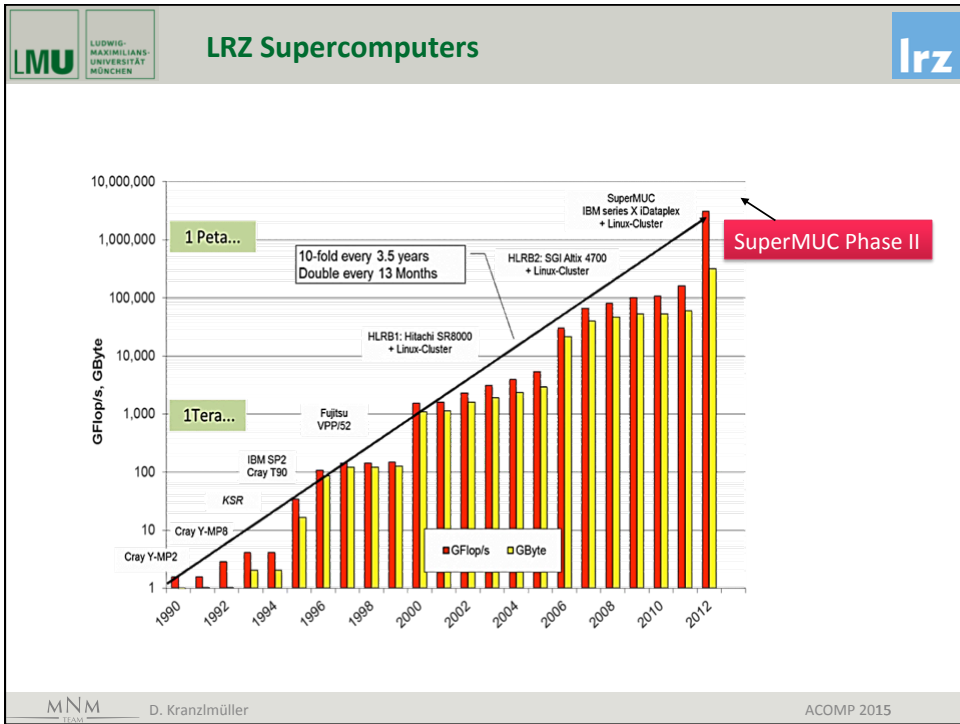
## Top 500 Supercomputer List (June 2012)

Rank	Site	Computer/Year Vendor	Cores	R <sub>max</sub>	R <sub>peak</sub>	Power
1	DOE/NNSA/LLNL United States	<b>Sequoia</b> - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom / 2011 IBM	1572864	16324.75	20132.66	7890.0
2	RIKEN Advanced Institute for Computational Science (AICS) Japan	<b>K computer</b> , SPARC64 VIIItx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
3	DOE/SC/Argonne National Laboratory United States	<b>Mira</b> - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	786432	8162.38	10066.33	3945.0
4	Leibniz Rechenzentrum Germany	<b>SuperMUC</b> - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR / 2012 IBM	147456	2897.00	3185.05	3422.7
5	National Supercomputing Center in Tianjin China	<b>Tianhe-1A</b> - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
6	DOE/SC/Oak Ridge National Laboratory United States	<b>Jaguar</b> - Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090 / 2009 Cray Inc.	298592	1941.00	2627.61	5142.0
7	CINECA Italy	<b>Fermi</b> - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	163840	1725.49	2097.15	821.9
8	Forschungszentrum Juelich (FZJ) Germany	<b>JuQUEEN</b> - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	131072	1380.39	1677.72	657.5
9	CEA/TGCC-GENCI France	<b>Curie thin nodes</b> - Bullx B510, Xeon E5- 2680 8C 2.700GHz, Infiniband QDR / 2012 Bull	77184	1359.00	1667.17	2251.0
10	National Supercomputing Centre in Shenzhen (NSCS) China	<b>Nebulae</b> - Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0

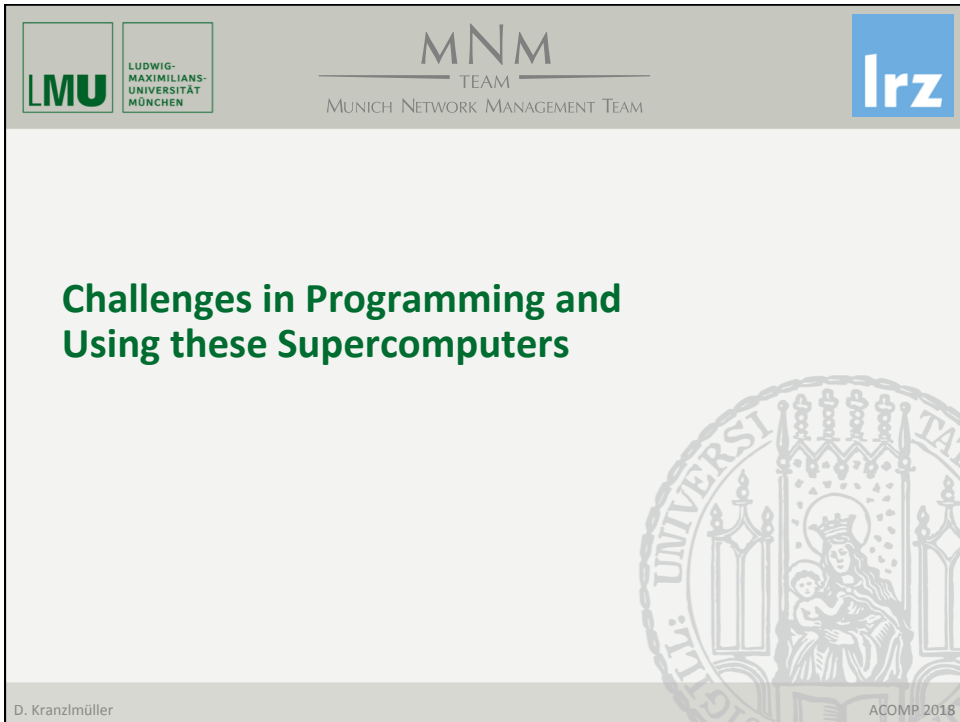
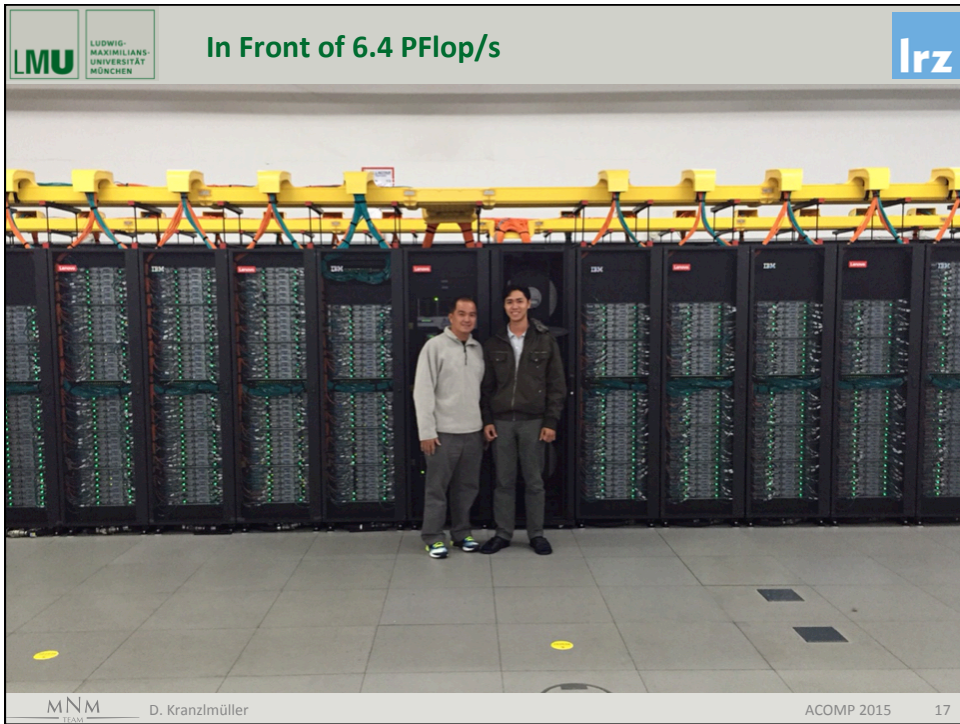
www.top500.org

D. Kranzlmüller

ACOMP 2015 14









**LMU** LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN **SuperMUC and its predecessors** **lrz**

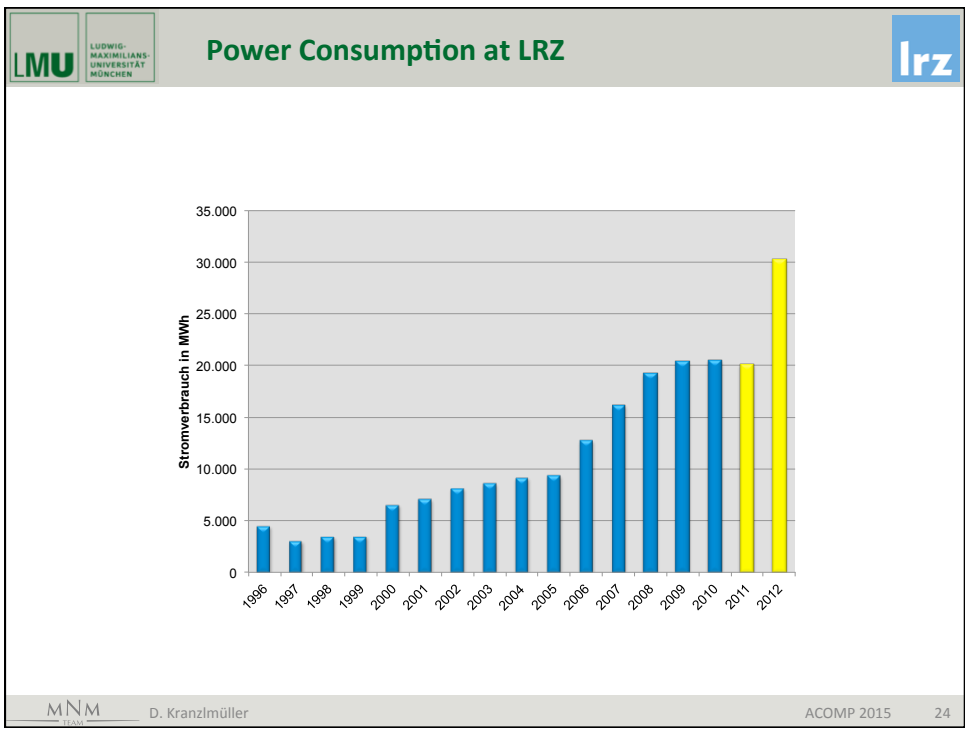
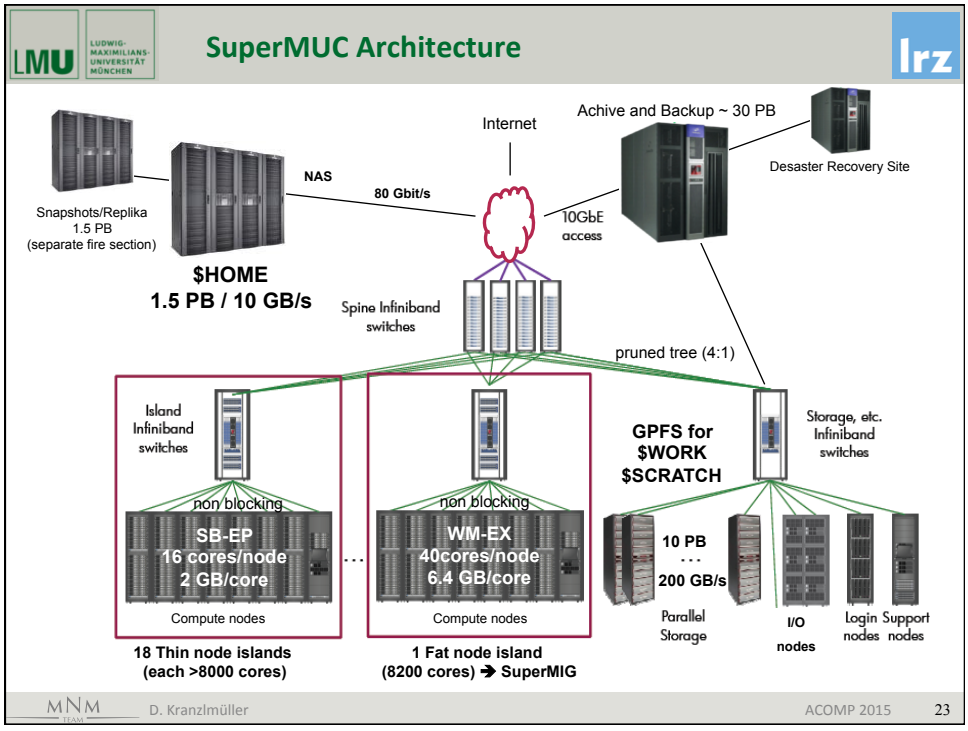
MNM D. Kranzmüller ACOMP 2015 21

**LMU** LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN **LRZ Building Extension** **lrz**

Picture: Horst-Dieter Steinhöfer

Figure: Herzog+Partner für StBAM2 (staatl. Hochbauamt München 2) Picture: Ernst A. Graf


MNM D. Kranzmüller ACOMP 2015 22



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

### Cooling SuperMUC

lrz




MNM D. Kranzmüller

ACOMP 2015 25

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

### SuperMUC Phase 1 & 2 @ LRZ

lrz



MNM D. Kranzmüller

ACOMP 2015 26

## LRZ Application Mix



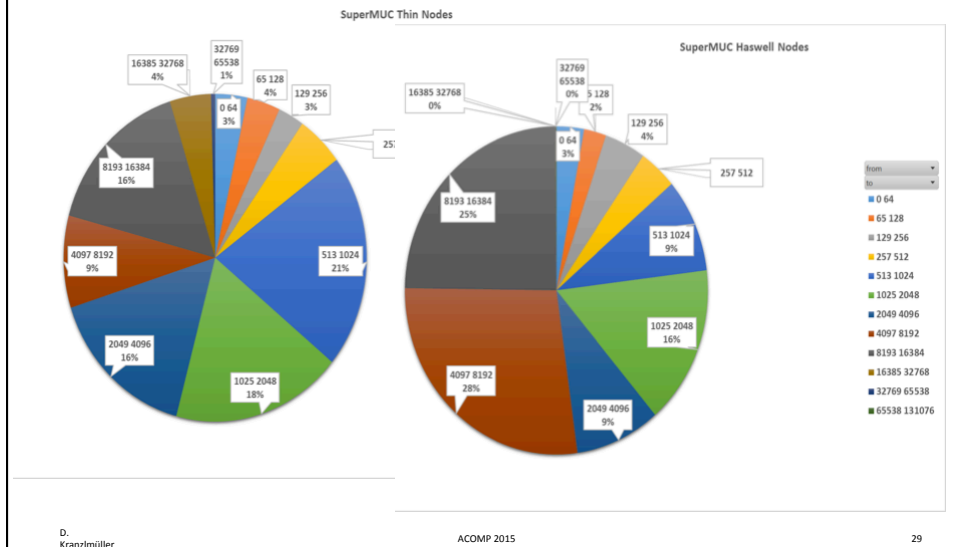
- Computational Fluid Dynamics: Optimisation of turbines and wings, noise reduction, air conditioning in trains**
- Fusion: Plasma in a future fusion reactor (ITER)**
- Astrophysics: Origin and evolution of stars and galaxies**
- Solid State Physics: Superconductivity, surface properties**
- Geophysics: Earth quake 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**

## 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)






LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

## 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




D. Kranzmüller

ACOMP 2015 30




**1<sup>st</sup> LRZ Extreme Scale Workshop**


- July 2013:
  - 1<sup>st</sup> 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


 D. Kranzmüller
 ACOMP 2015 31



**1<sup>st</sup> 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

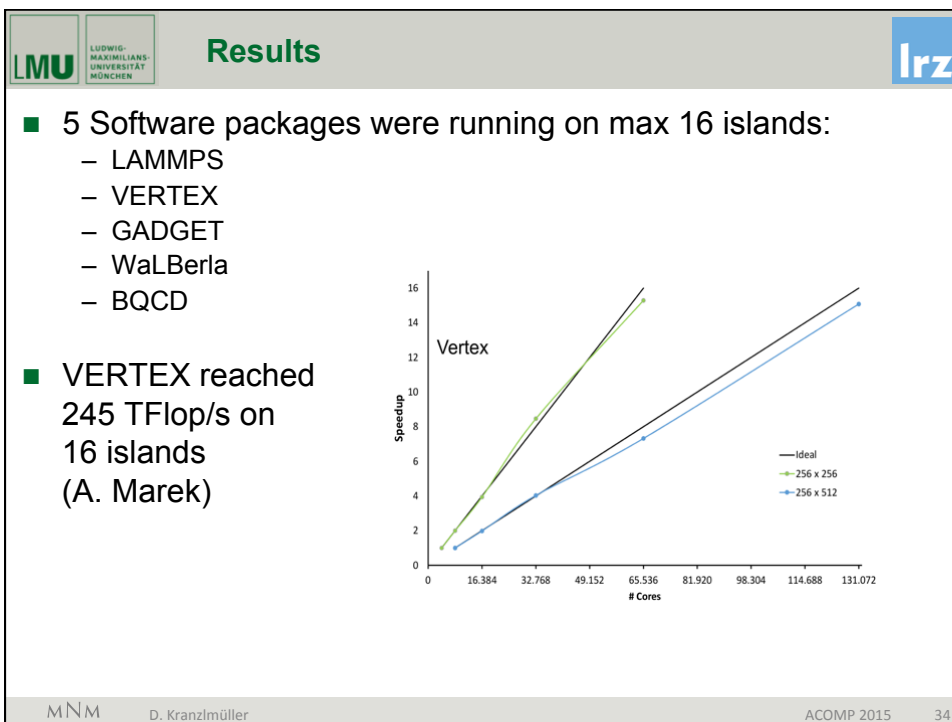

 D. Kranzmüller
 ACOMP 2015 32






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 ACOMP 2015 33





**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
- 2<sup>nd</sup> 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)
- Extreme Scale-Out with new SuperMUC Phase 2



 D. Kranzmüller
 ACOMP 2015 35






**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


 D. Kranzmüller
 ACOMP 2015 36

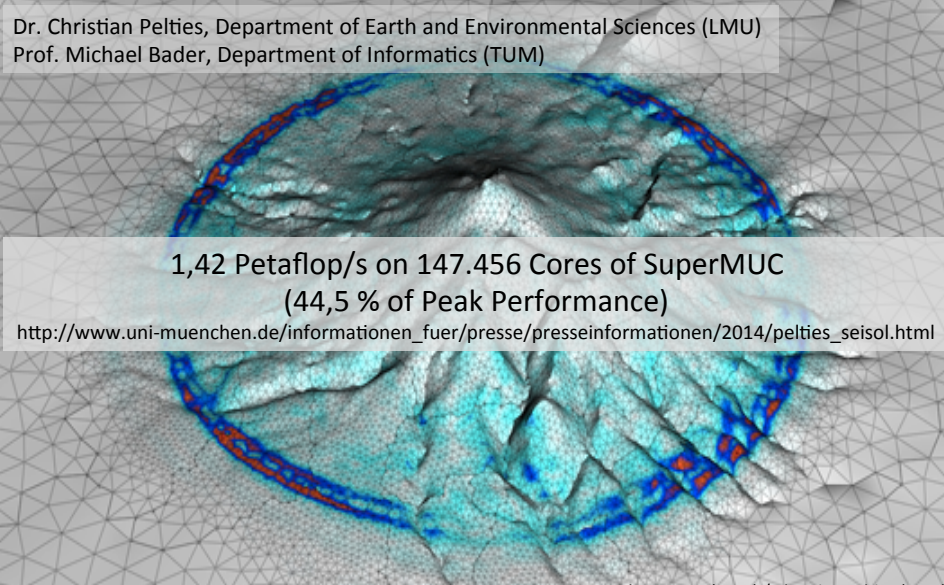


**Partnership Initiative**  
**Computational Sciences  $\pi$ CS**


- **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 CS and Math departments
- **Partner contribution**
  - Embedding IT experts in user groups
  - Joint research projects (including funding)
  - Scientific partnership – equal footing – joint publications
- **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**


 D. Kranzmüller
 

 37



**SeisSol - Numerical Simulation of Seismic Wave Phenomena**




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

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


 D. Kranzmüller
 

 38

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

## Extreme Scaling - Conclusions

lrz

- The number of compute cores, the complexity (and heterogeneity) is steadily increasing
- 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 2016
- The lessons learned from the Extreme Scaling Workshop are very valuable for the operation of the center
- The LRZ Partnership Initiative Computational Science (piCS) tries to improve user support  
<http://www.sciencedirect.com/science/article/pii/S1877050914003433>

MNM D. Kranzmüller ACOMP 2015 39

## Extreme Scaling vs. Energy Efficiency - A Challenge for Computer Science

Dieter Kranzmüller  
[kranzmuller@lrz.de](mailto:kranzmuller@lrz.de)

lrz MCSC bgce ETR Network of Europe KONWIHR GCS GA Gauß-Allianz PRACE prospect-hpc ETR 4 HPC