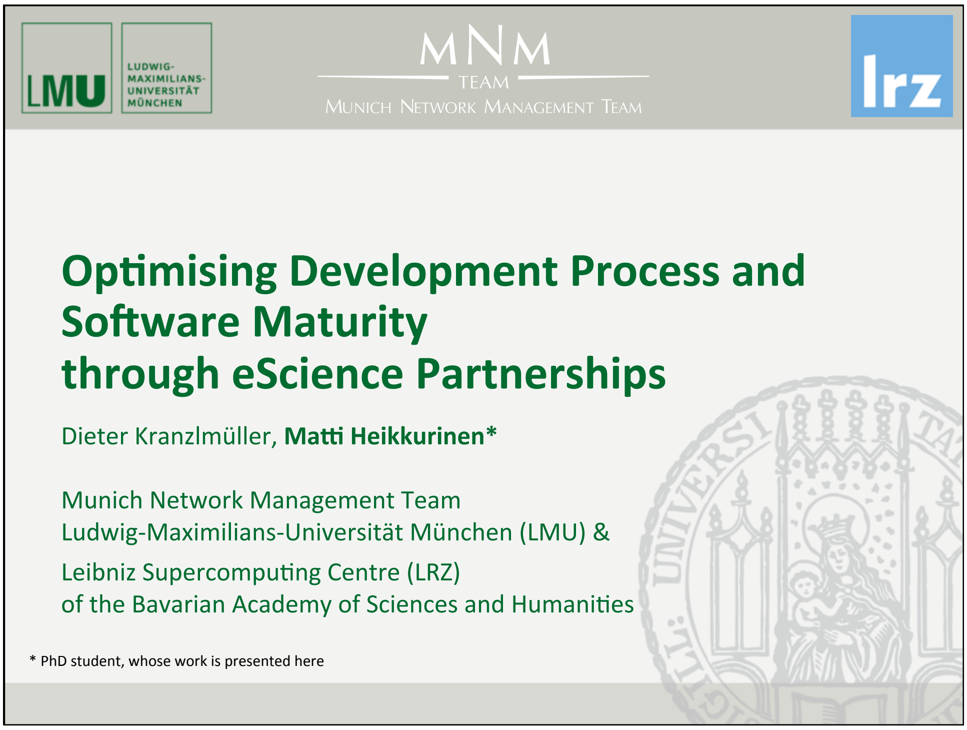


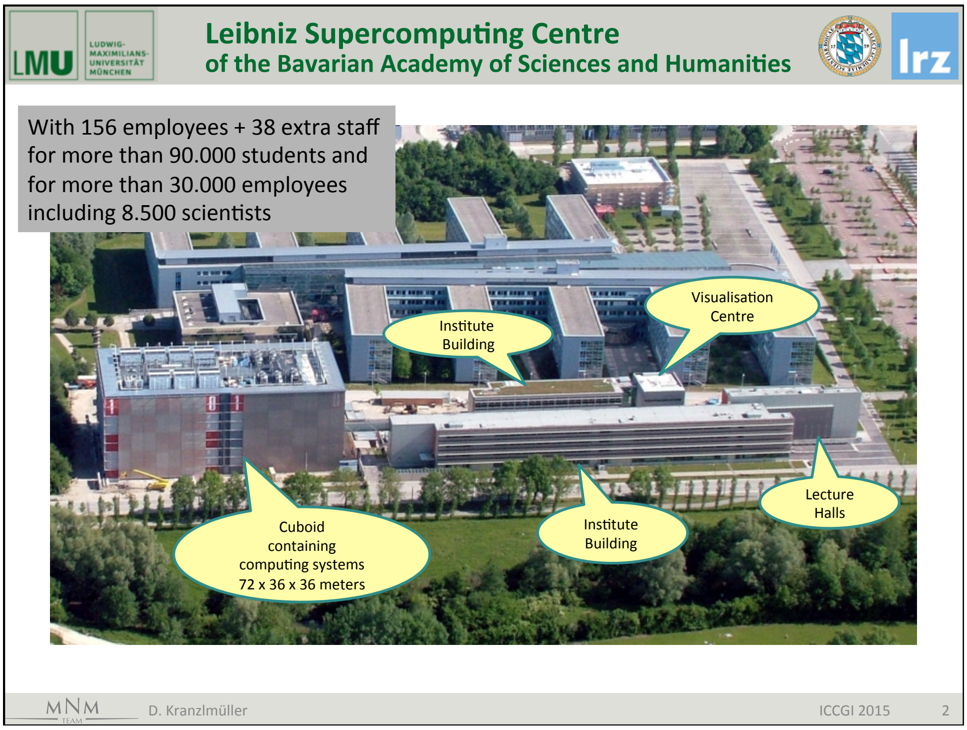
MNM TEAM
MUNICH NETWORK MANAGEMENT TEAM

Optimising Development Process and Software Maturity through eScience Partnerships

Dieter Kranzlmüller, **Matti Heikkurinen***

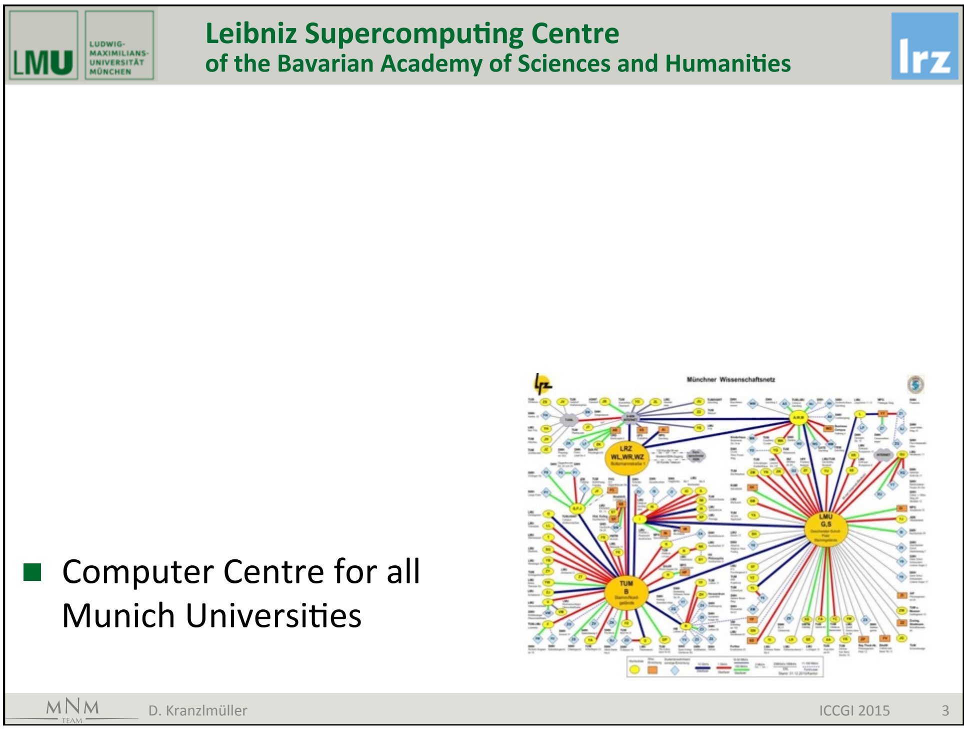
Munich Network Management Team
Ludwig-Maximilians-Universität München (LMU) &
Leibniz Supercomputing Centre (LRZ)
of the Bavarian Academy of Sciences and Humanities

* PhD student, whose work is presented here



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



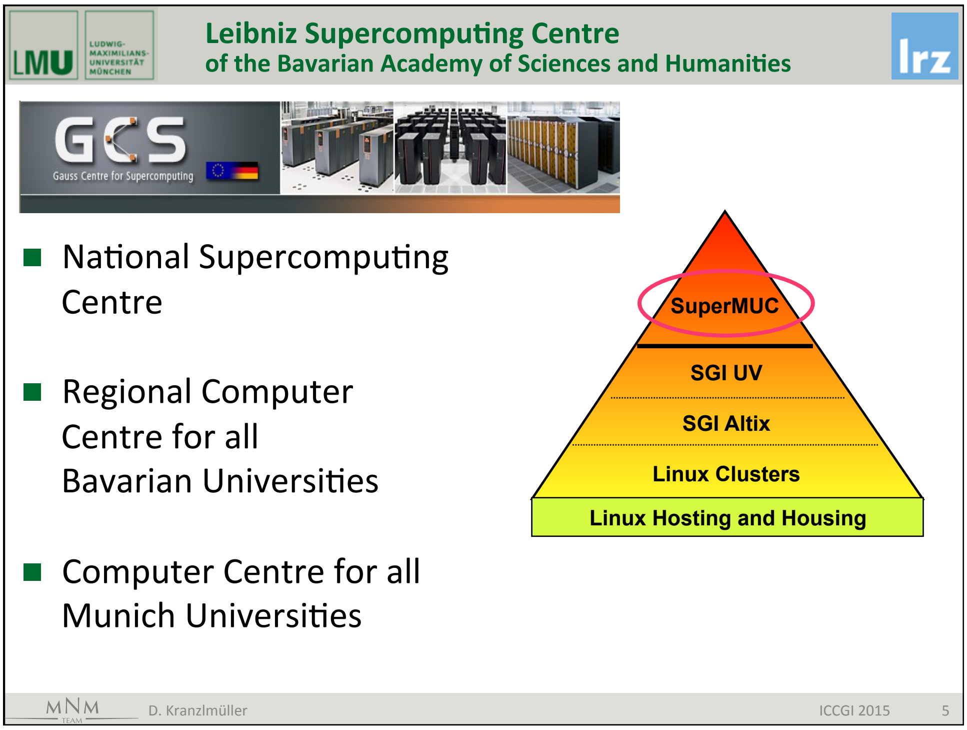
Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

- Computer Centre for all Munich Universities

Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

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

Powerwall
5-Sided Projection Installation
Tracking System Powerwall
6 Tracking Cameras
Mirror
Projectors
Entrance
Server Room
Tribüne
Control Terminals
Tracking System
5-Sided Projection Installation
6 Tracking Cameras



Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

GCS
Gauss Centre for Supercomputing

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

SuperMUC

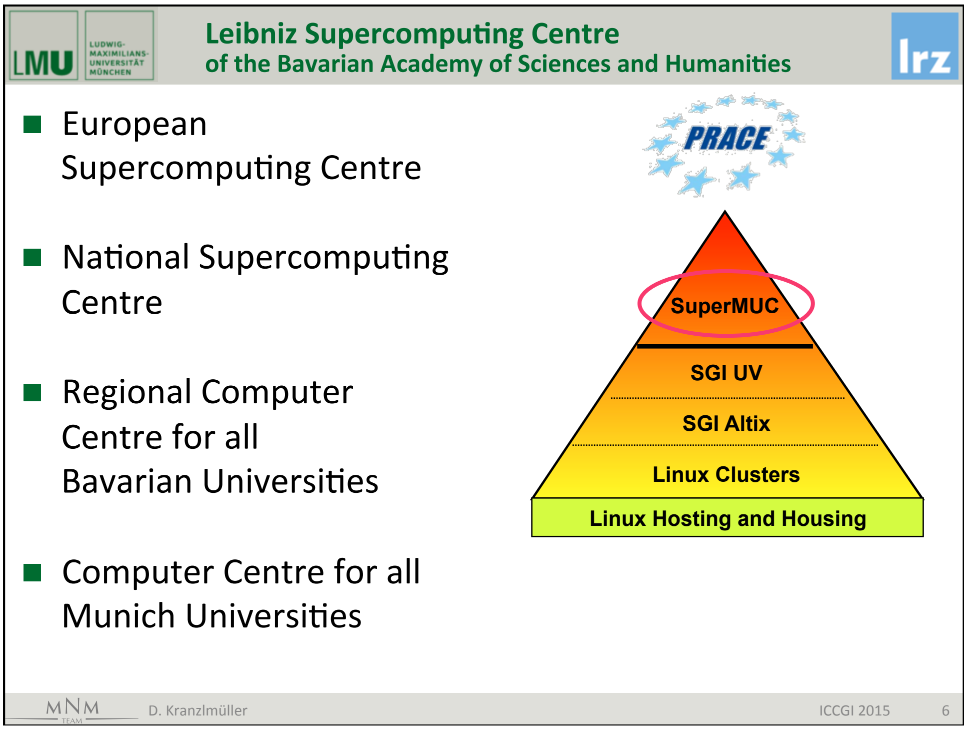
SGI UV

SGI Altix

Linux Clusters

Linux Hosting and Housing

D. Kranzlmüller ICCGI 2015 5



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

SGI UV

SGI Altix

Linux Clusters

Linux Hosting and Housing

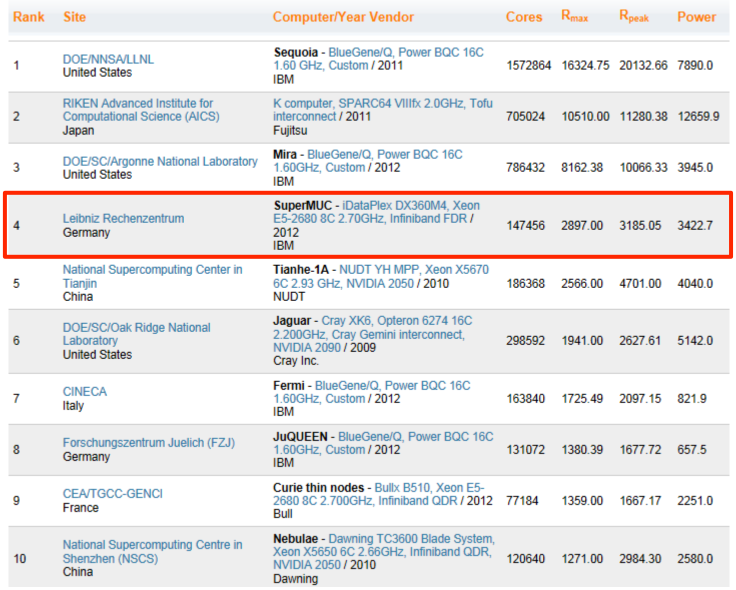
D. Kranzlmüller ICCGI 2015 6

SuperMUC @ LRZ

Video: **SuperMUC rendered on SuperMUC by LRZ**

http://youtu.be/OlAS6iiqWrQ

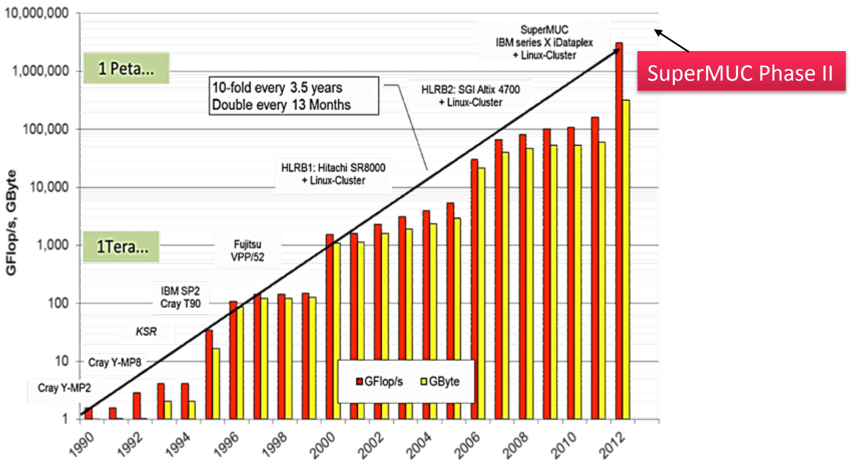
Top 500 Supercomputer List (June 2012)



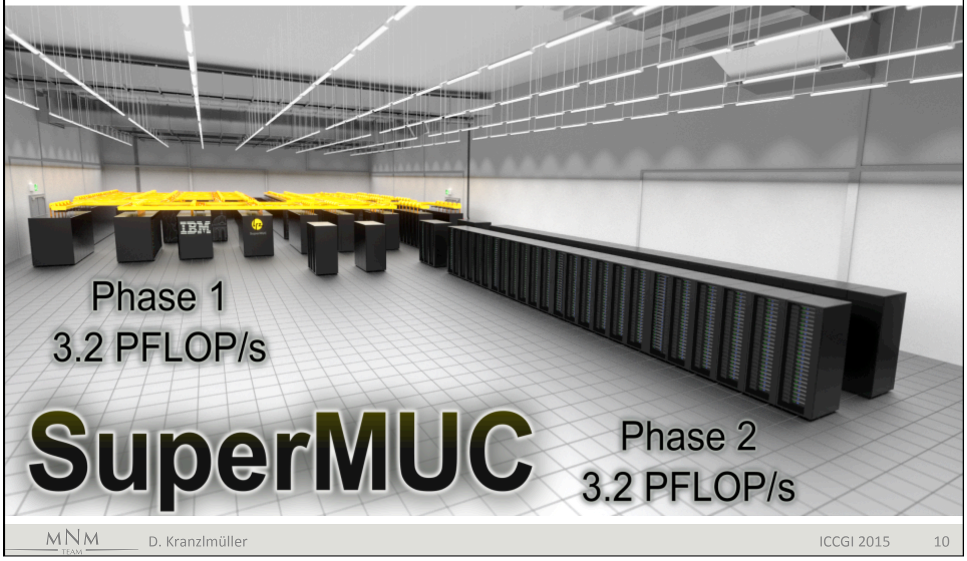
Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom / 2011 IBM	1572864	16324.75	20132.66	7890.0
2	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
3	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	786432	8162.38	10066.33	3945.0
4	Leibniz Rechenzentrum Germany	**SuperMUC** - 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	**Tianhe-1A** - 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	**Jaguar** - 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	**Fermi** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	163840	1725.49	2097.15	821.9
8	Forschungszentrum Juelich (FZJ) Germany	**JuQUEEN** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	131072	1380.39	1677.72	657.5
9	CEA/TGCC-GENCI France	**Curie thin nodes** - 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	**Nebulae** - Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0

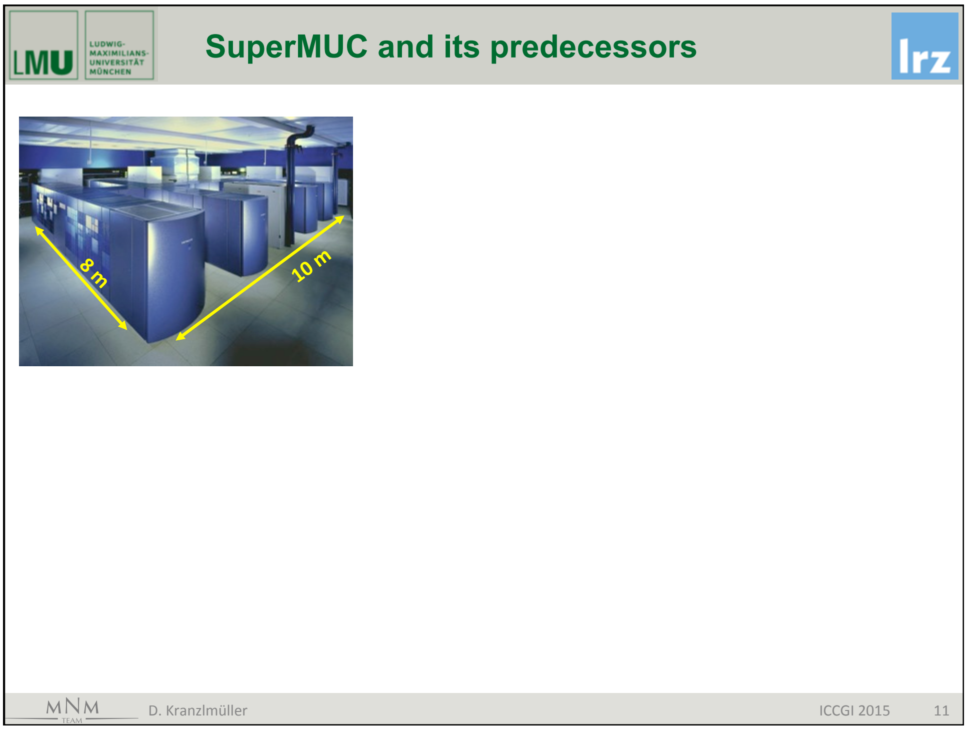
www.top500.org

LRZ Supercomputers

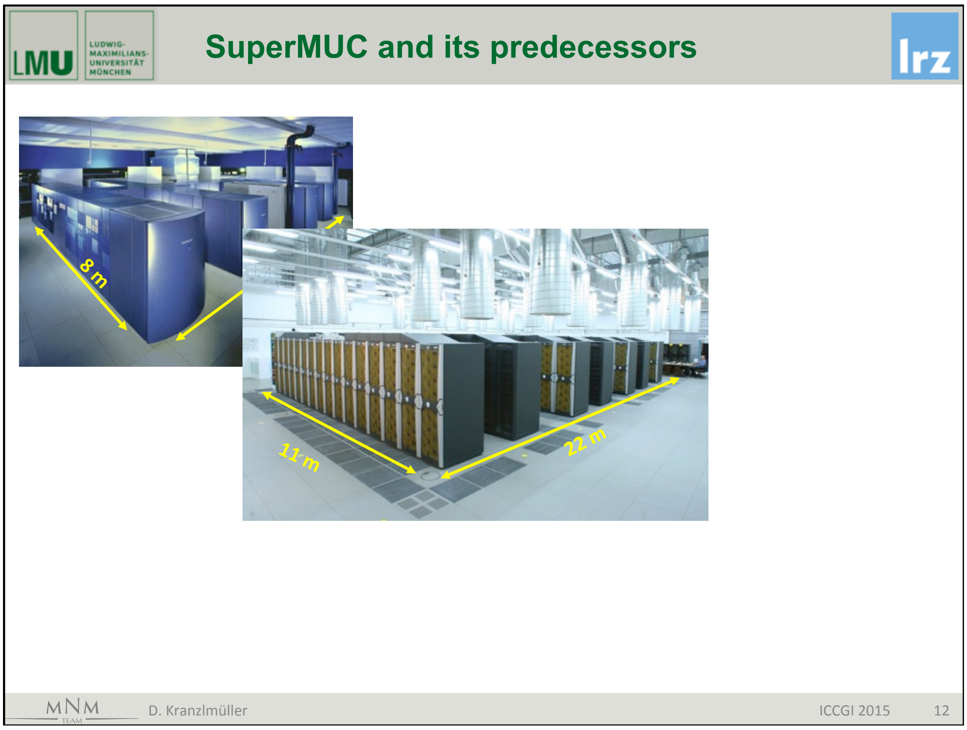


SuperMUC Phase 1 + 2



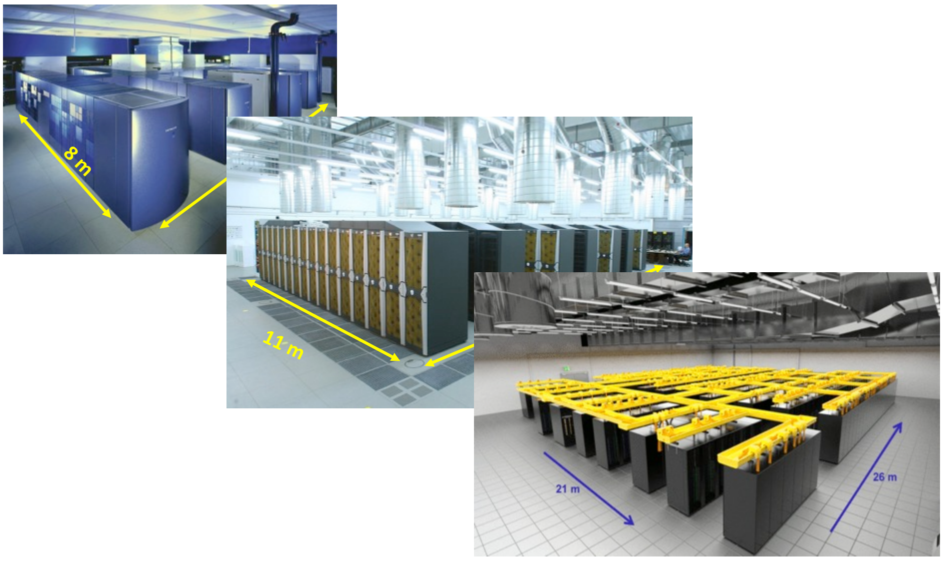


SuperMUC and its predecessors



SuperMUC and its predecessors

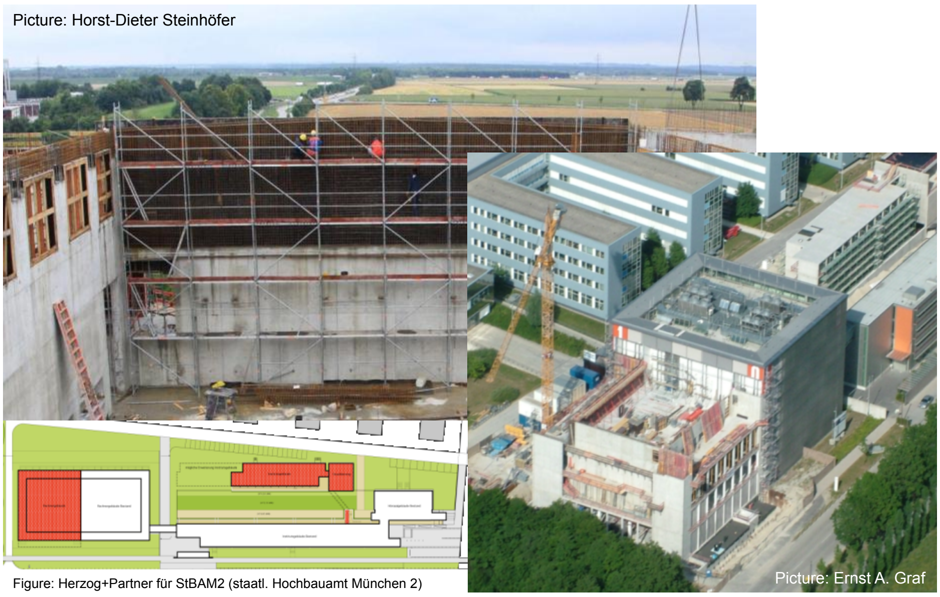
SuperMUC and its predecessors



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

LRZ Building Extension

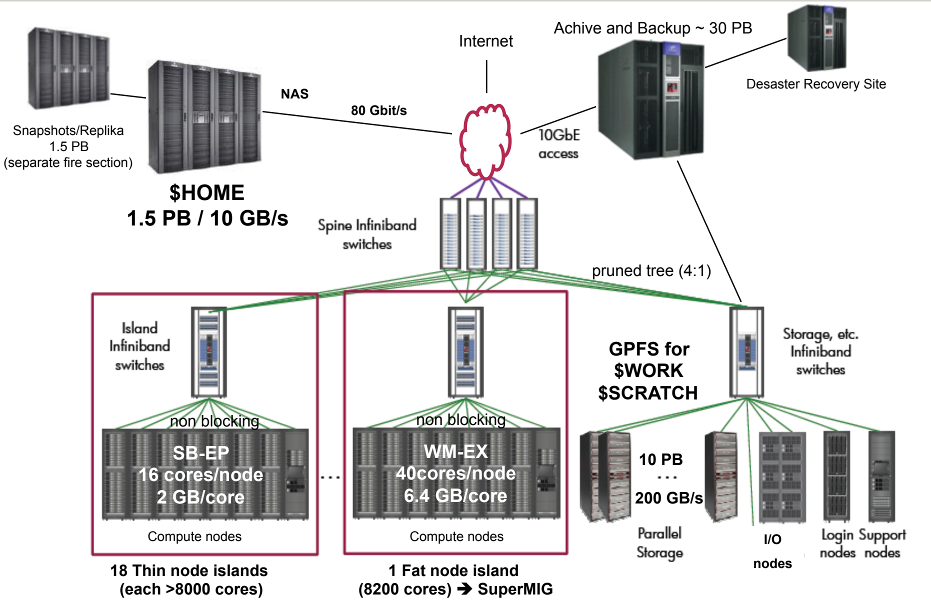


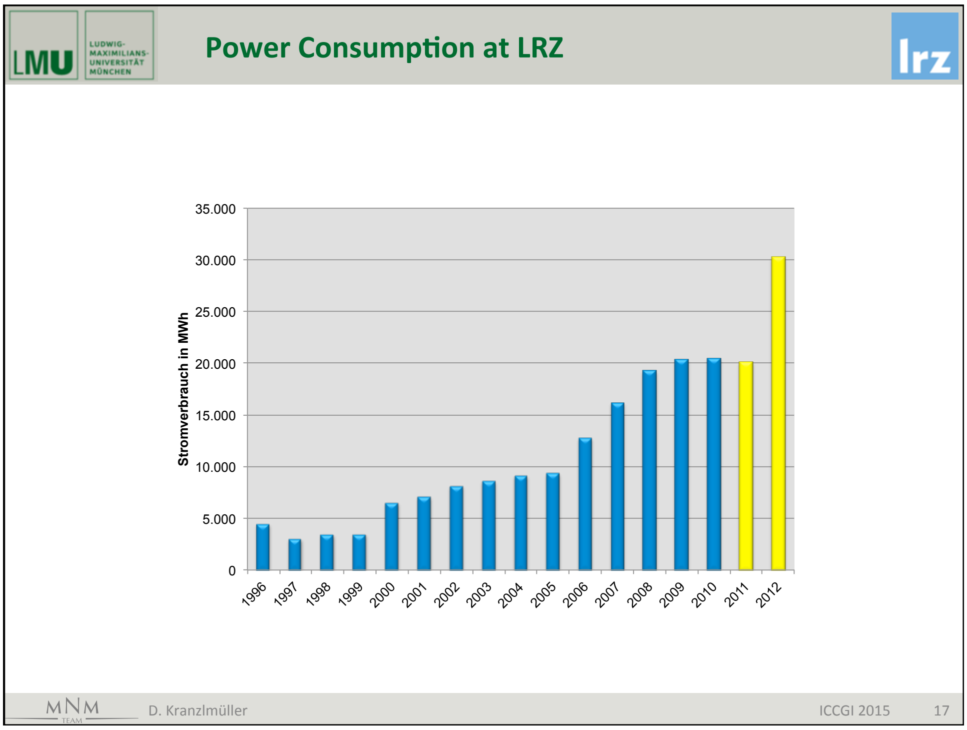
Picture: Horst-Dieter Steinhöfer

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

Picture: Ernst A. Graf

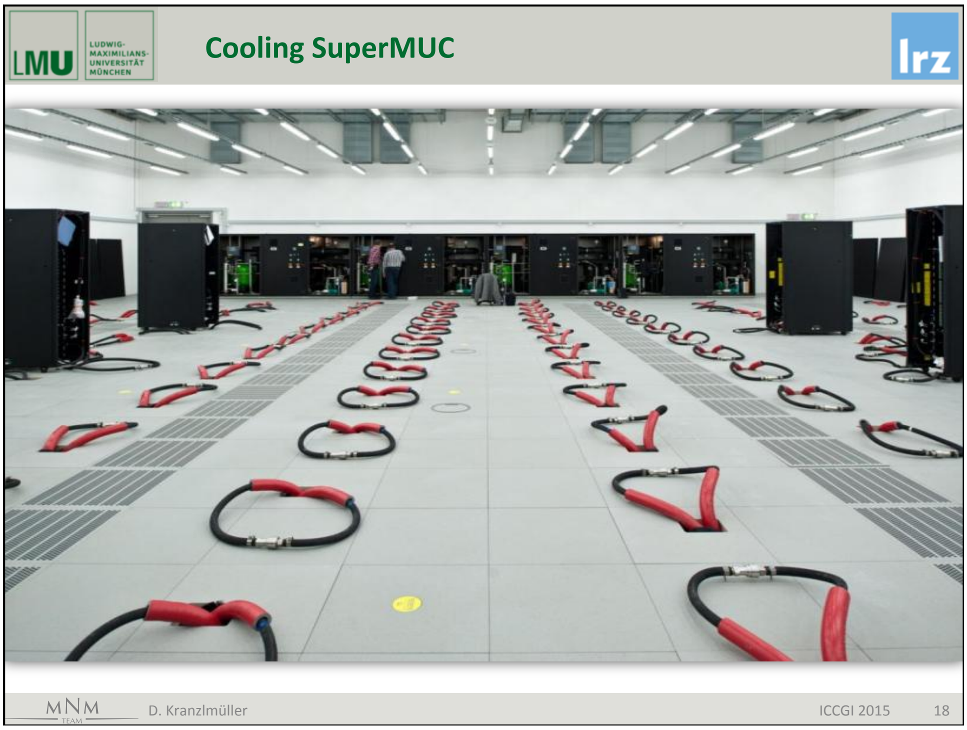
SuperMUC Architecture





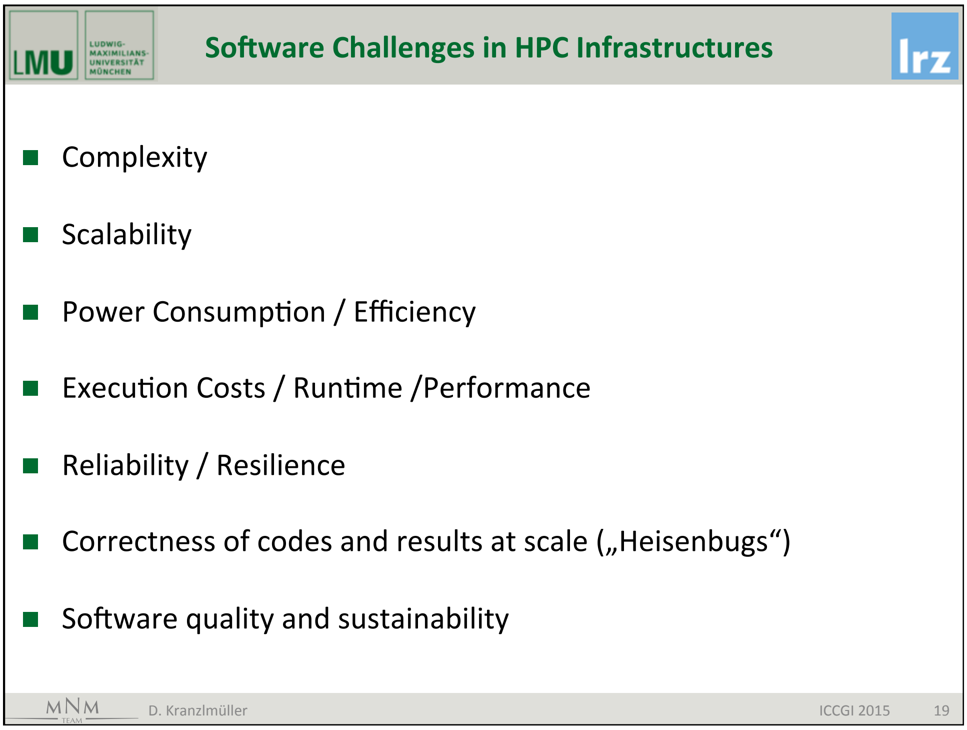
Power Consumption at LRZ

D. Kranzlmüller ICCGI 2015 17



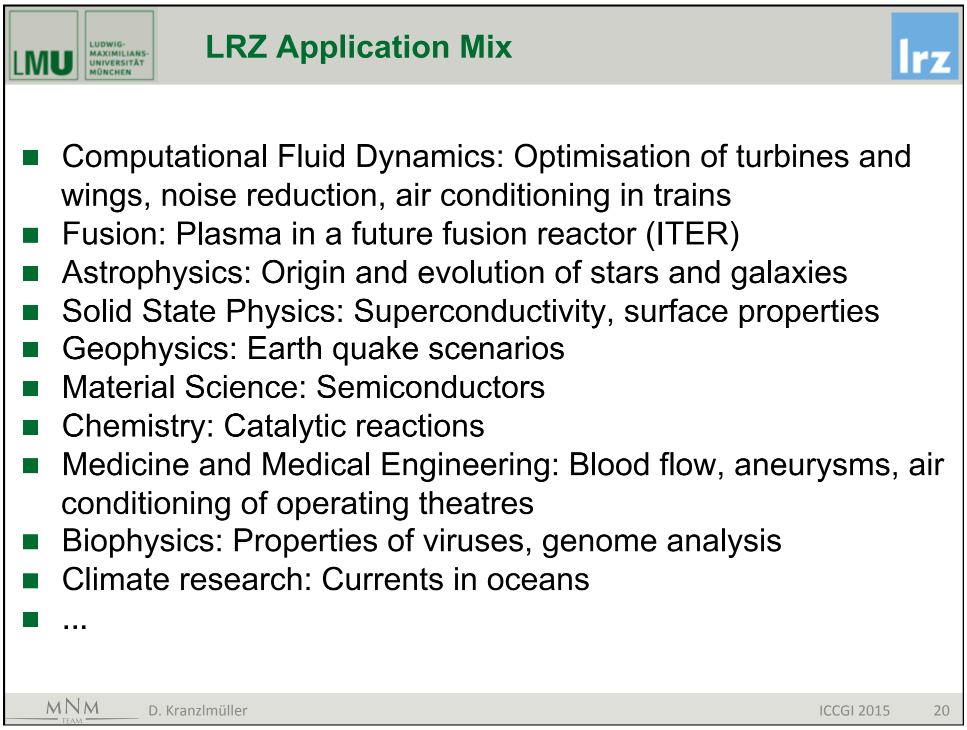
Cooling SuperMUC

D. Kranzlmüller ICCGI 2015 18



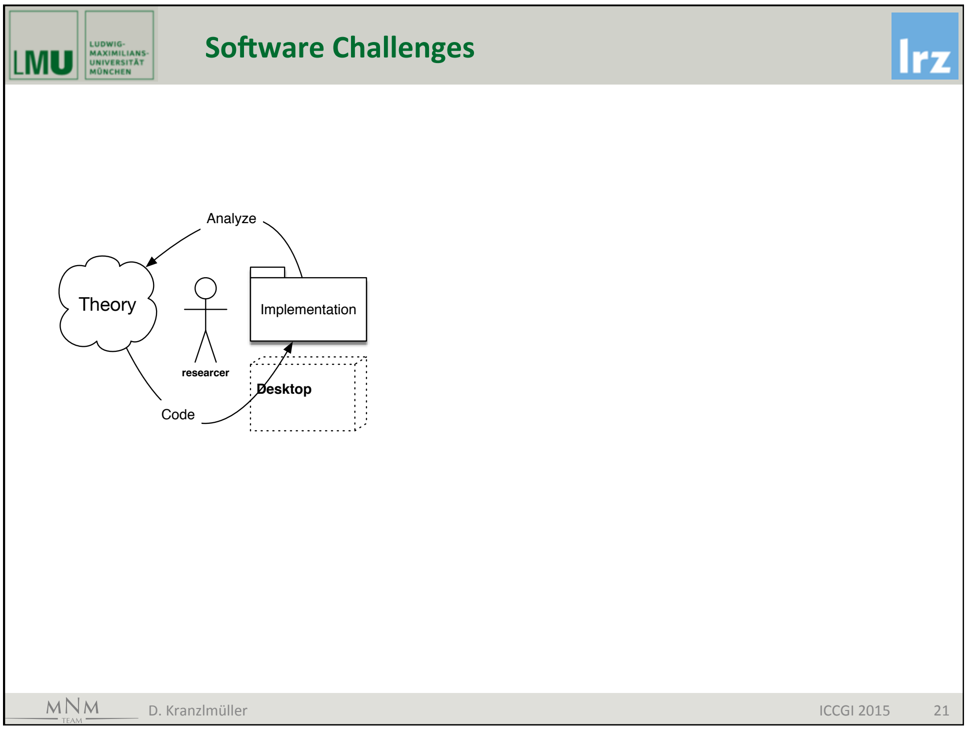
Software Challenges in HPC Infrastructures

- Complexity
- Scalability
- Power Consumption / Efficiency
- Execution Costs / Runtime /Performance
- Reliability / Resilience
- Correctness of codes and results at scale („Heisenbugs“)
- Software quality and sustainability

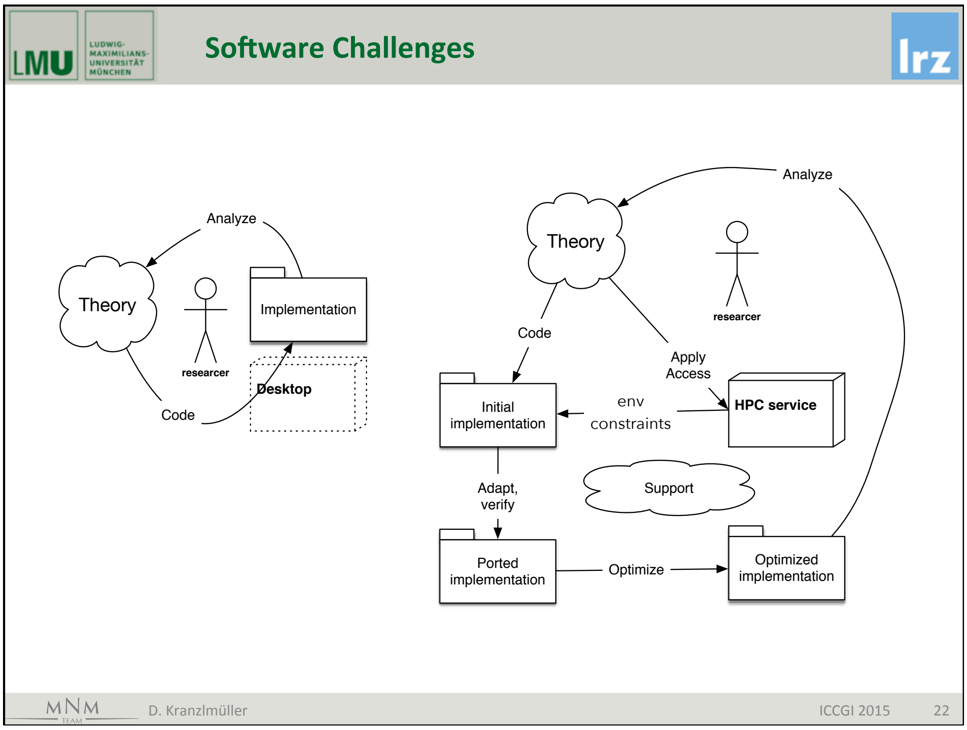


LRZ Application Mix

- 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
- ...

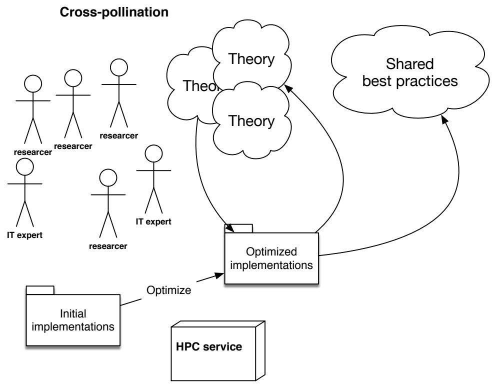


Software Challenges



Software Challenges

Realising the full potential of a HPC centre as a fulcrum



Cross-pollination

researcer

researcer

researcer

IT expert

researcer

IT expert

Theory

Theo

Theory

Shared best practices

Optimized implementations

Optimize

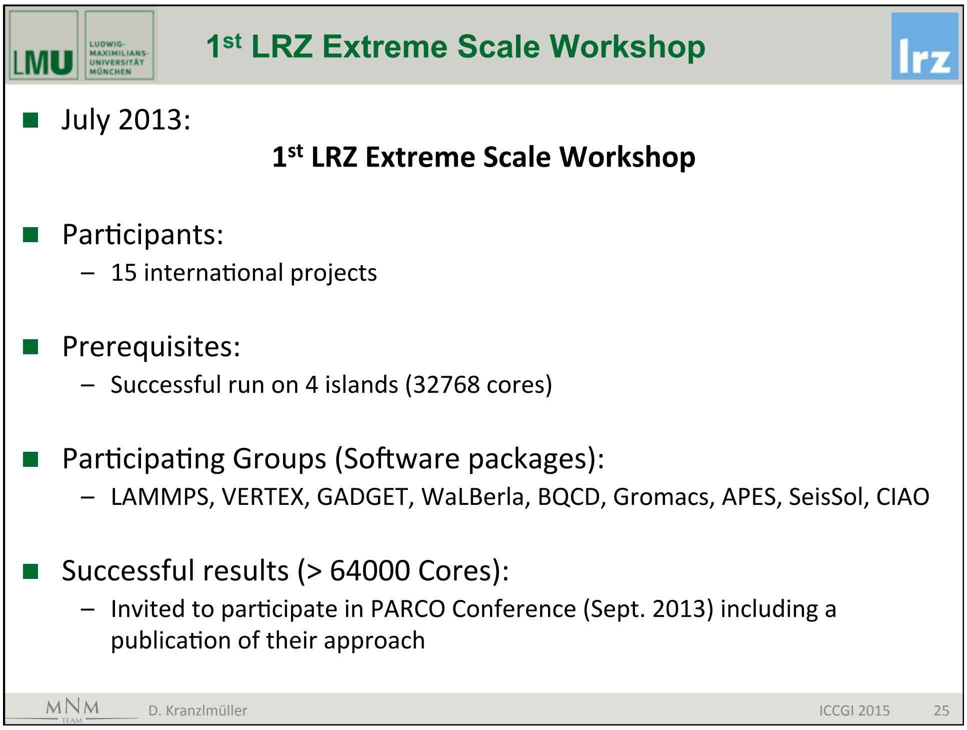
Initial implementations

HPC service

- Many challenges are domain independent
- Reuse of solutions is efficient (common libraries?)
- Fulcrum catalysing interdisciplinary action
- Support for multi-scale, multi-model approaches

How to Measure Success?

- Publications?
  - How to relate software to publications?
  - What about good implementation of a bad theory?
  - Compare Panel „Challenges in Knowledge Sharing“ on Monday
- Software or software development process maturity?
  - No easy/cheap way to measure (think ISO Xk audit...)
  - Penalises new applications and interdisciplinary collaborations?
- Approach: use scalability as proxy indicator
  - Maximum scalability correlates with SW maturity
  - Speed of scalability improvement with process maturity

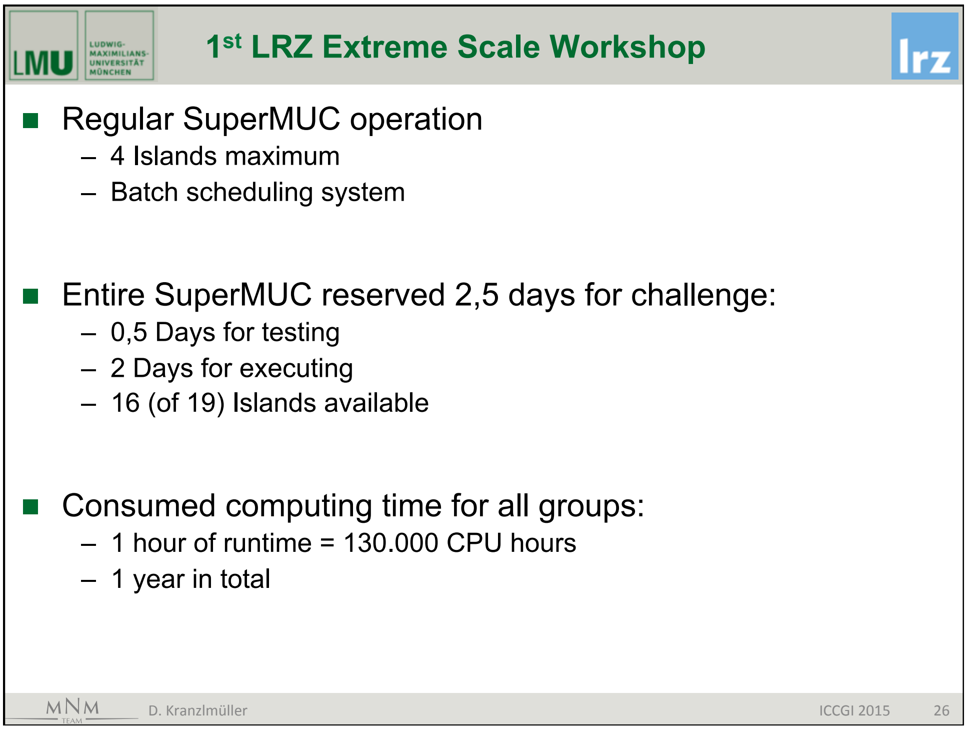


1st LRZ Extreme Scale Workshop

- 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

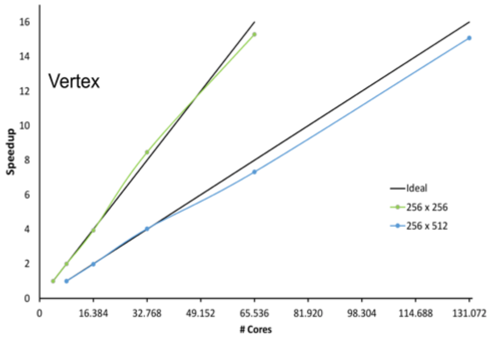
- Regular SuperMUC operation
  - 4 Islands maximum
  - Batch scheduling system
- Entire SuperMUC reserved 2,5 days for challenge:
  - 0,5 Days for testing
  - 2 Days for executing
  - 16 (of 19) Islands available
- Consumed computing time for all groups:
  - 1 hour of runtime = 130.000 CPU hours
  - 1 year in total

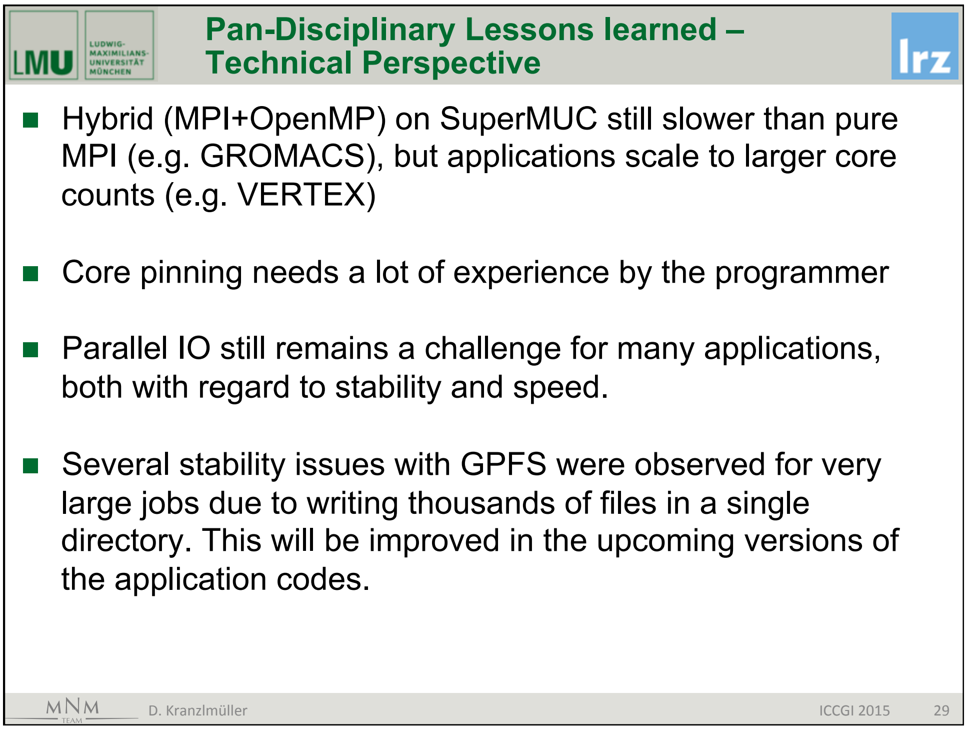
Results (Sustained TFlop/s on 128000 cores)

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

Results

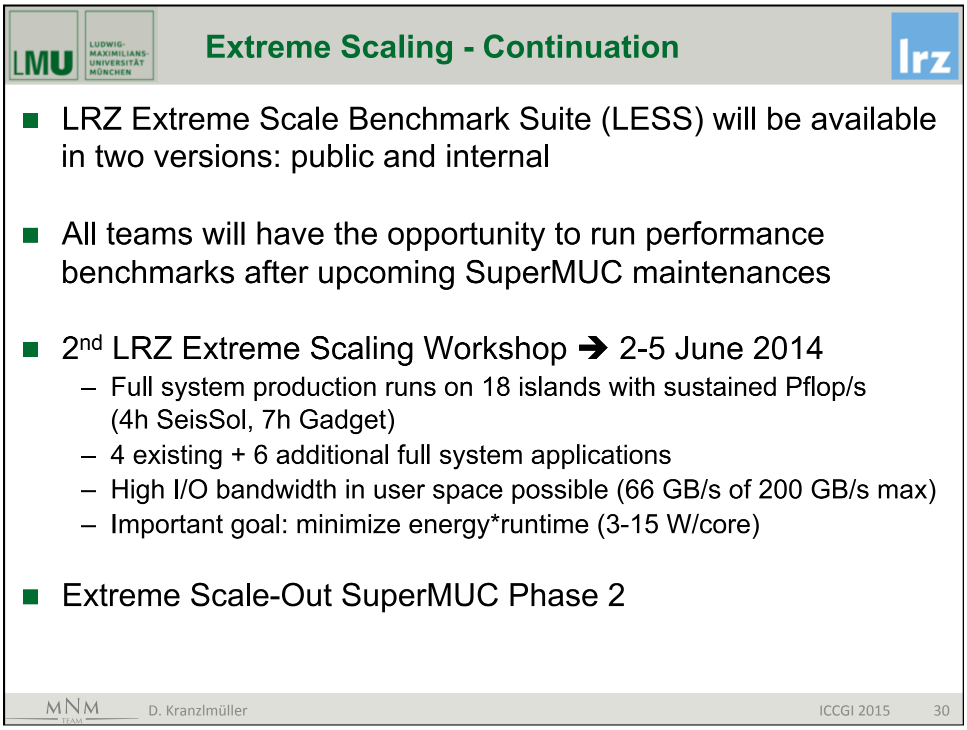
- 5 Software packages were running on max 16 islands:
  - LAMMPS
  - VERTEX
  - GADGET
  - WaLBerla
  - BQCD
- VERTEX reached 245 TFlop/s on 16 islands (A. Marek)





Pan-Disciplinary 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.



Extreme Scaling - Continuation

- 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)
- Extreme Scale-Out 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)

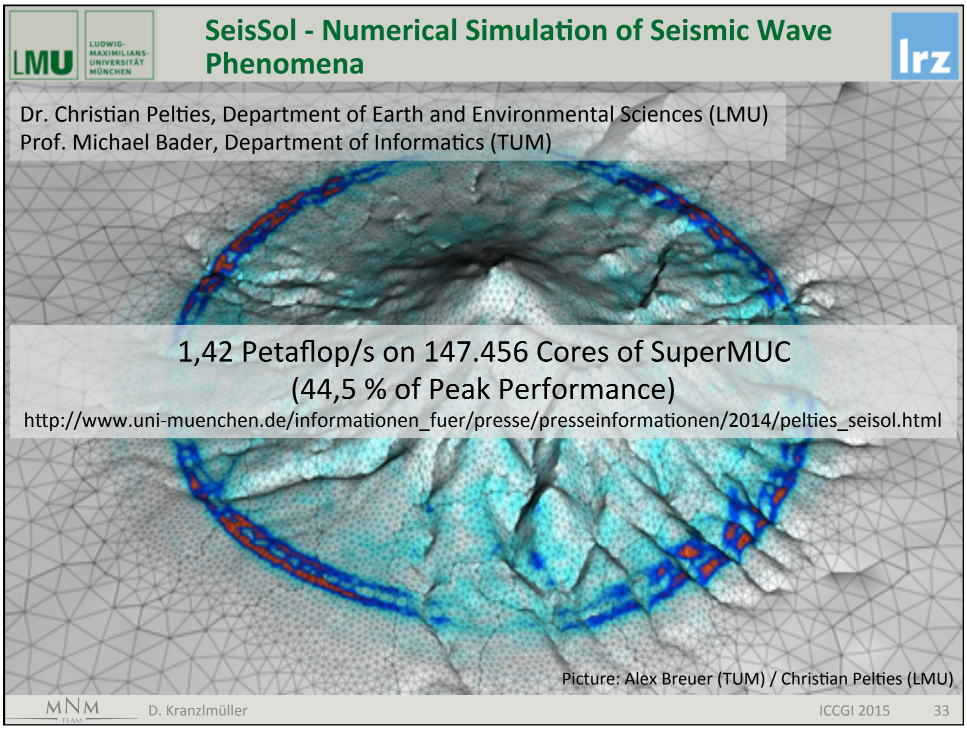
- Total available: 63,432,000 core hours
- Total used: 43,758,430 core hours (Utilisation: 68.98%)

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

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 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**



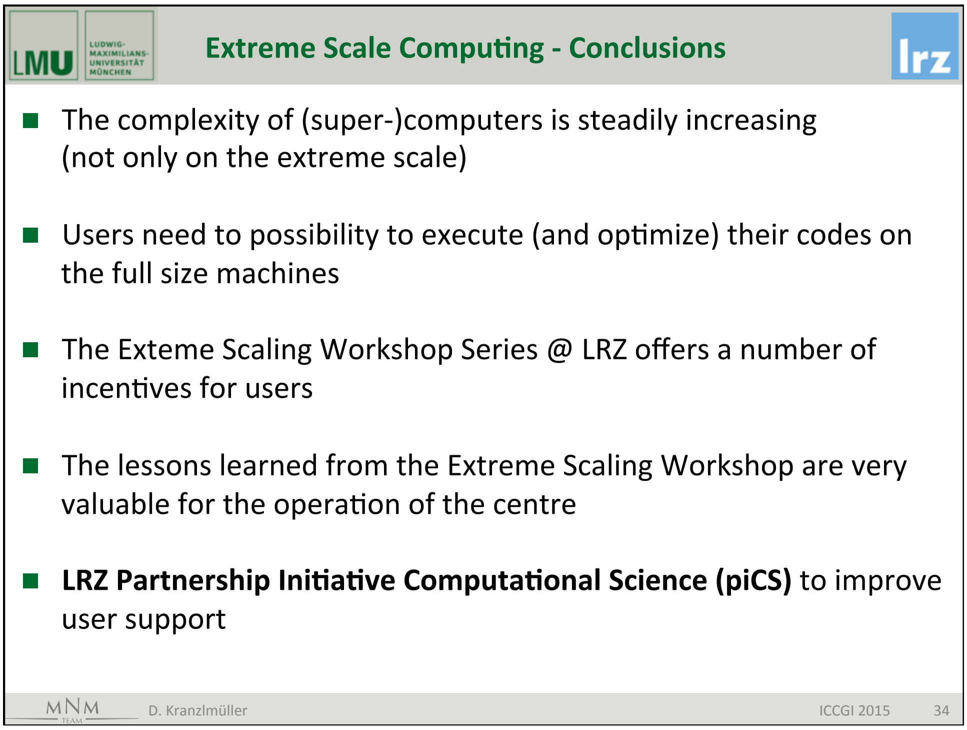
SeisSol - Numerical Simulation of Seismic Wave Phenomena

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

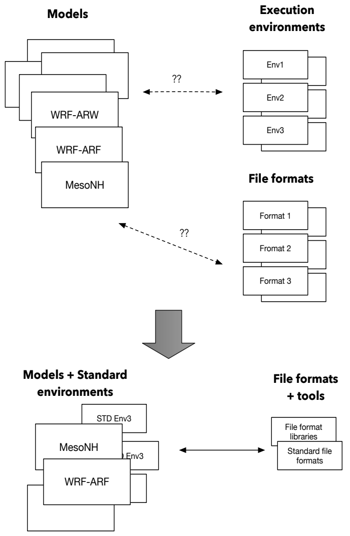
Picture: Alex Breuer (TUM) / Christian Pelties (LMU)



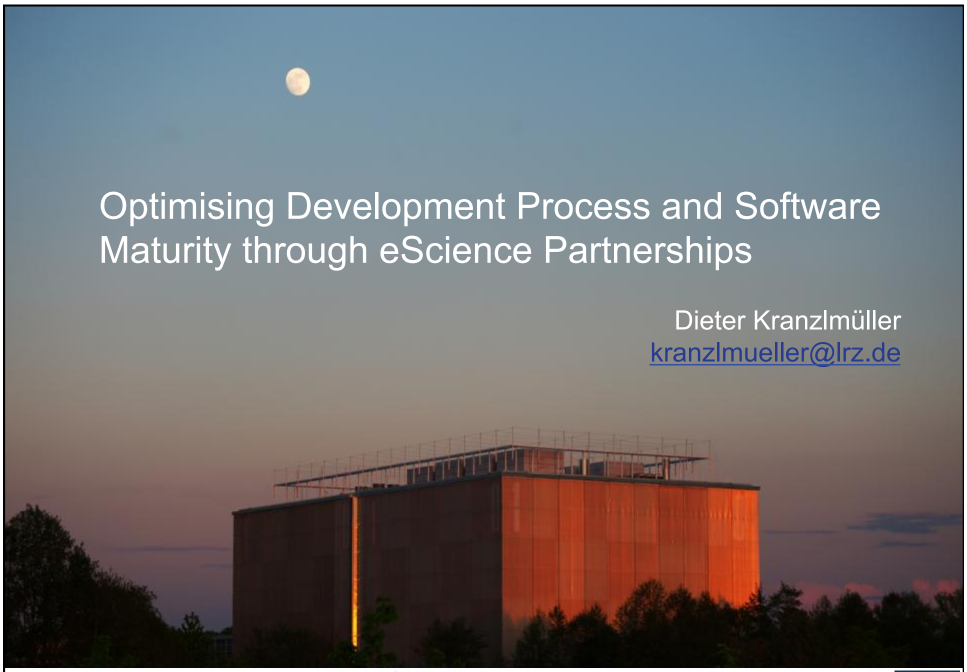
Extreme Scale Computing - Conclusions

- The complexity of (super-)computers is steadily increasing (not only on the extreme scale)
- Users need to possibility to execute (and optimize) their codes on the full size machines
- The Exteme Scaling Workshop Series @ LRZ offers a number of incentives for users
- The lessons learned from the Extreme Scaling Workshop are very valuable for the operation of the centre
- **LRZ Partnership Initiative Computational Science (piCS)** to improve user support

πCS in Action (EU Project DRIHM)



- Starting point: manual multi-model, multi-data
  - Execution time: weeks
- Partnership with delopers
  - Standardise metadata
  - Identify dependencies
  - Build workflow system
- End point: single click execution
  - Execution time: hours



Optimising Development Process and Software Maturity through eScience Partnerships

Dieter Kranzlmüller
kranzlmueller@lrz.de