Environmental Exascale Computing

Dieter Kranzlmüller

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

Gauss Centre for Supercomputing (GCS)

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)

PRACE Research Infrastructure Created

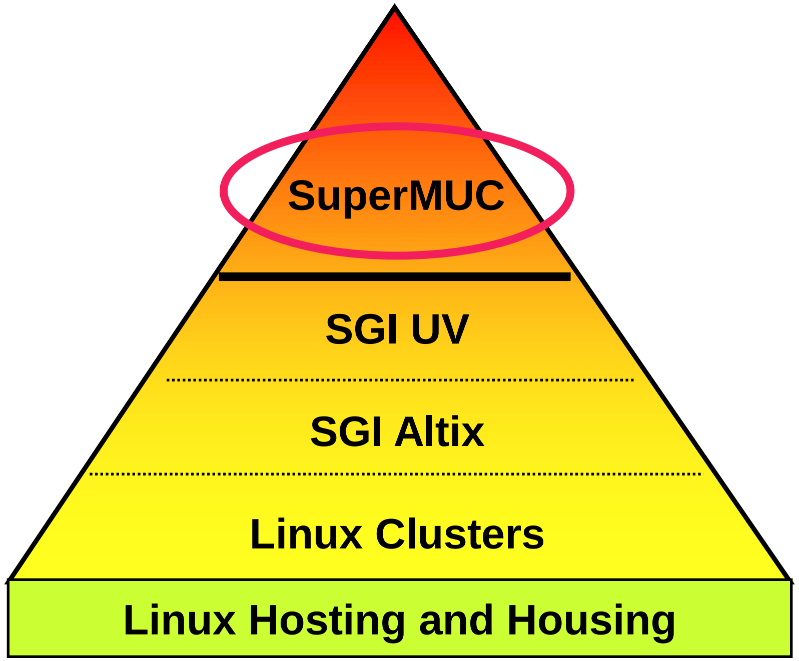
- Establishment of the legal framework
  - PRACE AISBL created with seat in Brussels in April (Association Internationale Sans But Lucratif)
  - 20 members representing 20 European countries
  - Inauguration in Barcelona on June 9

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

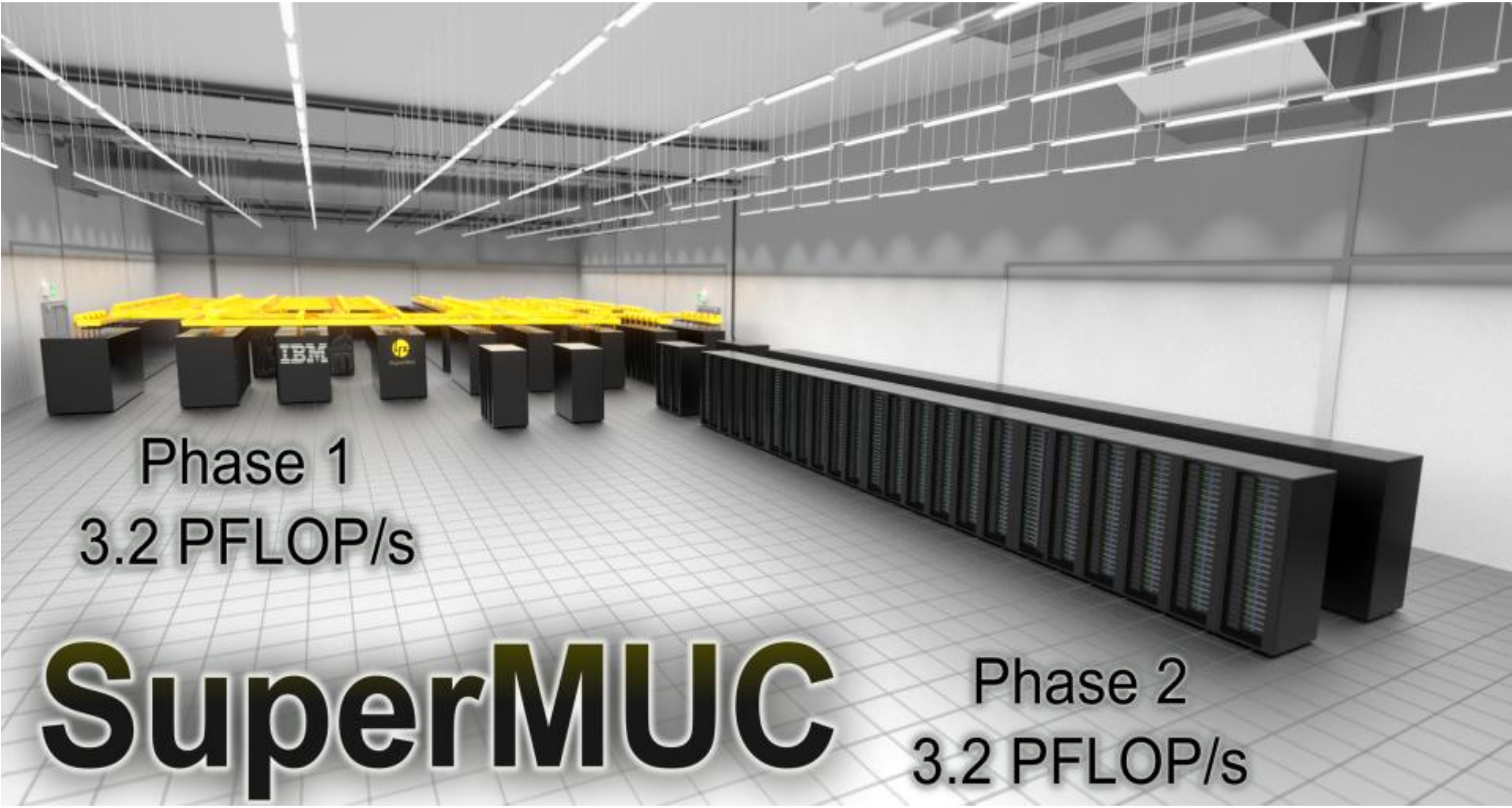
PRACE Tier-0 Systems

- **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

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

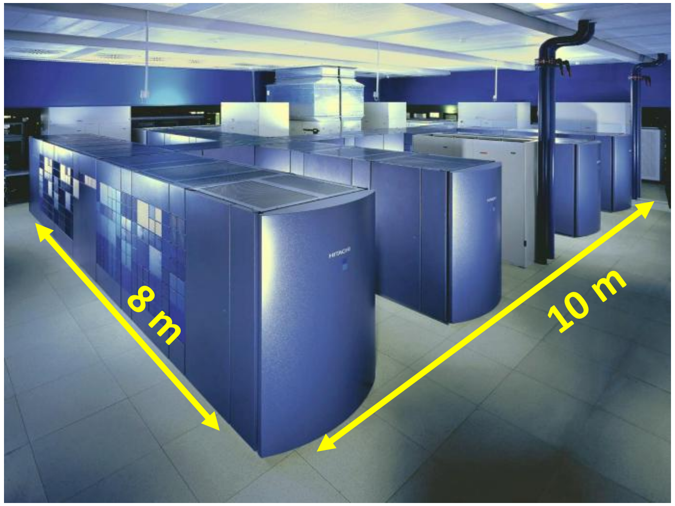


SuperMUC Phase 1 + 2

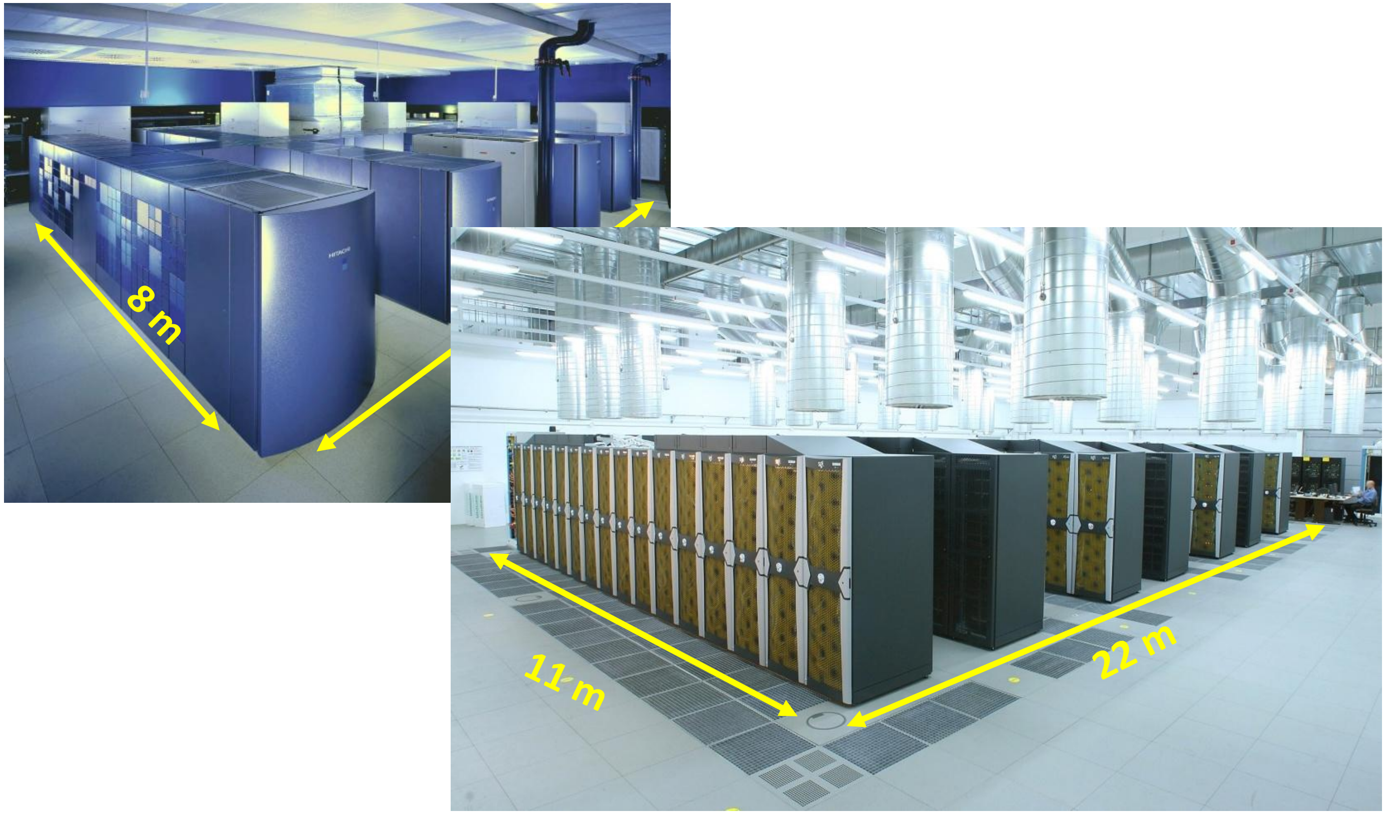


Challenges in Programming and Using these Supercomputers

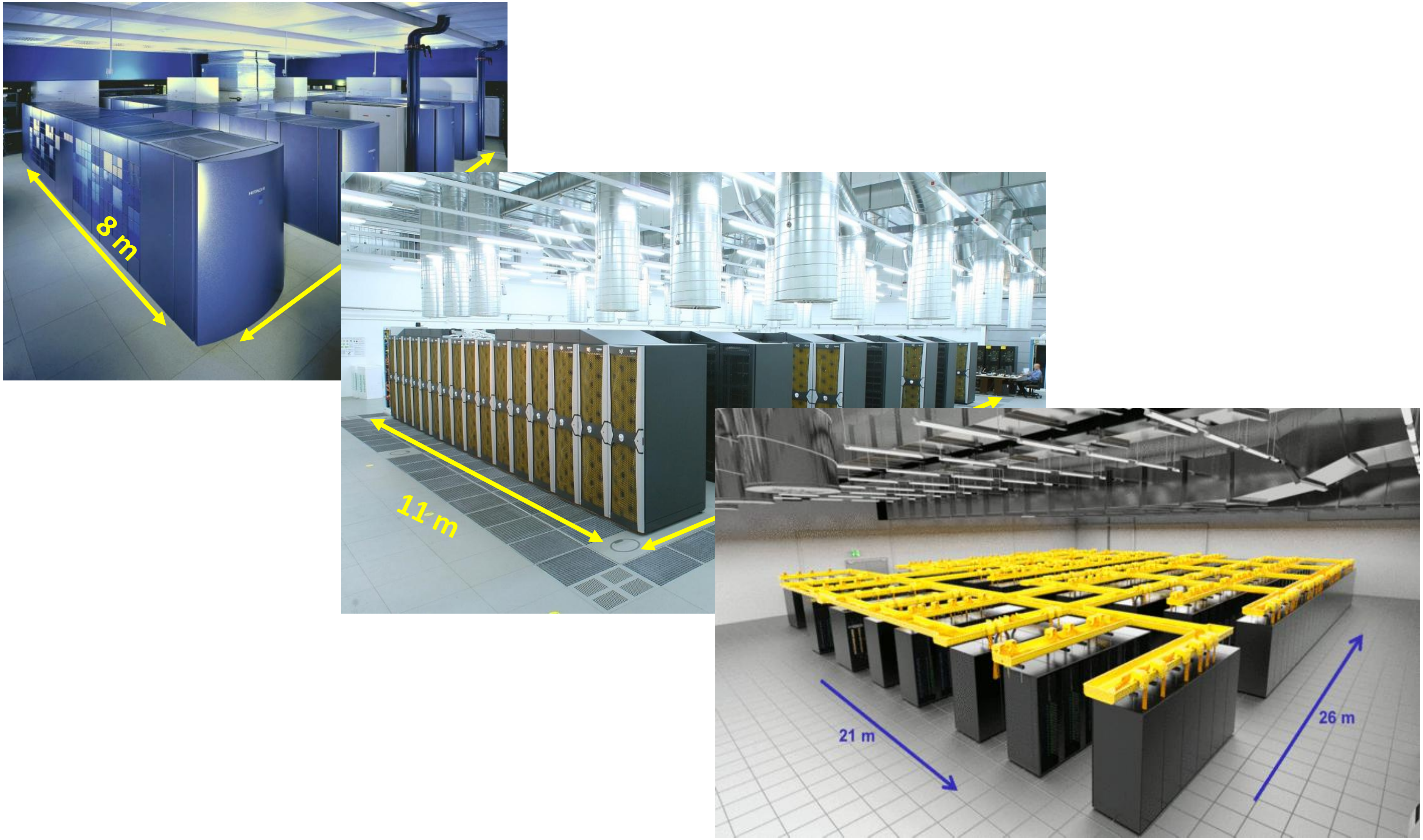
SuperMUC and its predecessors



SuperMUC and its predecessors



SuperMUC and its predecessors



LRZ Building Extension

Picture: Horst-Dieter Steinhöfer

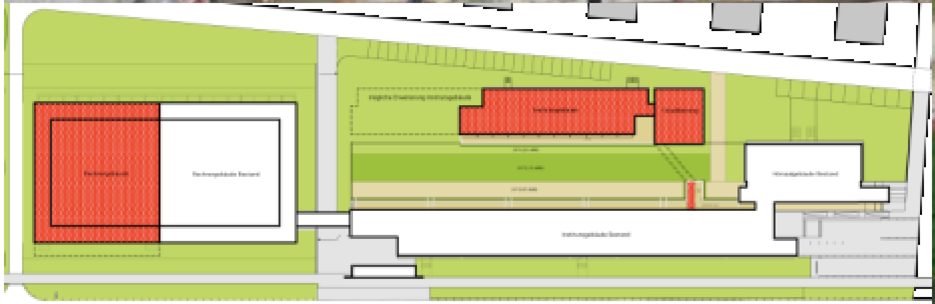
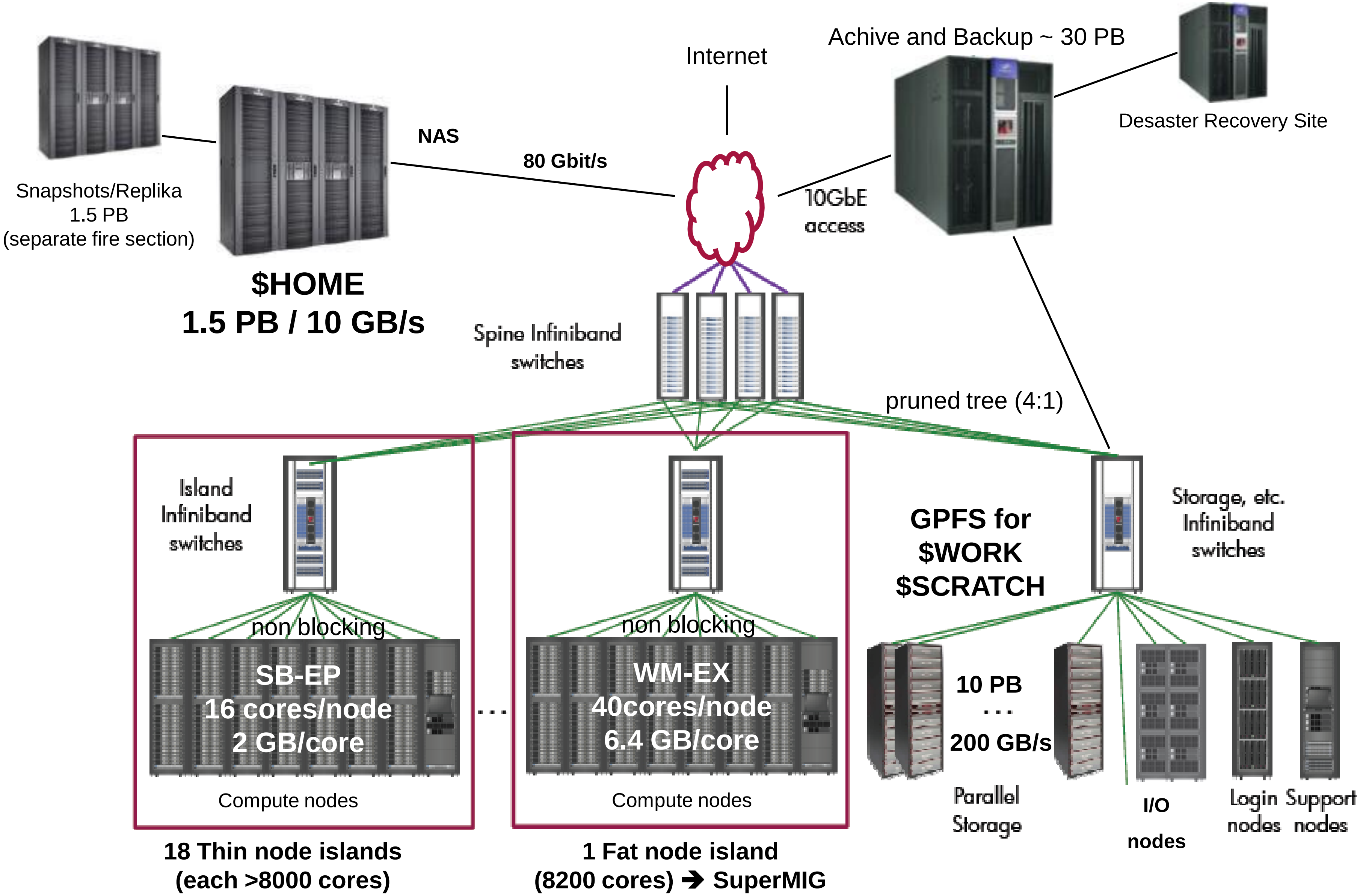


Figure: Herzog+Partner für StBAM2 (staatl. Hochbauamt München 2)

Picture: Ernst A. Graf

SuperMUC Architecture



Snapshots/Replika
1.5 PB
(separate fire section)

NAS

80 Gbit/s

Internet

Achive and Backup ~ 30 PB

Desaster Recovery Site

10GbE
access

$HOME
1.5 PB / 10 GB/s

Spine Infiniband
switches

pruned tree (4:1)

Island
Infiniband
switches

non blocking

SB-EP
16 cores/node
2 GB/core

Compute nodes

. . .

non blocking

WM-EX
40cores/node
6.4 GB/core

Compute nodes

GPFS for
$WORK
$SCRATCH

Storage, etc.
Infiniband
switches

10 PB
. . .
200 GB/s

Parallel
Storage

I/O

nodes

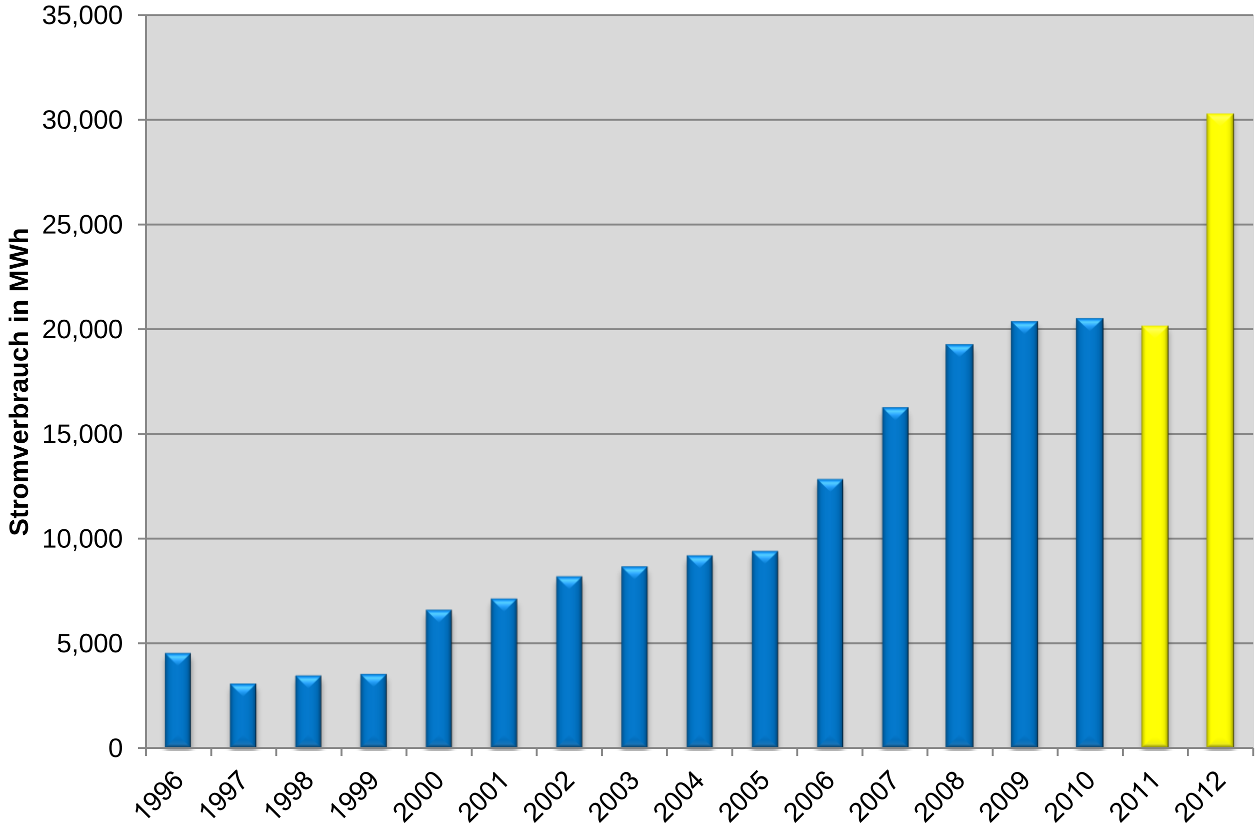
Login
nodes

Support
nodes

18 Thin node islands
(each >8000 cores)

1 Fat node island
(8200 cores) ➔ SuperMIG

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

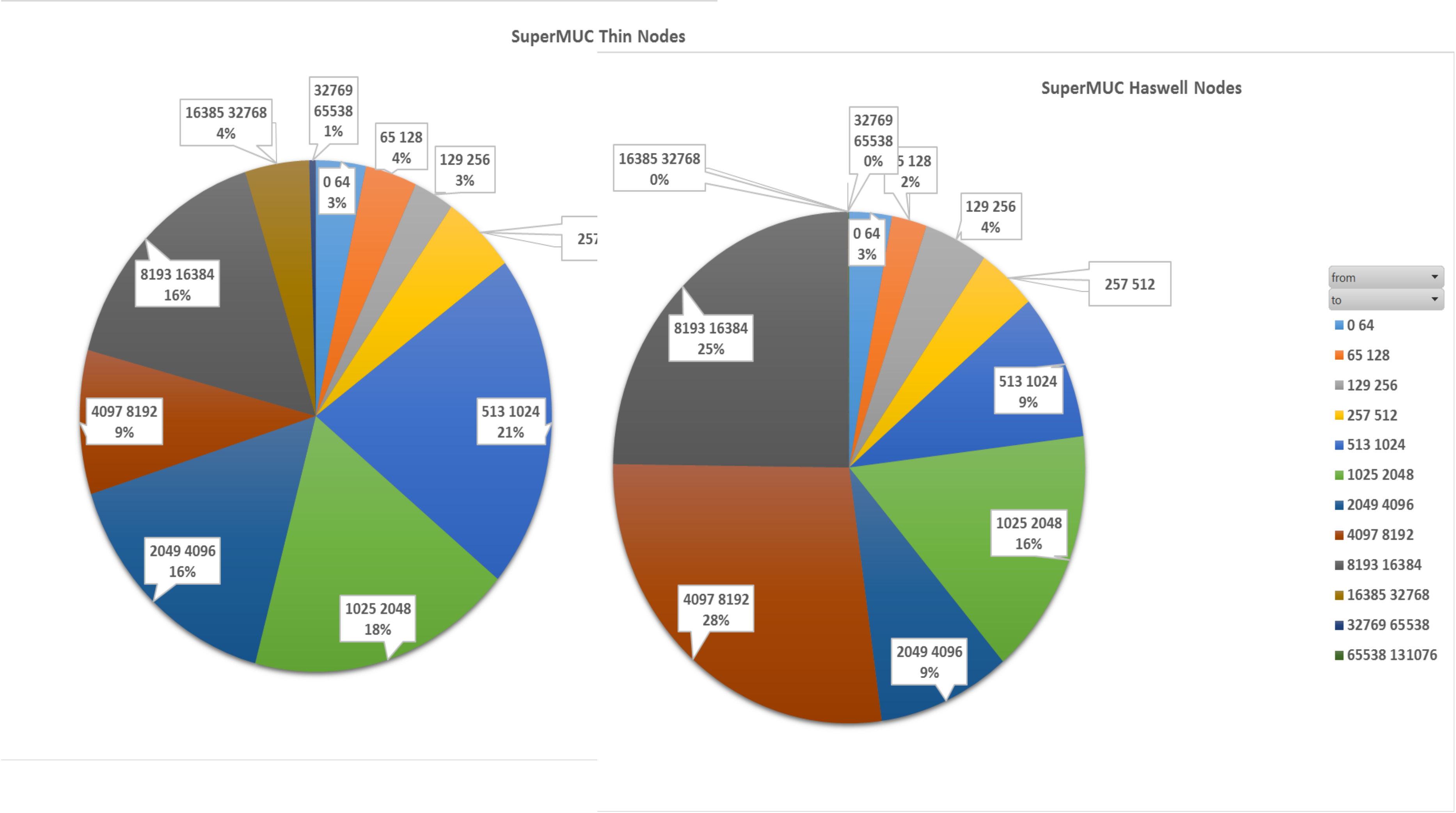
- ✓ Energy-aware scheduling
  - – 6% power advantage
  - – ~40% power advantage
  - – Total annual savings of ~2 Mio. € for SuperMUC Phase 1 and 2

Slide: Herbert Huber

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

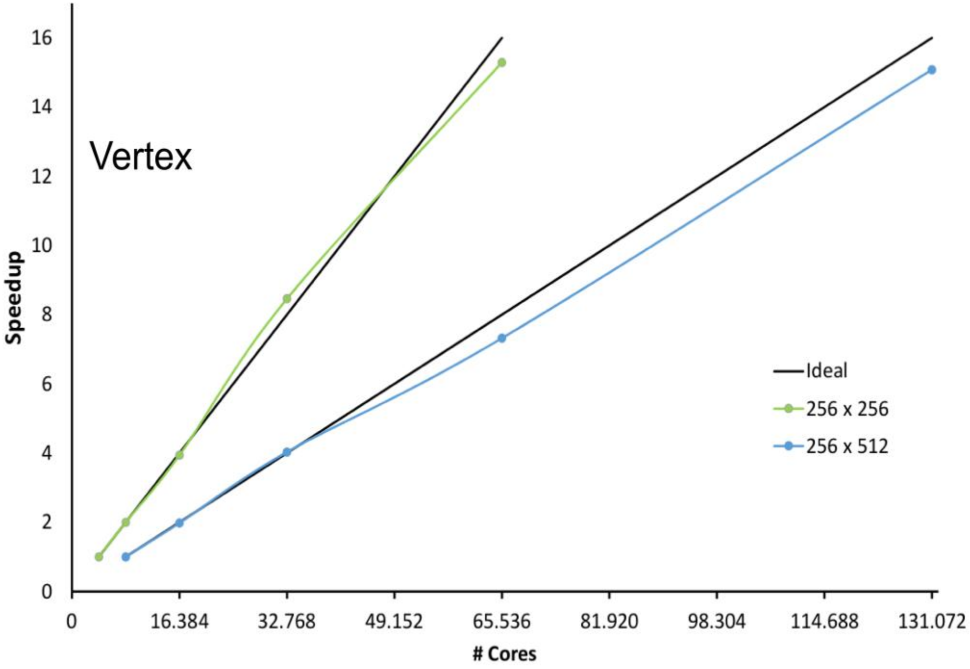
Results (Sustained TFlop/s on 128000 cores)

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

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)



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)
- 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):

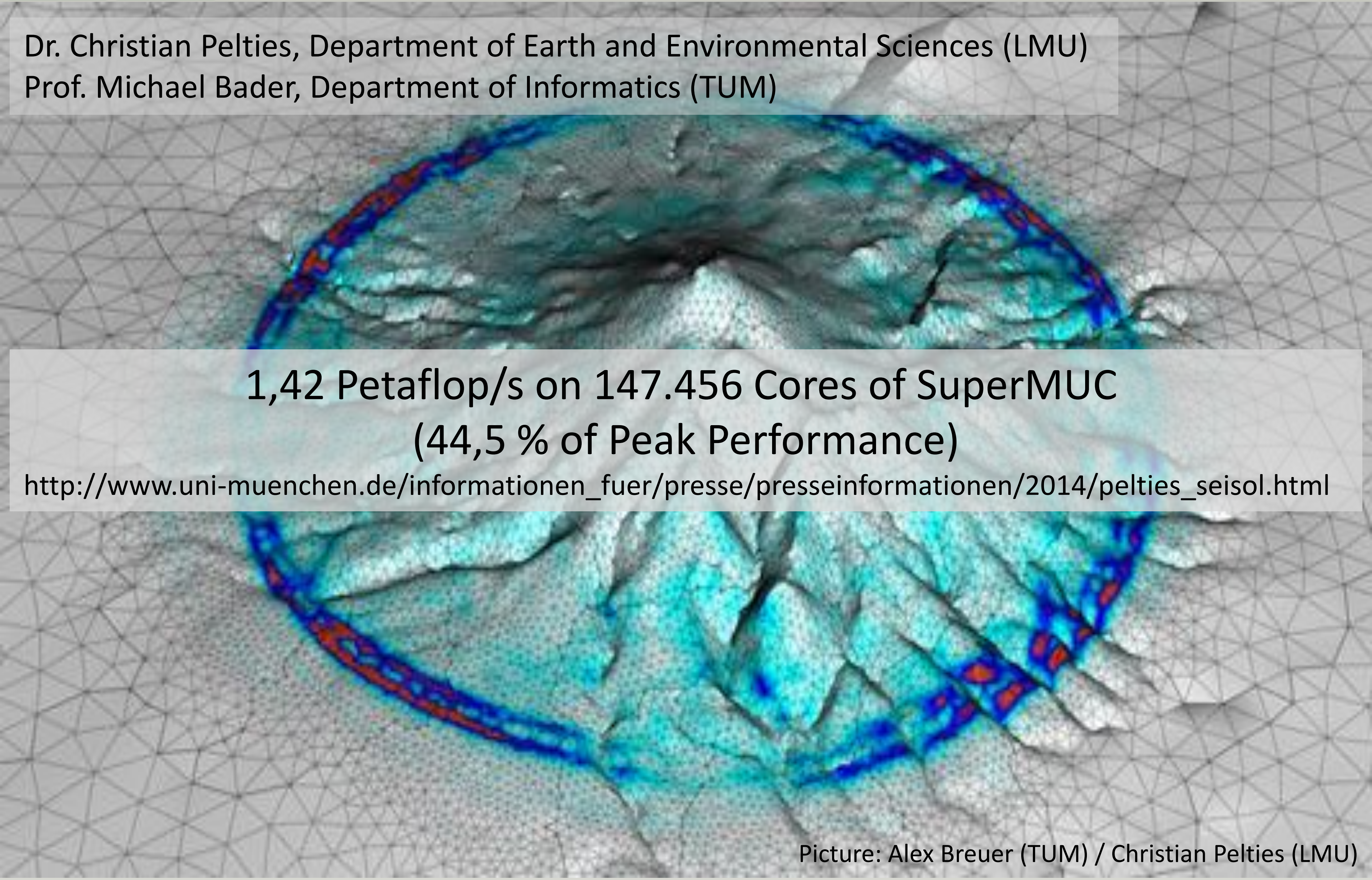
- Preparation is everything
- Finding Heisenbugs is difficult
- MPI is at its limits
- Hybrid (MPI+OpenMP) is the way to go
- I/O libraries getting even more important

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)

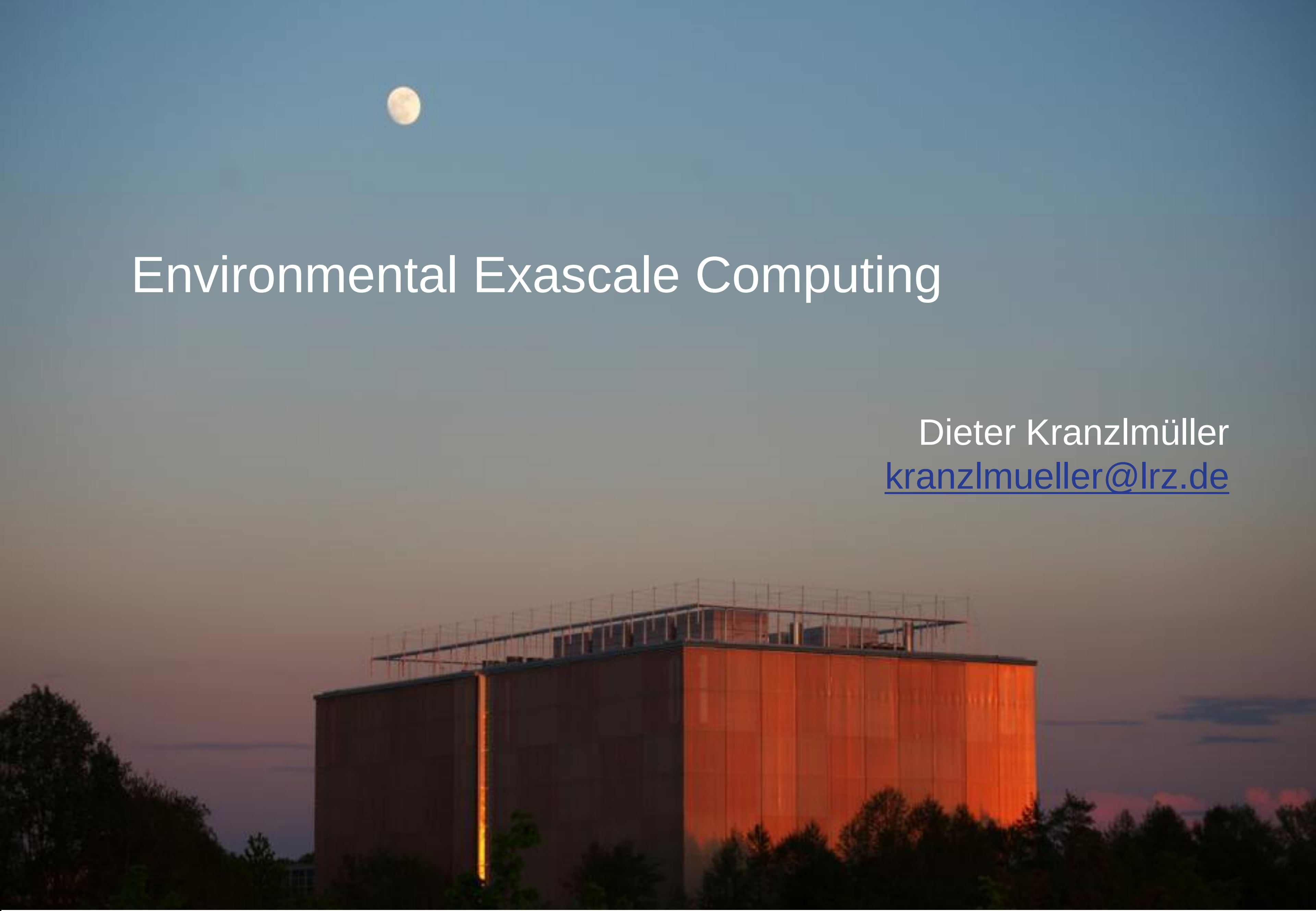
http://www.uni-muenchen.de/informationen_fuer/presse/presseinformationen/2014/pelties_seisol.html

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

Extreme Scaling - Conclusions

- 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



Environmental Exascale Computing

Dieter Kranzlmüller
kranzlmueller@lrz.de