LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MNM TEAM
MUNICH NETWORK MANAGEMENT TEAM

lrz

Extreme Scale Computing at LRZ

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

MNM TEAM
MUNICH NETWORK MANAGEMENT TEAM

lrz

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MNM TEAM
Munich Network Management Team

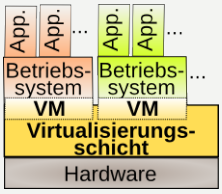
lrz

Networks

Grid computing

Cloud Computing

High Performance Computing



Virtualization

Service Management

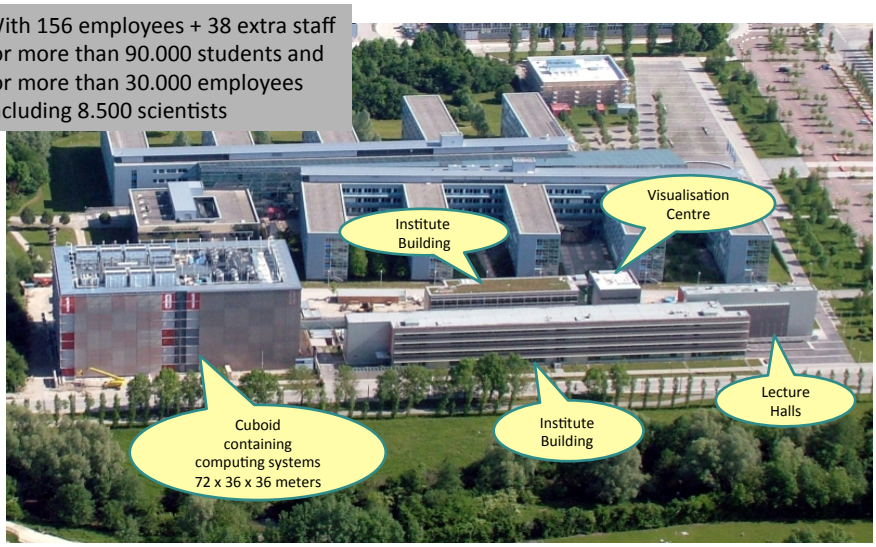
IT Security

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

lrz

With 156 employees + 38 extra staff for more than 90.000 students and for more than 30.000 employees including 8.500 scientists



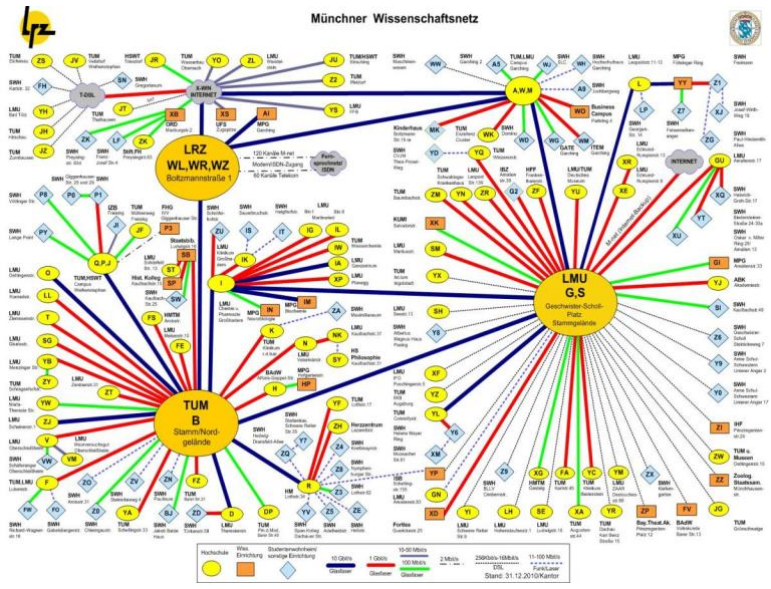
Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

- Computer Centre for all Munich Universities

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, scientifc computing, …)
- Courses (text editing, image processing, UNIX, Linux, HPC, …)

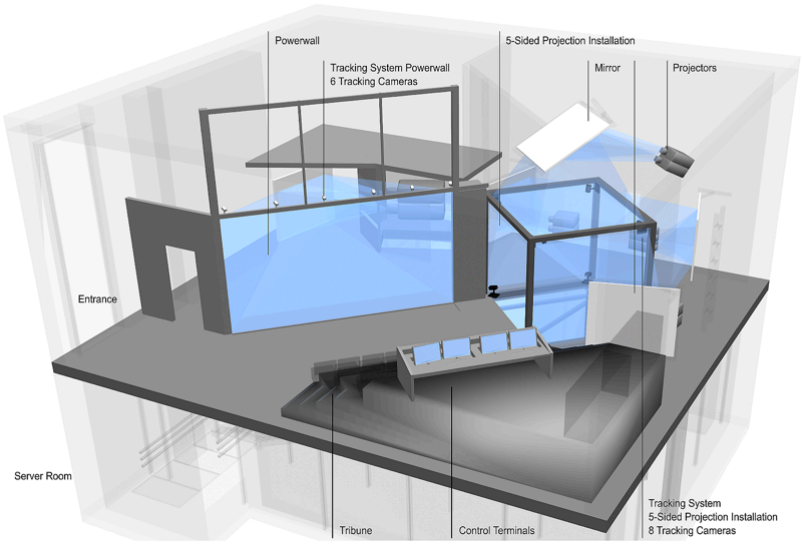
The Munich Scientific Network (MWN)



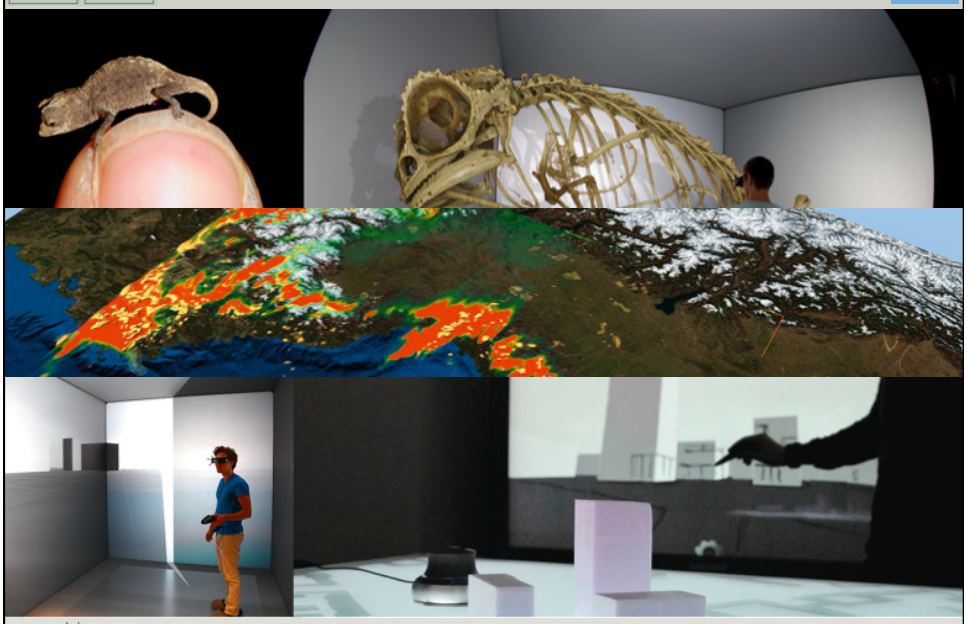
Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

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

Virtual Reality & Visualization Centre (LRZ)



Examples from the V2C



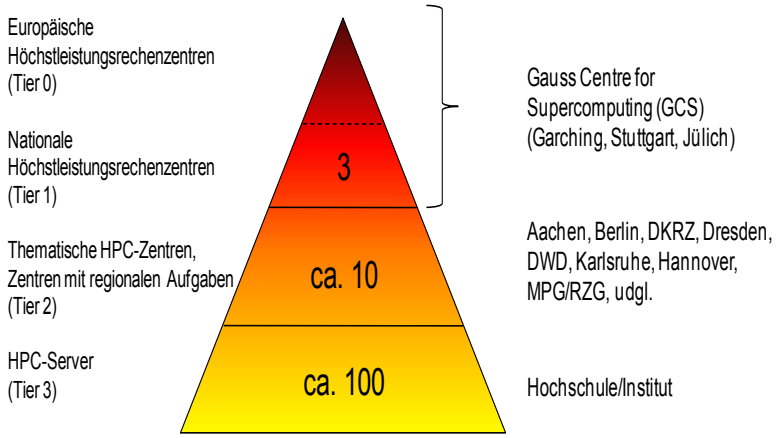
Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

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

Gauss Centre for Supercomputing (GCS)

- 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)

National HPC-Strategy



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

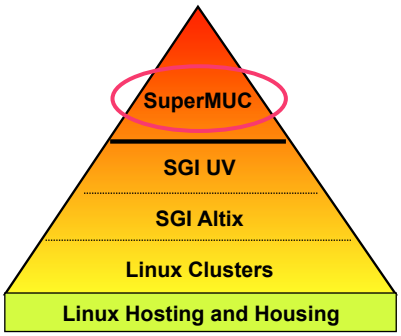
- **Curie @ GENCI**:
  Bull Cluster, 1.7 PFlop/s
- **FERMI @ CINECA**:
  IBM BG/Q, 2.1 PFlop/s
- **Hermit @ HLRS**:
  Cray XE6, 1 Pflop/s
- **JUQUEEN @ FZJ**:
  IBM Blue Gene/Q, 5.9 PFlop/s
- **MareNostrum @ BSC**:
  IBM System X iDataPlex, 1 PFlop/s
- **SuperMUC @ LRZ**:
  IBM System X iDataPlex, 3.2 PFlop/s

PRACE Tier-0 Access

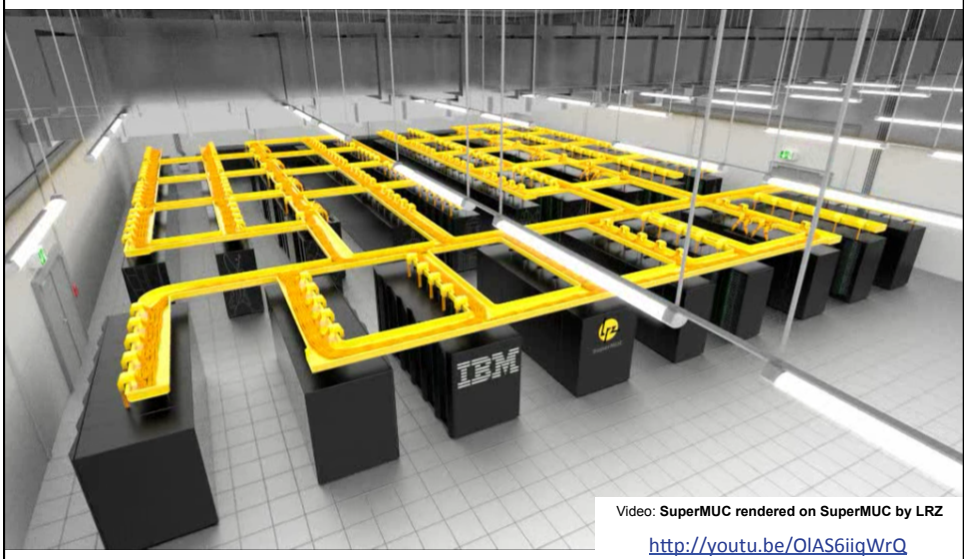
- Single pan-European Peer Review
- http://www.prace-project.eu/Call-Announcements?lang=en
- Early Access Call in May 2010
  - 68 proposals asked for 1870 Million Core hours
  - 10 projects granted with 328 Million Core hours
  - Principal Investigators from D (5), UK (2) NL (1), I (1), PT (1)
  - Involves researchers from 31 institutions in 12 countries
- Further calls being scheduled (every 6 months)
  - Call open February > Access September same year
  - Call open September > Access March next year
- Example from 8th Regular Call closed on 15 October 2013
  - Spatially adaptive radiation-hydrodynamical simulations of reionization
  - Project leader: Dr Andreas Pawlik, Max Planck Society, GERMANY
  - Research field: Universe Sciences
  - Resource Awarded: 33,800,000 core hours on SuperMUC, Germany;

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 @ LRZ



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

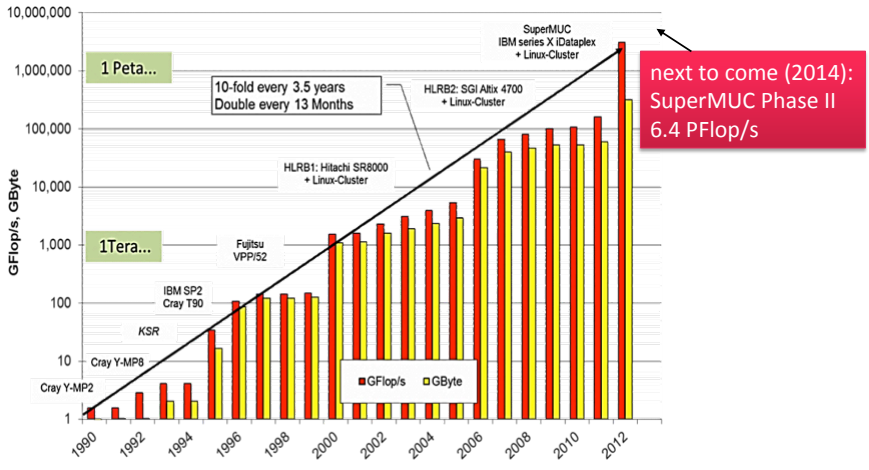
http://youtu.be/OlAS6iiqWrQ

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

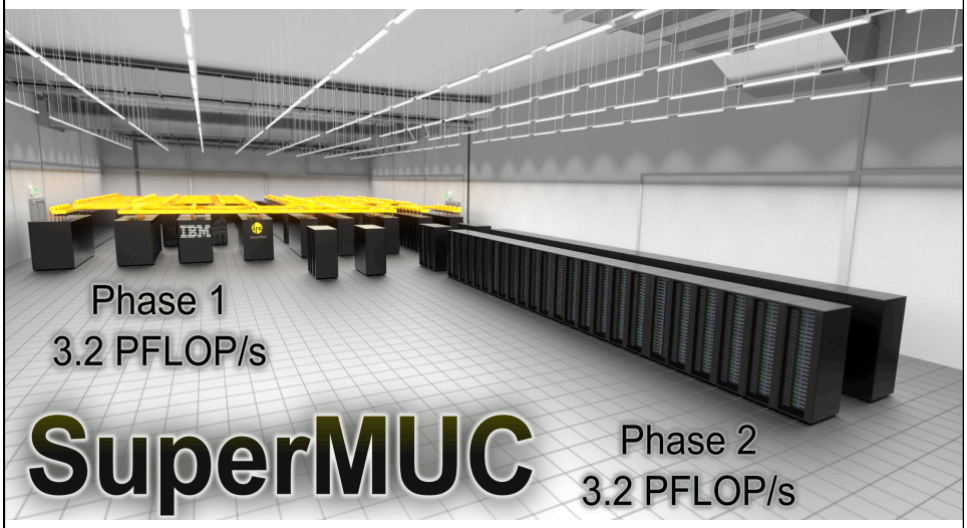
www.top500.org

LRZ Supercomputers

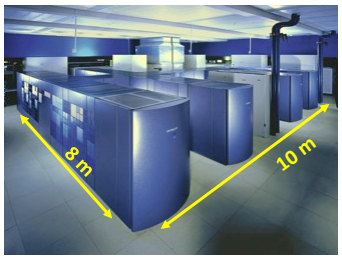


next to come (2014):
SuperMUC Phase II
6.4 PFlop/s

SuperMUC Phase 1 + 2

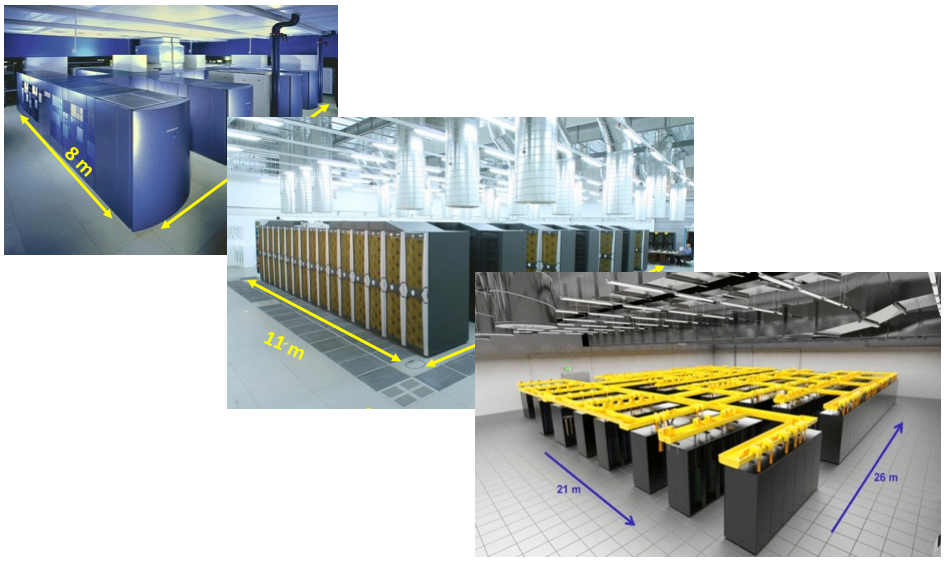


SuperMUC and its predecessors



SuperMUC and its predecessors

SuperMUC and its predecessors



LRZ Building Extension

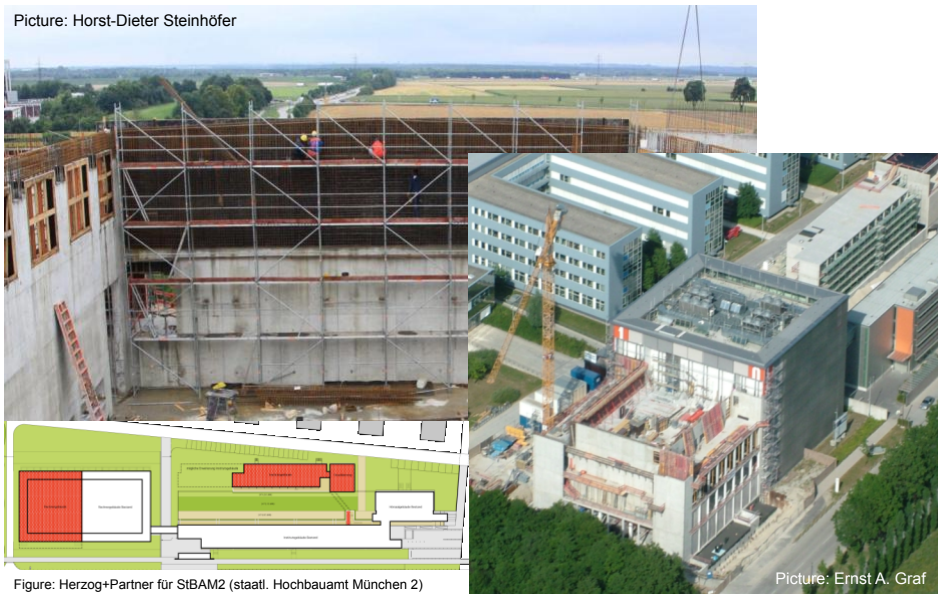
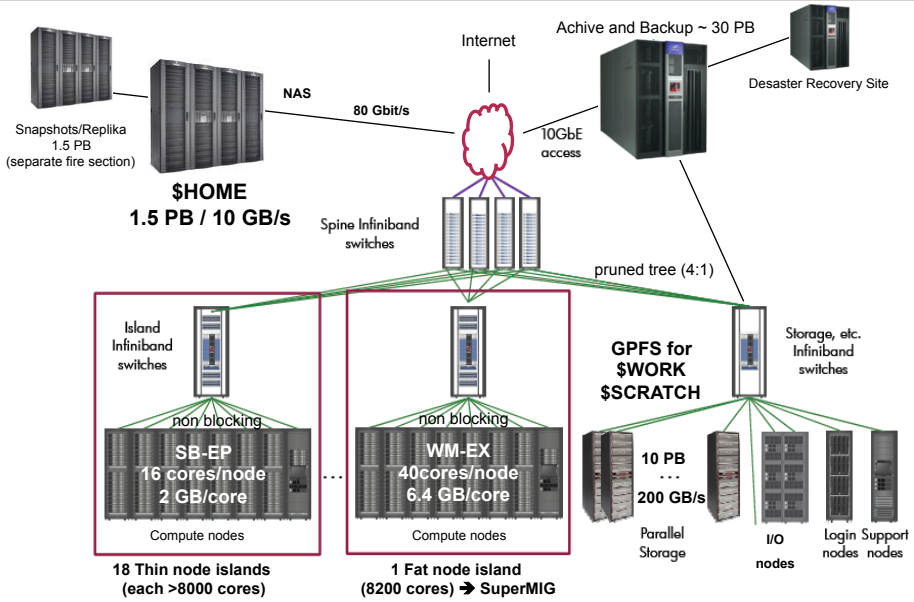


Figure: Herzog+Partner für StBAM2 (staatl. Hochbauamt München 2)

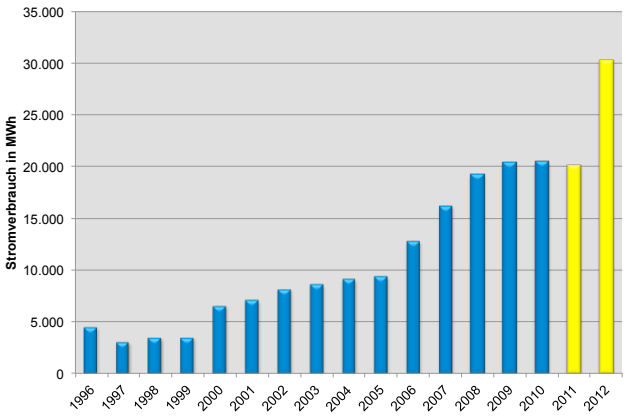
Increasing numbers

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
2014	SuperMUC Phase II	3.2 + 3.2 Pflop/s	229960

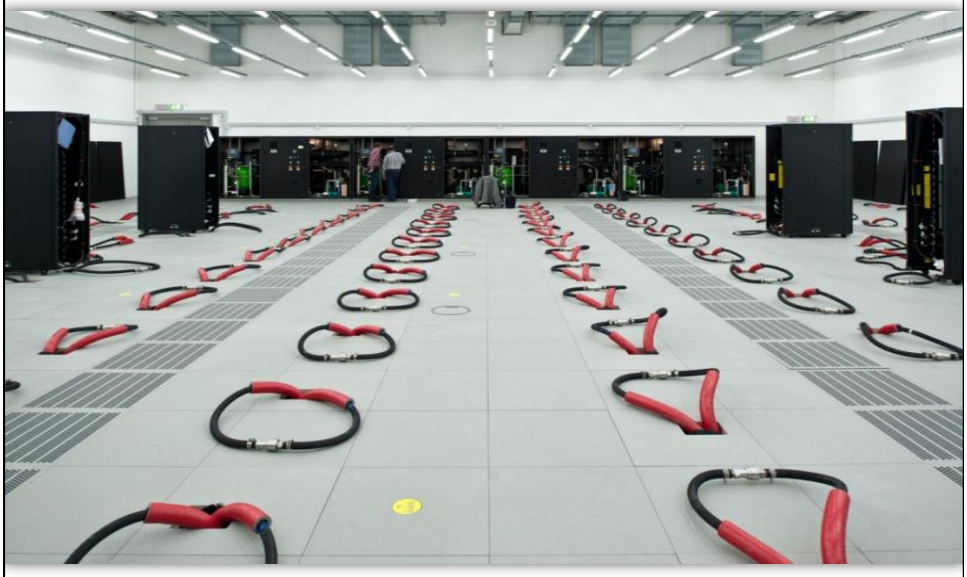
SuperMUC Architecture



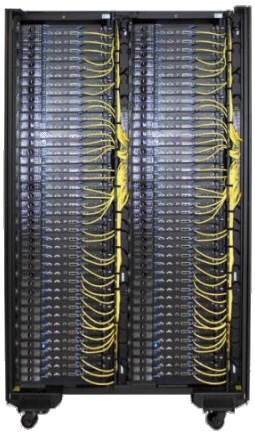
Power Consumption at LRZ



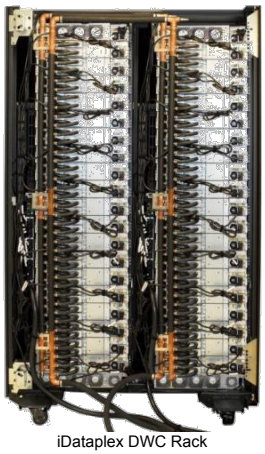
Cooling SuperMUC



IBM System x iDataPlex Direct Water Cooled Rack



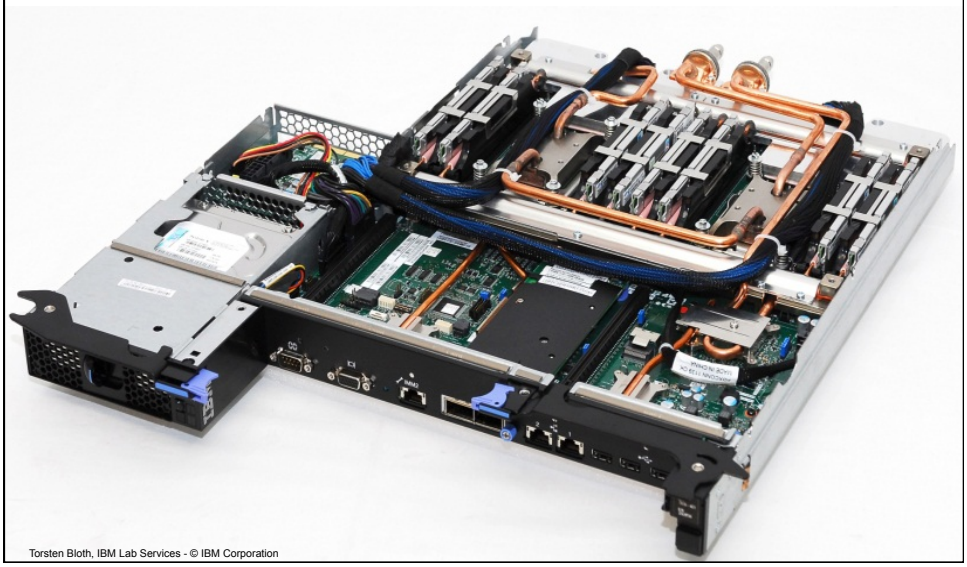
iDataplex DWC Rack
w/ water cooled nodes
(front view)



iDataplex DWC Rack
w/ water cooled nodes
(rear view of water manifolds)

Torsten Bloth, IBM Lab Services - © IBM Corporation

IBM iDataplex dx360 M4

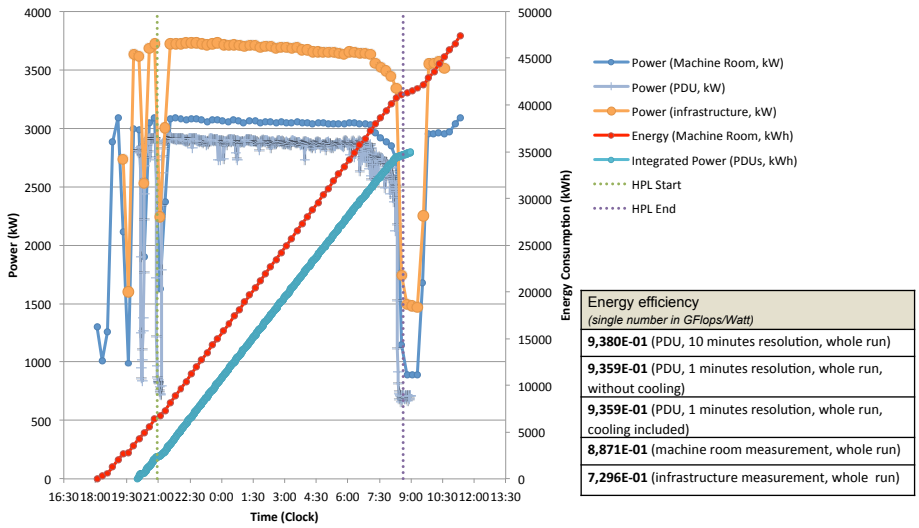


Torsten Bloth, IBM Lab Services - © IBM Corporation

Cooling Infrastructure

Photos: StBAM2 (staatl. Hochbauamt München 2)

SuperMUC HPL Energy Consumption



Energy efficiency *(single number in GFlops/Watt)*
9,380E-01 (PDU, 10 minutes resolution, whole run)
9,359E-01 (PDU, 1 minutes resolution, whole run, without cooling)
9,359E-01 (PDU, 1 minutes resolution, whole run, cooling included)
8,871E-01 (machine room measurement, whole run)
7,296E-01 (infrastructure measurement, whole run)

SuperMUC @ LRZ

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

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

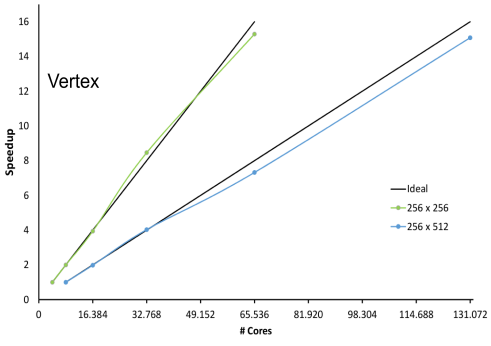
- Regular SuperMUC operation
  - 4 Islands maximum
  - Batch scheduling system
- Entire SuperMUC reserved 2,5 days for challenge:
  - 0,5 Days for testing
  - 2 Days for executing
  - 16 (of 19) Islands available
- Consumed computing time for all groups:
  - 1 hour of runtime = 130.000 CPU hours
  -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



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.

Next Steps

- 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)
- Initiation of the **LRZ Partnership Initiative πCS**

Partnership Initiative Computational Sciences πCS

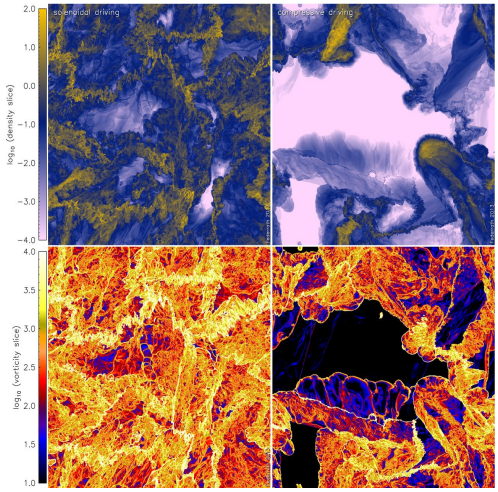
- **Individualized services** for selected scientific groups – flagship role
  - Dedicated point-of-contact
  - Individual support and guidance and targeted training &education
  - Planning dependability for use case specific optimized IT infrastructures
  - Early access to latest IT infrastructure (hard- and software) developments and specification of future requirements
  - Access to IT competence network and expertise at Computer Science and Mathematics departments
- **Partner contribution**
  - Embedding IT experts in user groups
  - Joint research projects (including funding)
  - Scientific partnership – joint publications
- **LRZ benefits**
  - Understanding the (current and future) needs and requirements of the respective scientific domain
  - Developing future services for all user groups

Partnerschaftsinitiative Computational Sciences πCS

Goals for LRZ:

- Thematic focusing – **Environmental Computing**
- Strengthening science through innovative, high performance IT technologies and modern IT infrastructures and IT services
- Interdisciplinary integration (technical and personnel) of scientists and (international) research groups
- Novel requirements and research results at the interface of scientific computing and computer-based sciences
- Increased prospects for attracting research funding through established IT expertise as contribution to application projects
- Outreach and exploitation

Astrophysics: world's largest simulations of supersonic, compressible turbulence



Slices through the three-dimensional gas density (top panels) and vorticity (bottom panels) for fully developed, highly compressible, supersonic turbulence, generated by solenoidal driving (left-hand column) and compressive driving (right-hand column), and a grid resolution of 4096^3 cells.

Federrath C MNRAS 2013;mnras.stt1644

MONTHLY NOTICES *of the* Royal Astronomical Society

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
at the Leibniz Supercomputing Centre (LRZ)
Dieter Kranzlmüller
kranzlmueller@lrz.de