

Extreme Scaling on SuperMUC

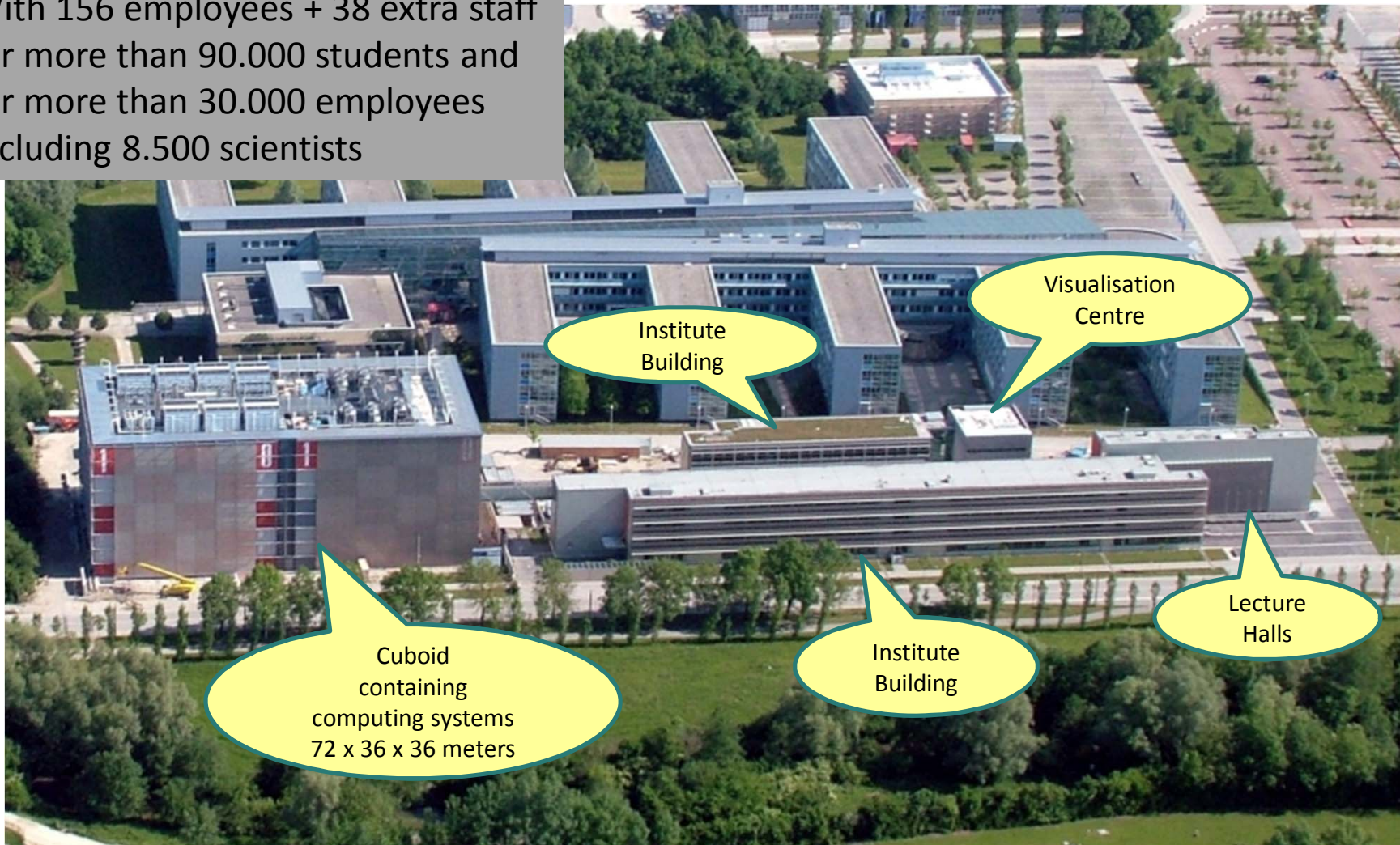
Dieter Kranzlmüller

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 90.000 students and
for more than 30.000 employees
including 8.500 scientists



- 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)

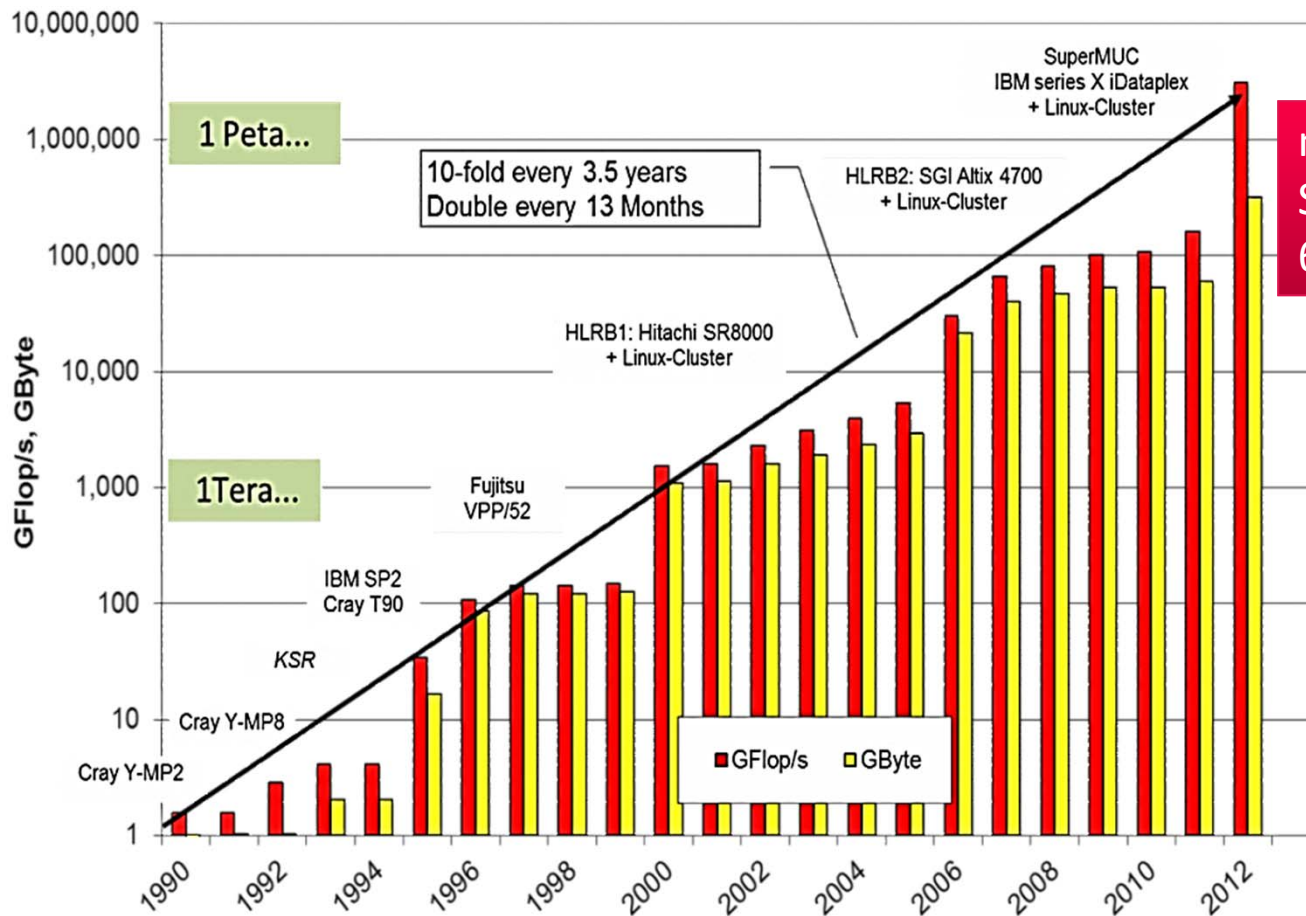




Video: **SuperMUC** rendered on SuperMUC by LRZ

<http://youtu.be/OIAS6iiqWrQ>

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



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

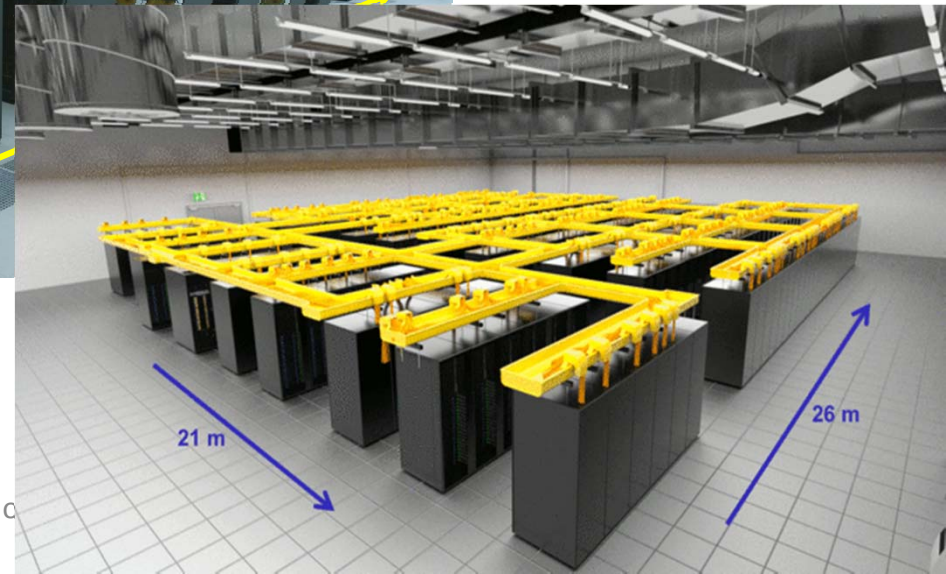
SuperMUC and its predecessors



SuperMUC and its predecessors



SuperMUC and its predecessors



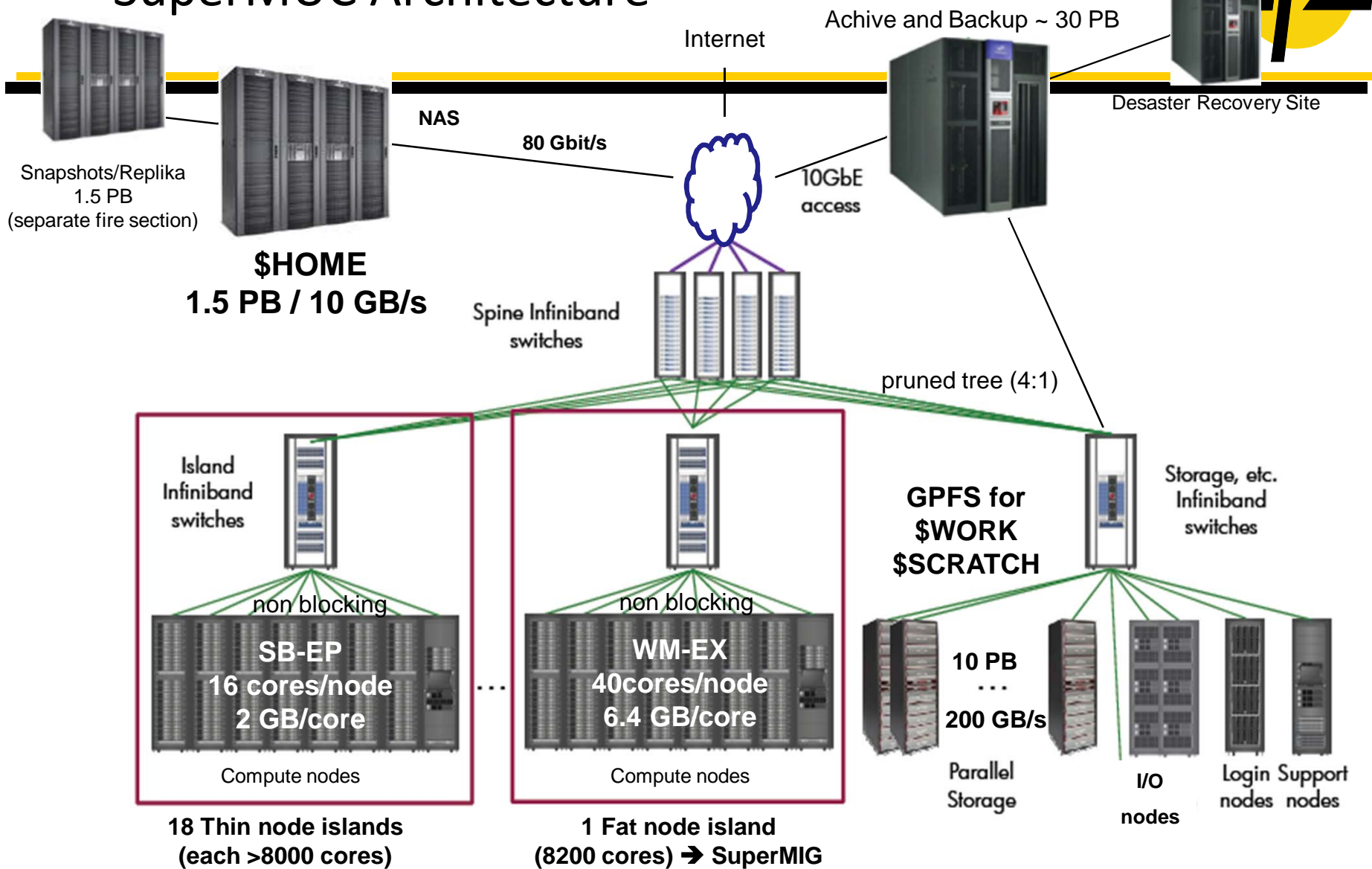
Increasing numbers



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
2014	SuperMUC Phase II	3.2 + 3.2 Pflop/s	229960



SuperMUC Architecture



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**



□ July 2013:

First SuperMUC 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

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
Linpac	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

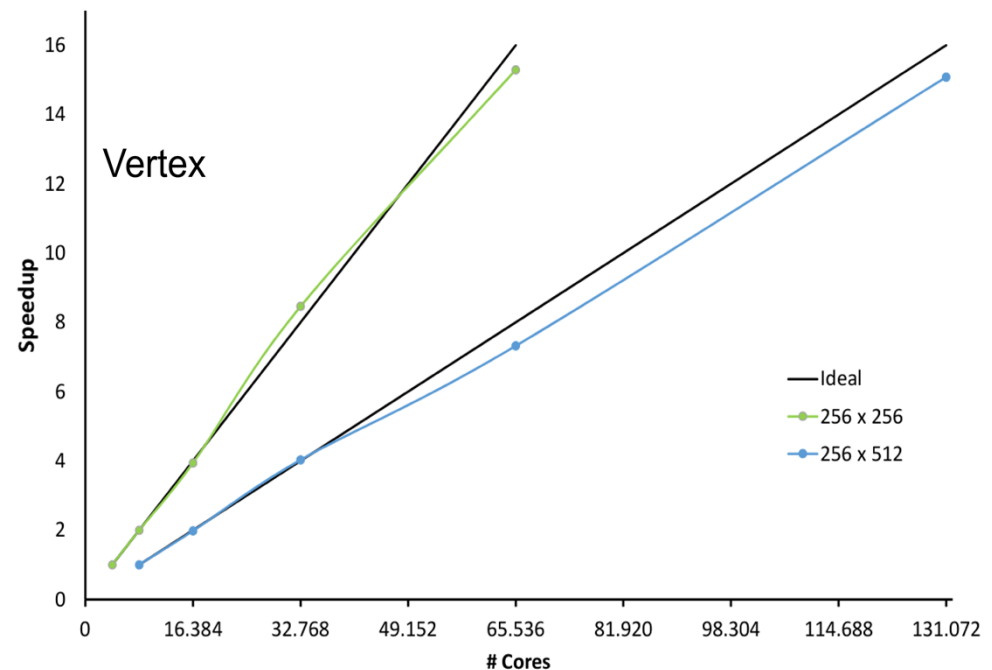
Results



❑ 5 Software packages were running on max 16 islands:

- LAMMPS
- VERTEX
- GADGET
- WaLBerla
- BQCD

❑ VERTEX reached 245 TFlop/s on 16 islands (A. Marek)



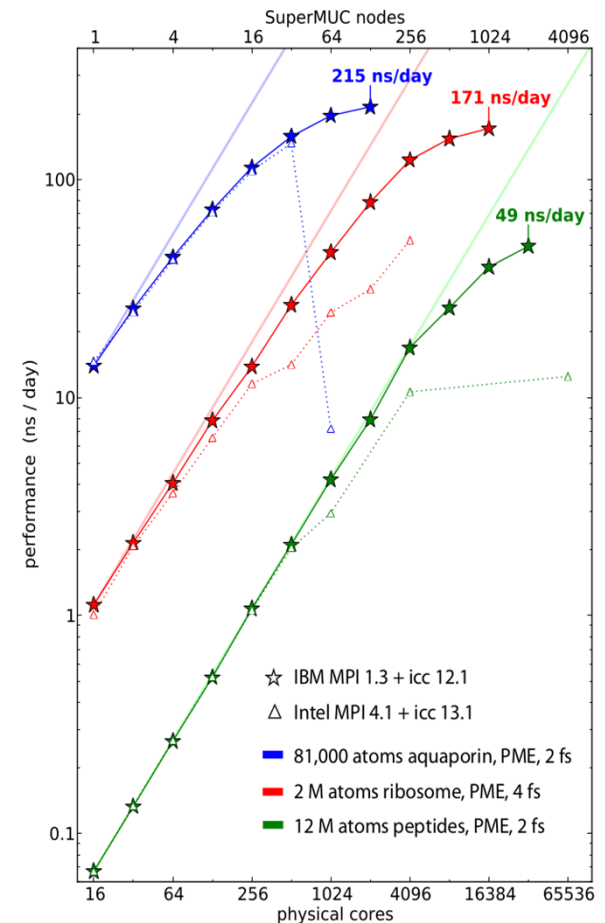
Results



4 Software packages were running on max 8 islands:

- Gromacs
- APES
- SeisSol
- CIAO

GROMACS reached 201 TFlop/s on 8 islands (C. Kutzner)



Lessons learned



- ❑ **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.**

LRZ Extreme Scale Suite (LESS)



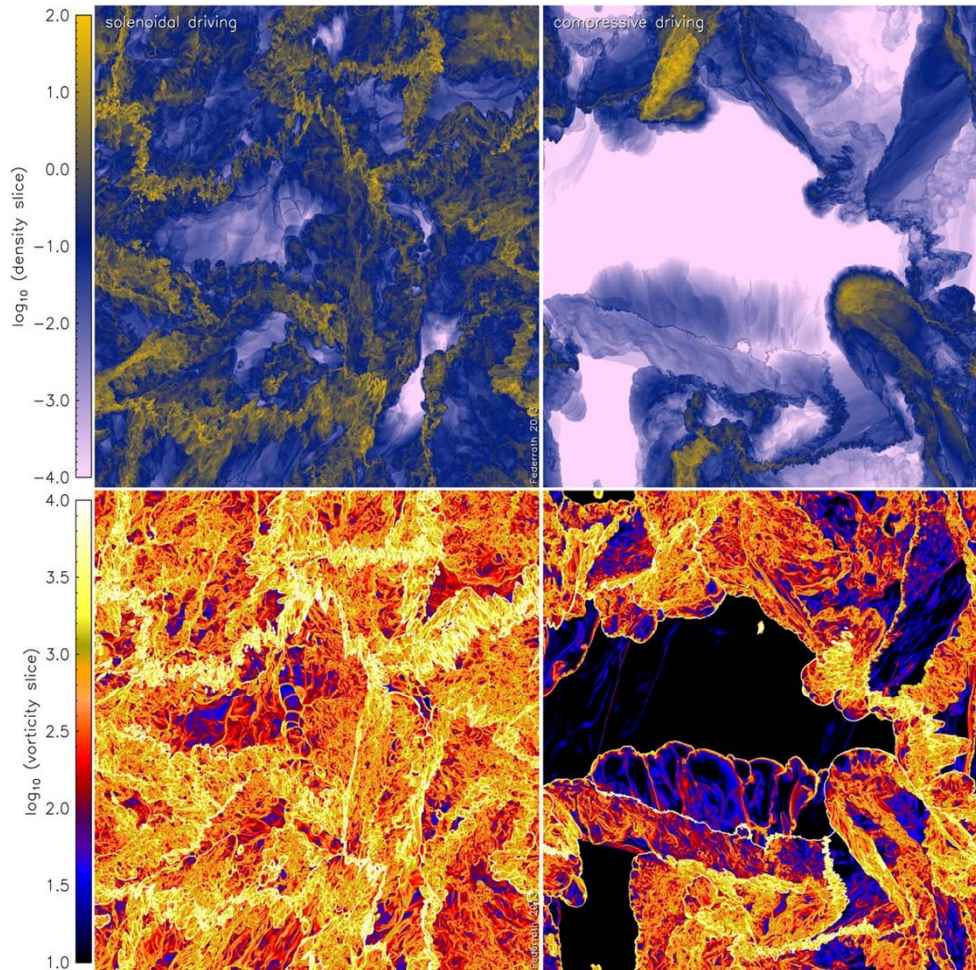
- ❑ **Platform and architecture agnostic framework for automatic compilation and submission of validation and test jobs**
- ❑ **Framework source originates from DEISA benchmark suite**
- ❑ **Extended in EU Project Scalalife to Gromacs, Dalton and Discrete**
- ❑ **Implemented in XML and perl**
- ❑ **System architecture is described as XML file**
- ❑ **Many system architectures available**
 - **Hardware: SGI Altix, UV, ice, IBM dataplex, cell, CRAY, generic x86**
 - **Compilers: icc, gcc, xlc**
 - **Batch systems: PBS, SLURM, Loadleveller, ...**
- ❑ **Software Packages: BQCD, GROAMCS, Lammps, Gadget, APES, CIAO, SeisSol, GPI, pbdMPI, doRedis, Blender**

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**
- Next workshop will be June/July 2014**

Astrophysics: world's largest simulations of supersonic, compressible turbulence with a numerical grid resolution of 4096^3 points.



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 – Earthquake Simulation at Petascale



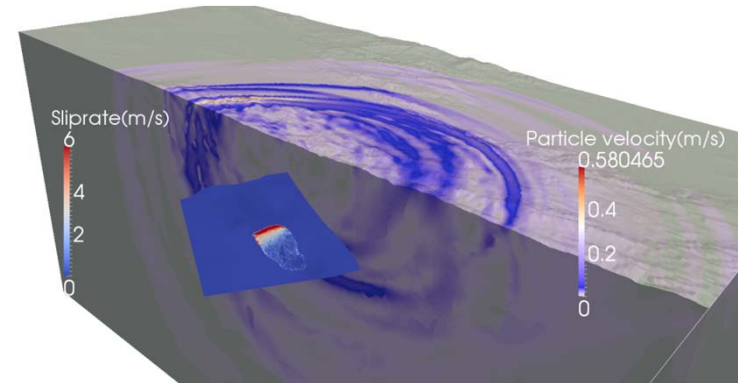
Key Features

- ❑ ADER-DG: high approximation order in space and time
- ❑ Adaptive tetrahedral meshes for highly complex geometries
- ❑ Dynamic rupture simulation coupled to seismic wave propagation

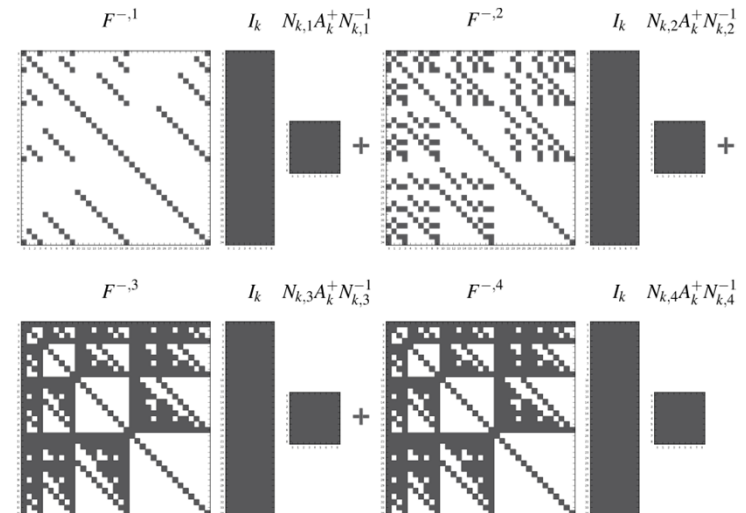
Code Generation for Matrix Kernels:

- ❑ Optimal code generation in an offline pre-compile phase
- ❑ Generation of vector instruction when the auto-vectorizer fails
- ❑ Selection of the optimal kernel (sparse or dense) for every matrix and numerical order

Check-out detailed poster @SC'13!



1994 Northridge Earthquake (A. Gabriel)



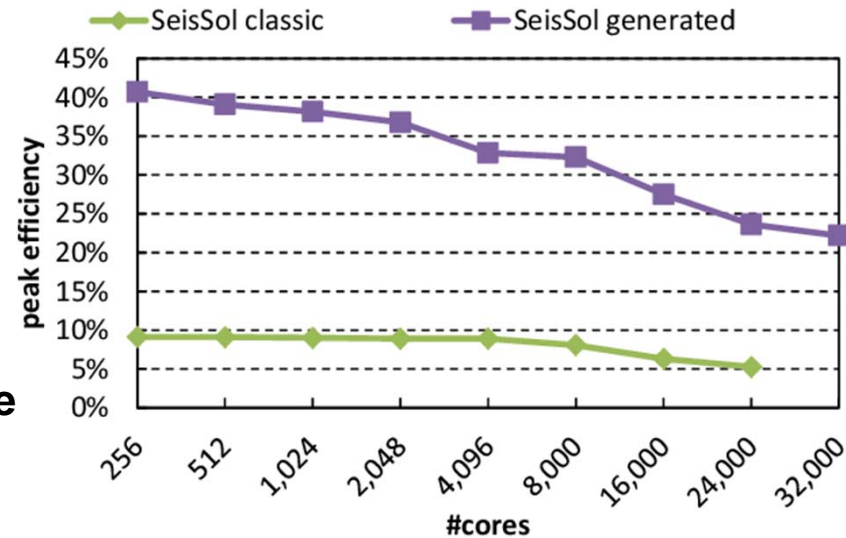
Several sparsity patterns of a 5th-order discretization

~1 PF Sustained Performance on SuperMUC



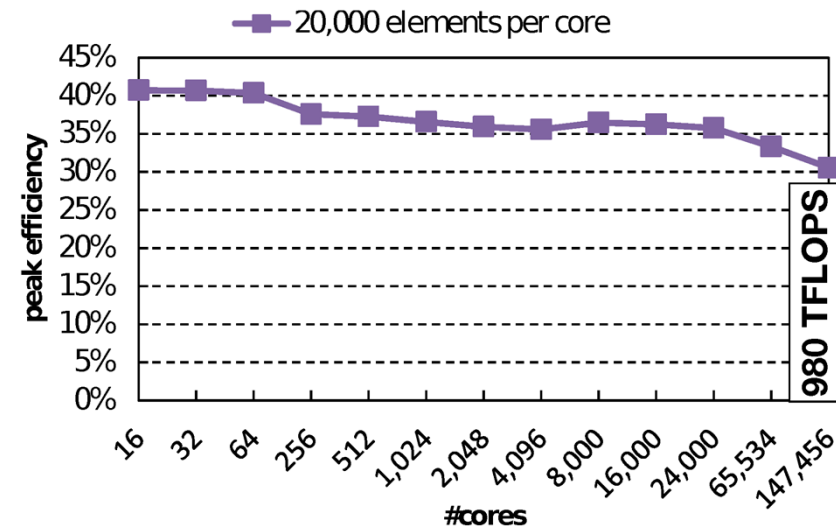
Strong Scaling Benchmark Run:

- ❑ 20,000 elements per core (recursively generated mesh on cube domain)
- ❑ 6th order (1.5 trillion unknowns) using 30% of SuperMUC's memory
- ❑ 0.98 PF, more than 30% of peak performance



Weak Scaling Study:

- SCEC LOH.1 , 7,252,482 cells
- 6th order (3.6 billion unknowns)
- 2.25 TF on 256 cores (40.6% peak)
153 TF on 32K cores (21.6% peak)
- 4.5x speedup in time to solution due to kernel generation



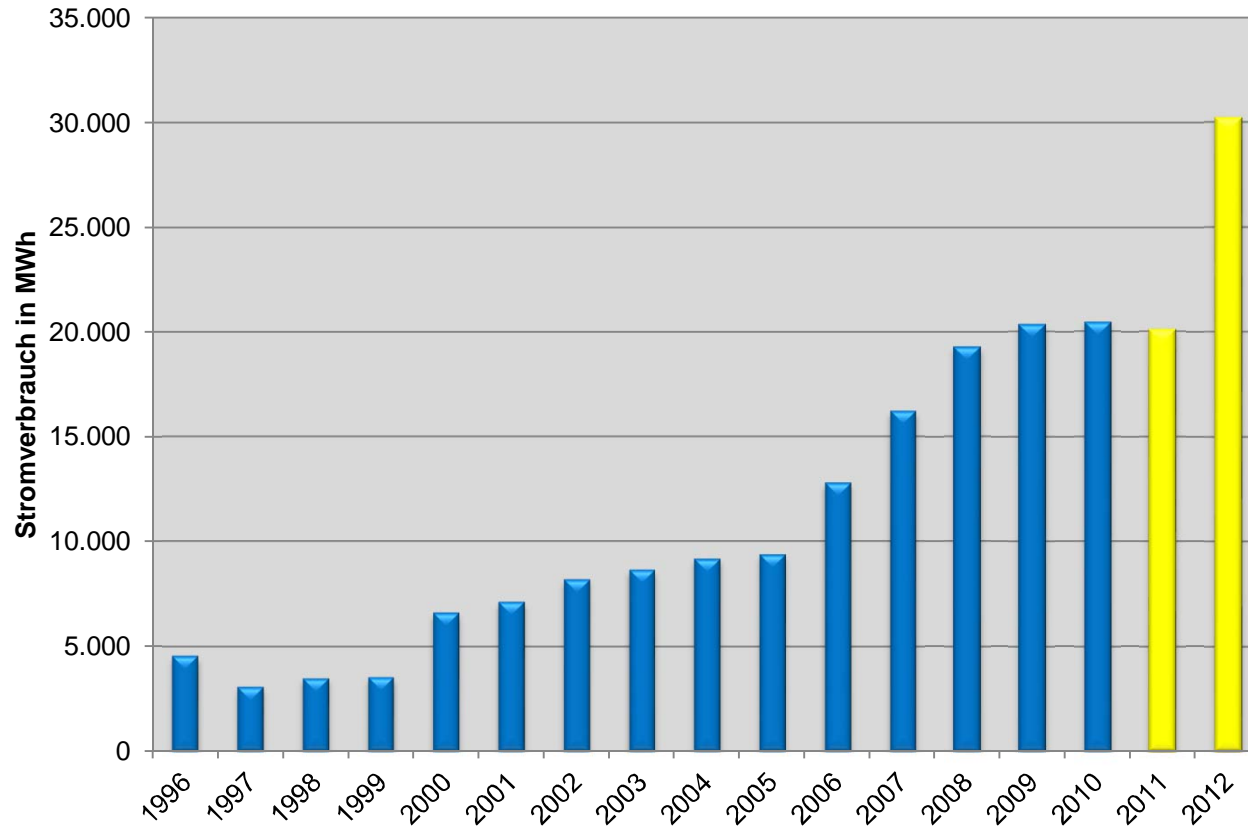
Check-out detailed poster @SC'13!

SuperMUC @ LRZ

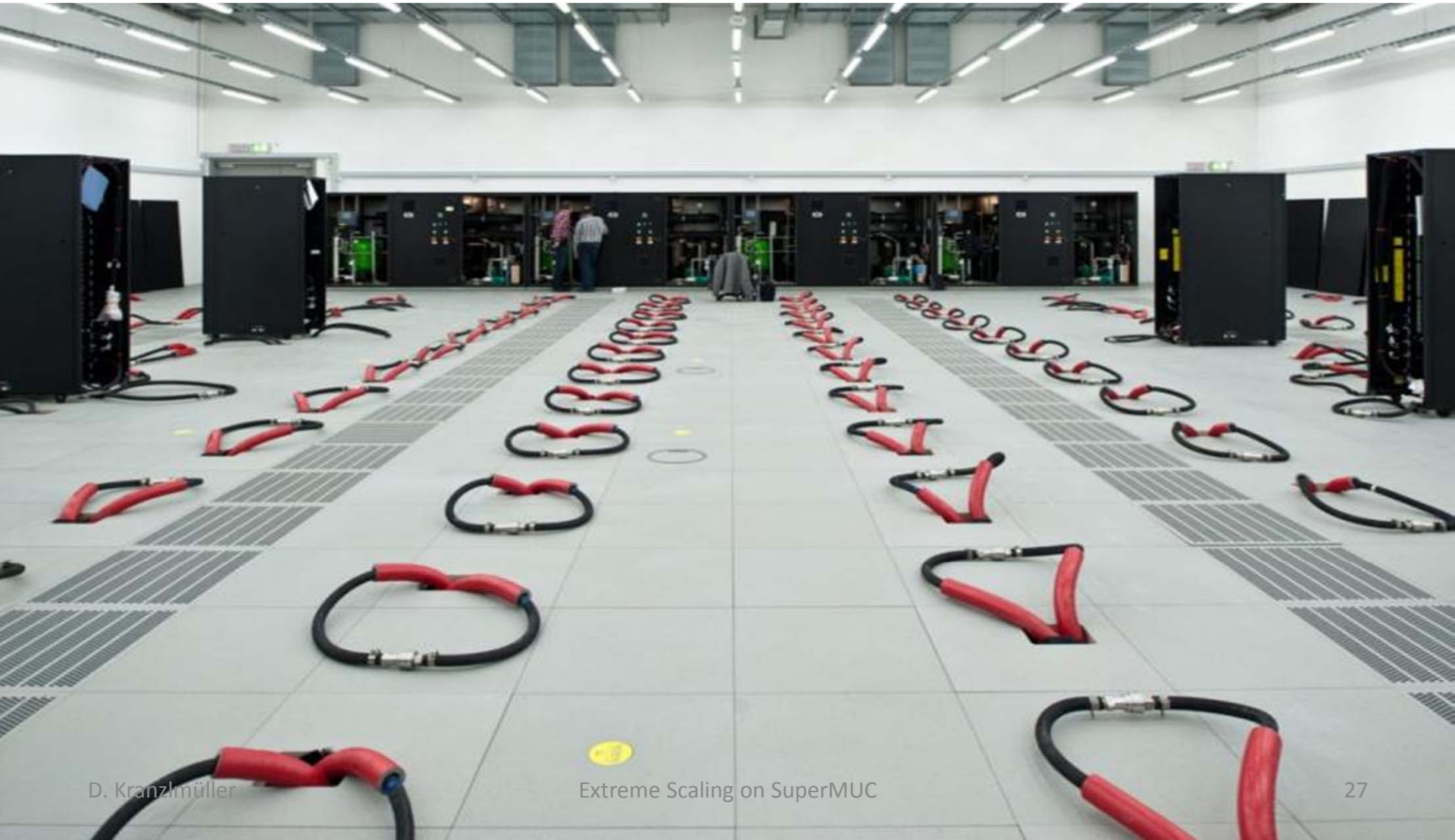


rendered on SuperMUC by LRZ

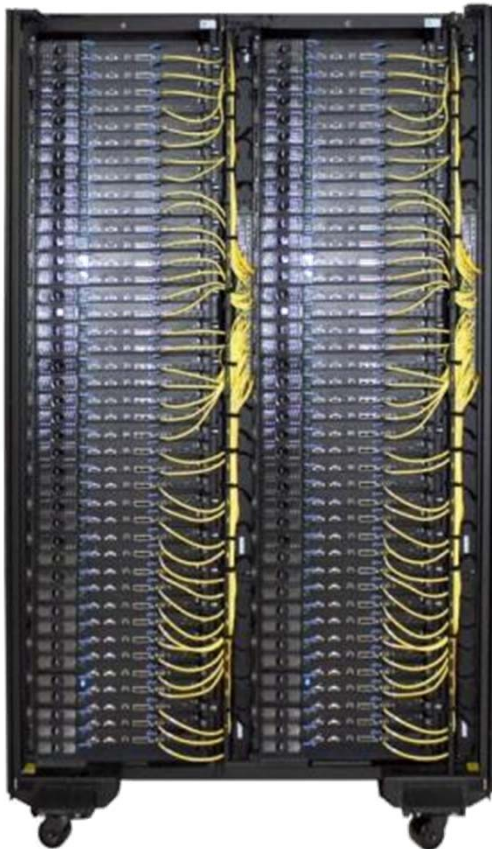
Power Consumption at LRZ



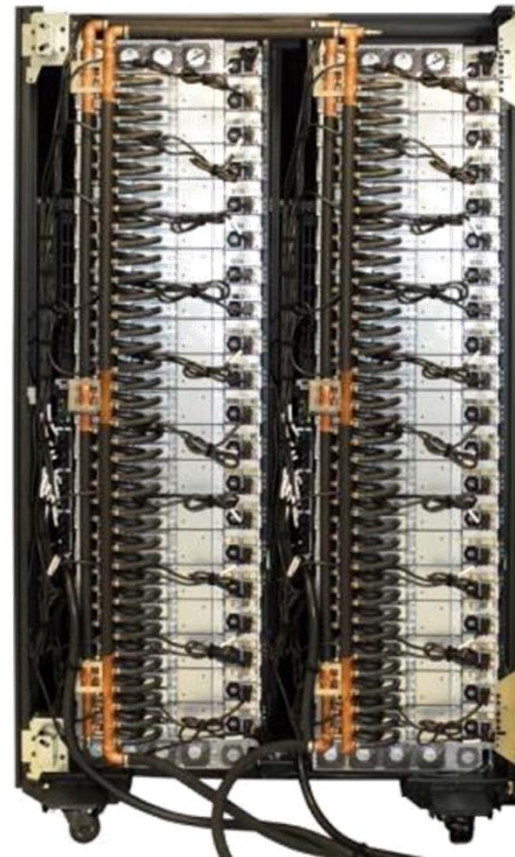
Cooling SuperMUC



IBM System x iDataPlex Direct Water Cooled Rack

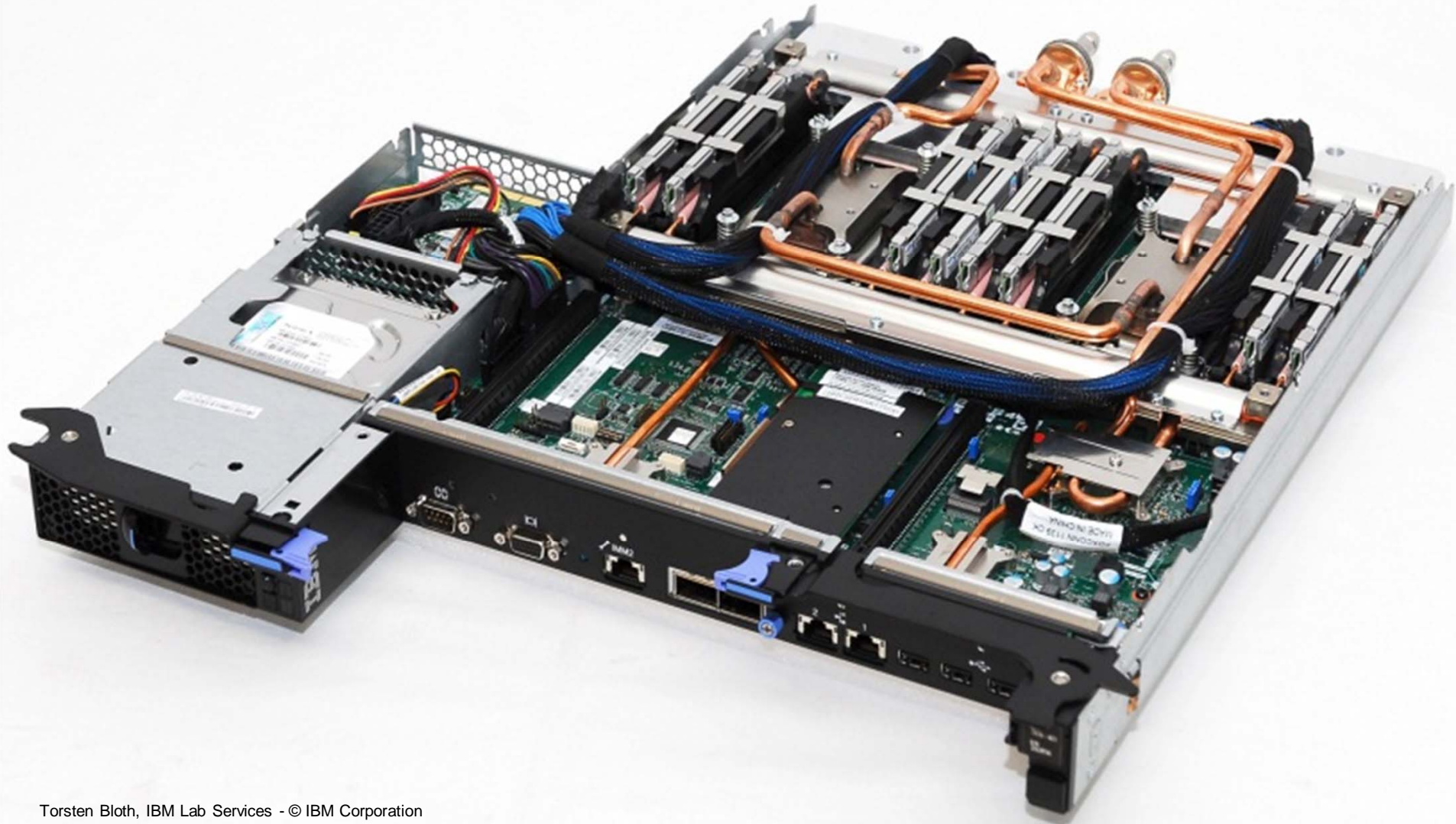


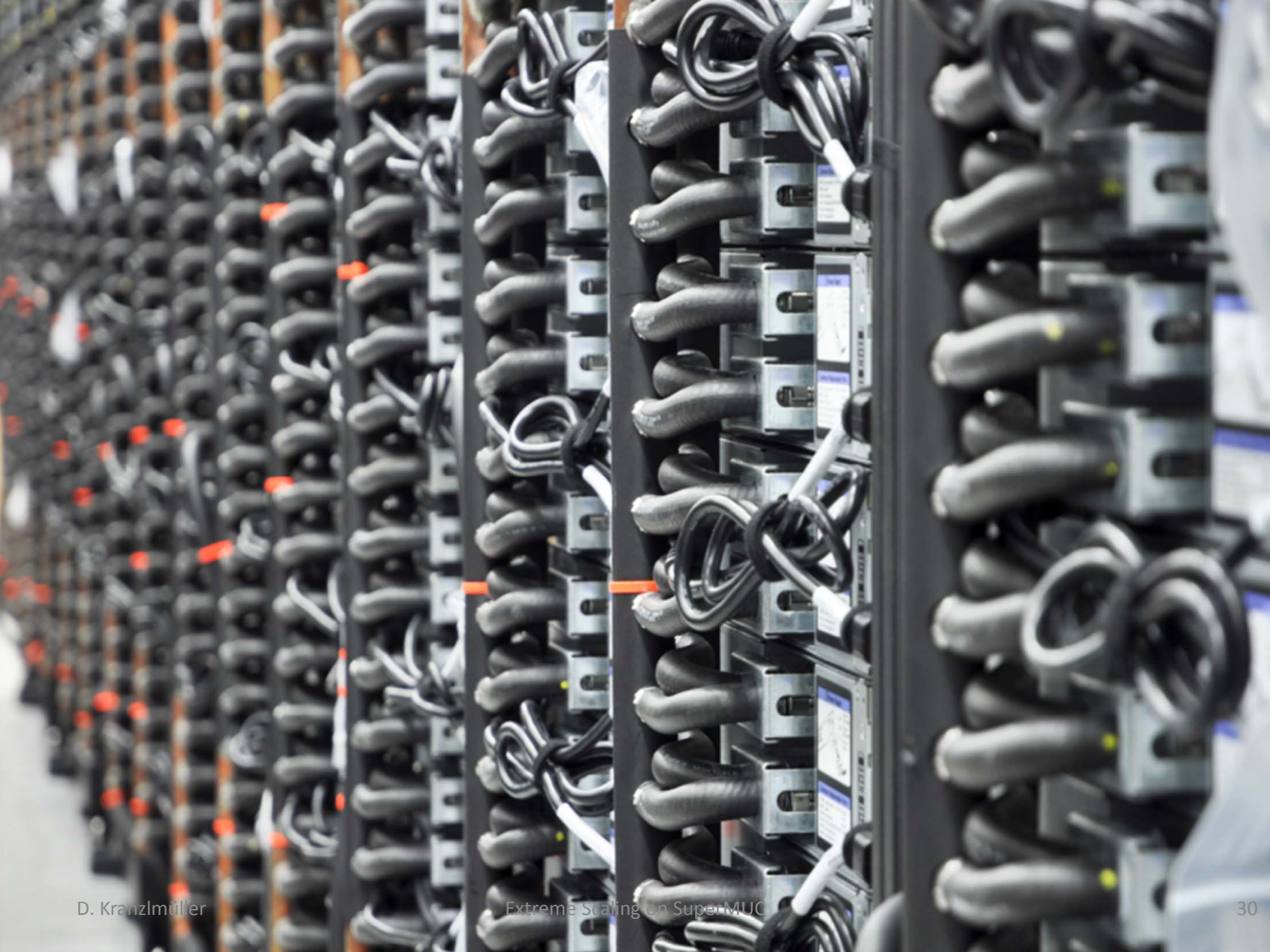
iDataplex DWC Rack
w/ water cooled nodes
(front view)



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

IBM iDataplex dx360 M4





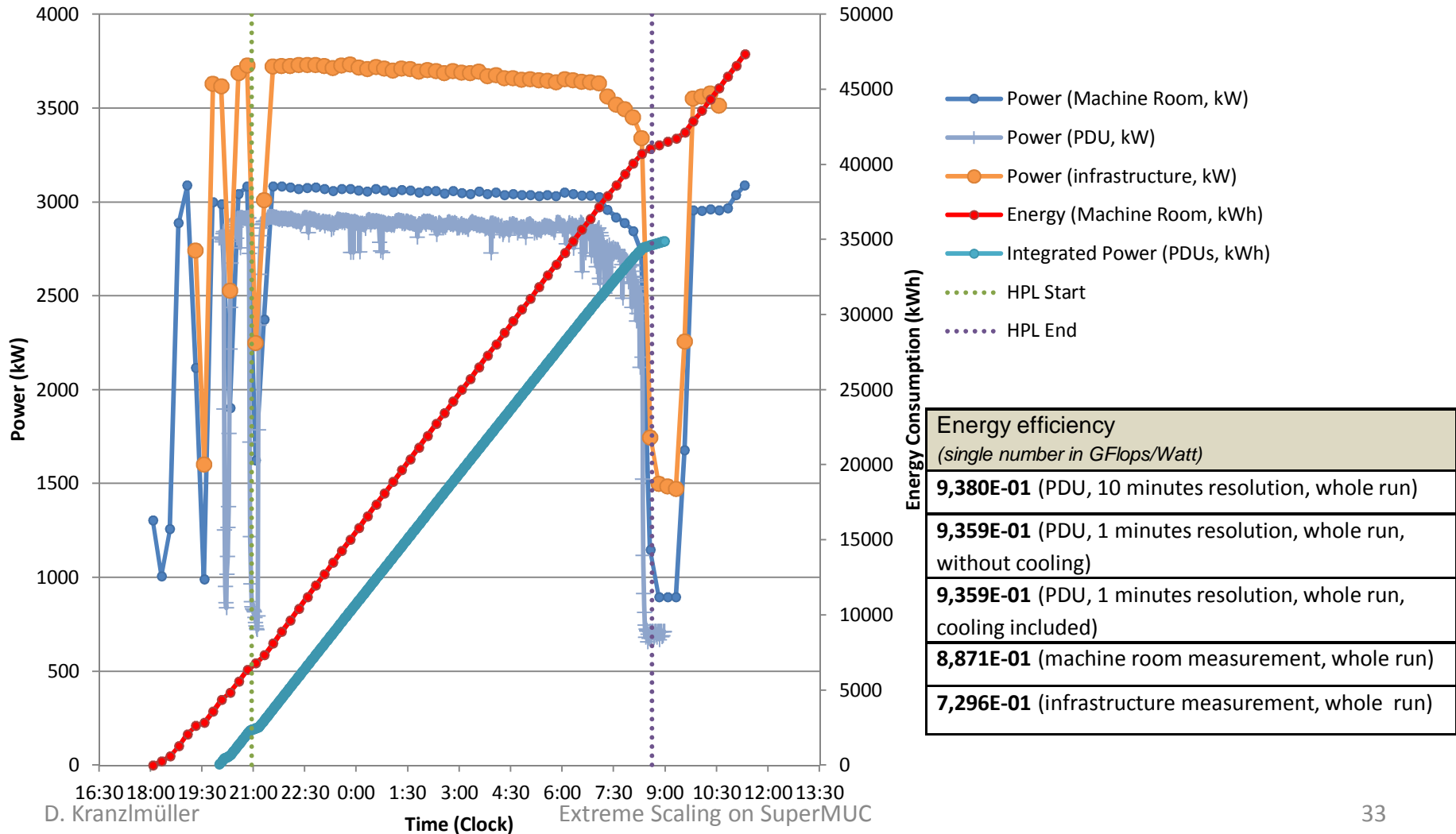
Cooling Infrastructure



Cooling Infrastructure (Roof)



SuperMUC HPL Energy Consumption



Extreme Scaling on SuperMUC

Dieter Kranzlmüller
kranzlmue@lrz.de

Visit us at SC13 Booth # 3241

