Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

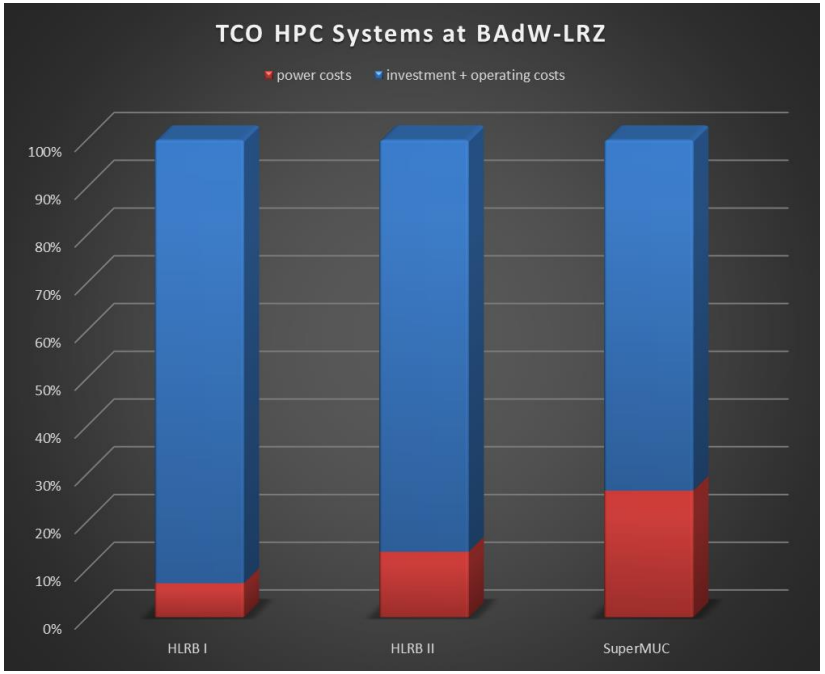
Energy Efficient data centers
A holistic approach and best practice at LRZ

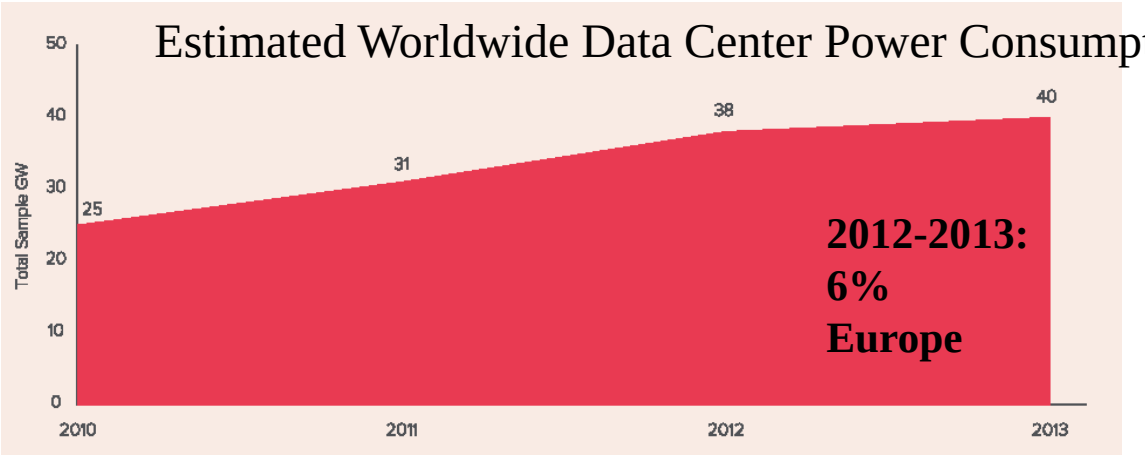
CERN / ERF / ESS Workshop – Energy for Sustainable Science at Research Infrastructures
DESY Hamburg, October 29th, 2015

Arndt Bode Chairman of the Board, Leibniz-Rechenzentrum of the Bavarian Academy of Sciences and Humanities and Technische Universität München

LRZ TCO

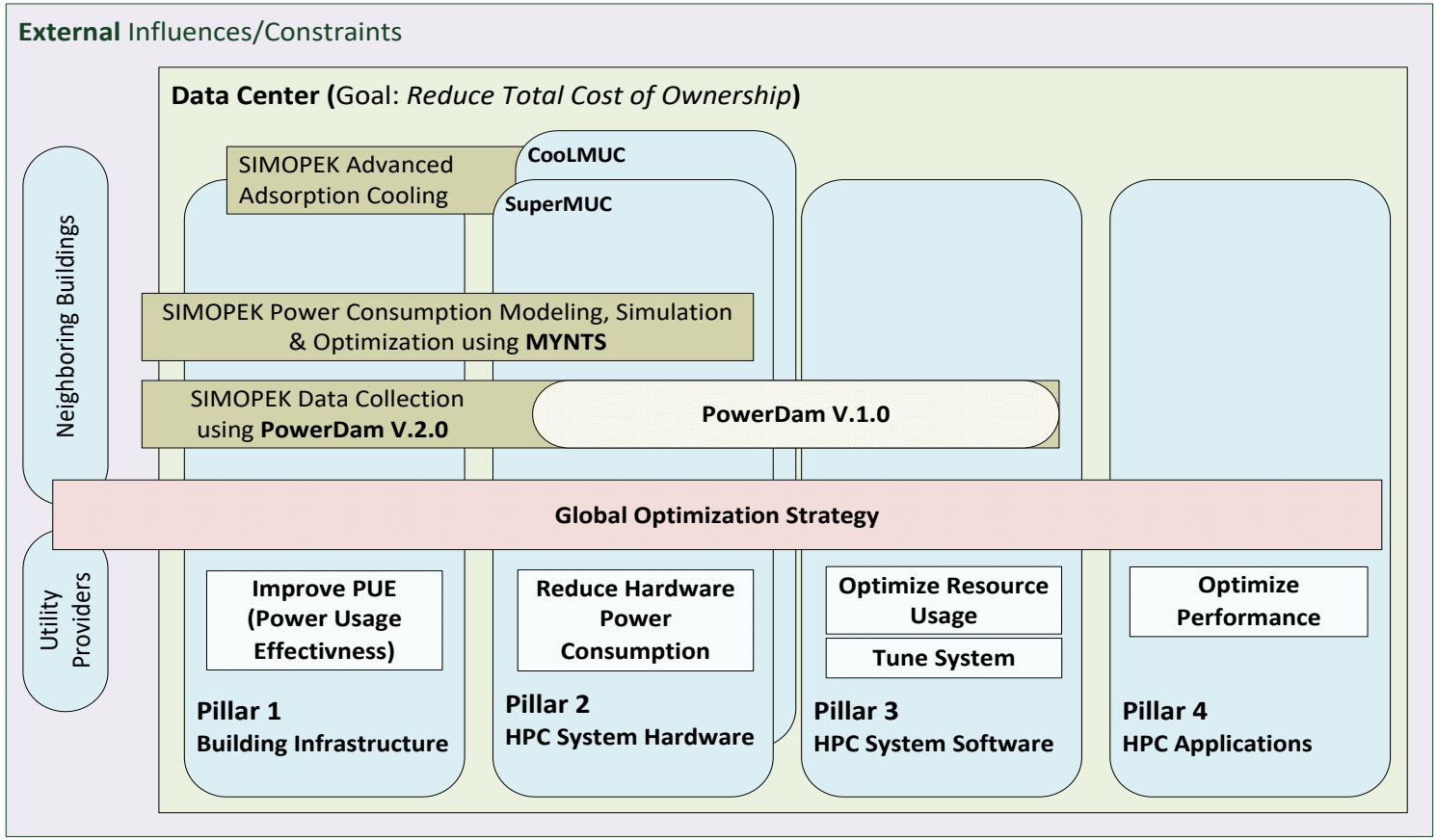
Normalized TCO:





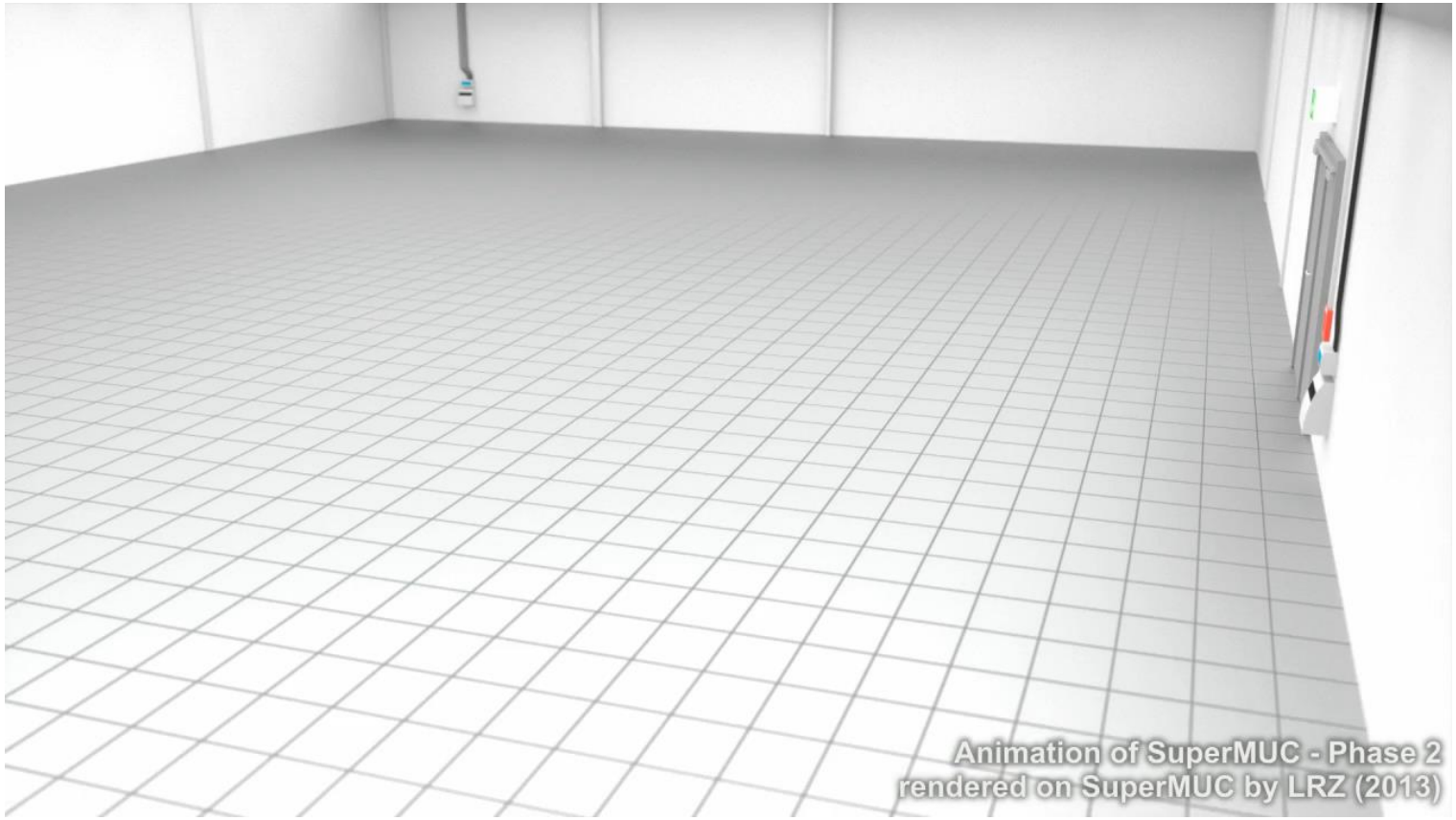
Source: DataCenterDynamics Focus, Volume 3, Issue 33, Jan/Feb 2014

SIMOPEK Project Coverage

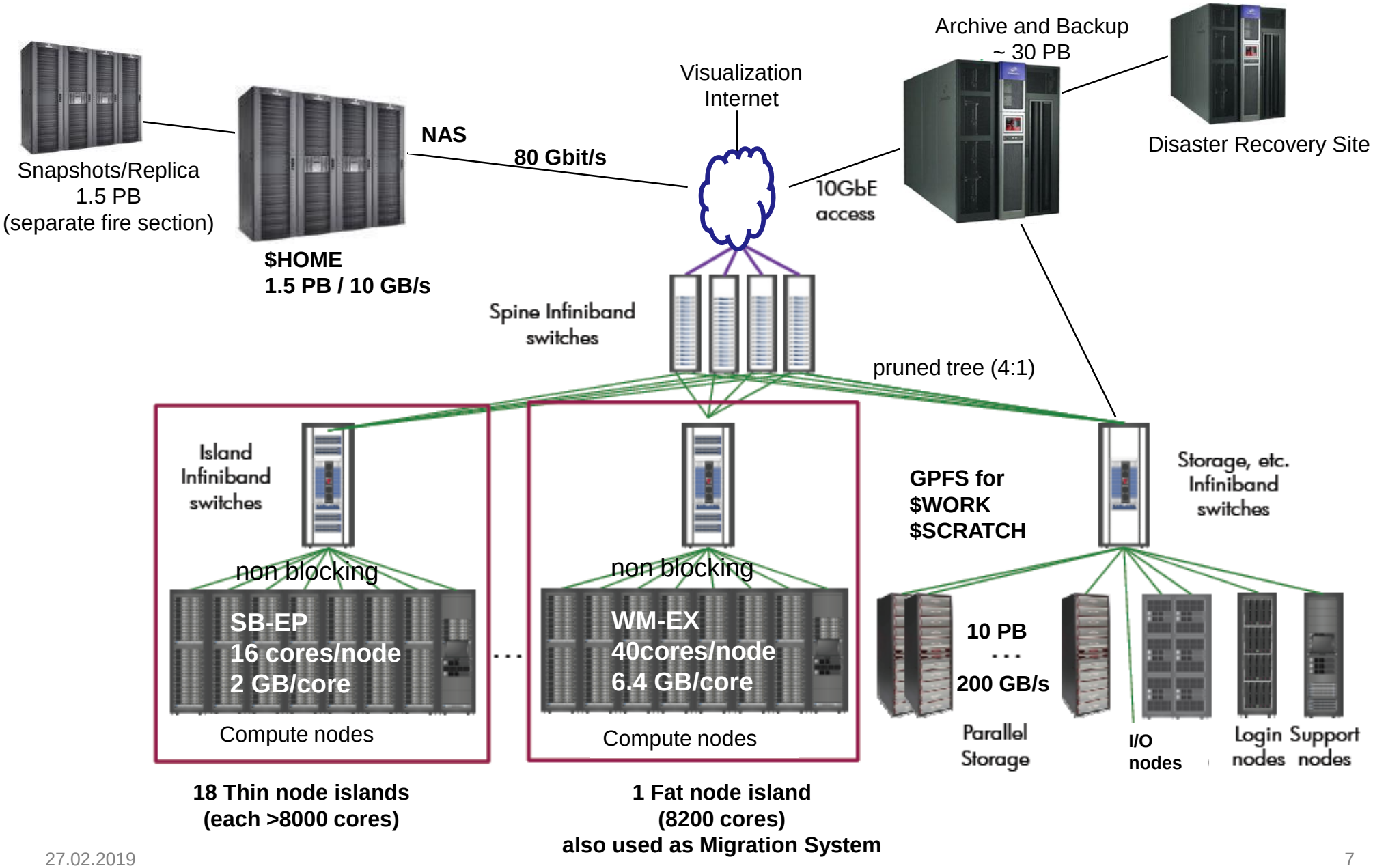


Open Access 4 Pillar Framework Paper: http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s00450-013-0244-6

SuperMUC – Phase 1/2



SuperMUC General Configuration – Phase 1



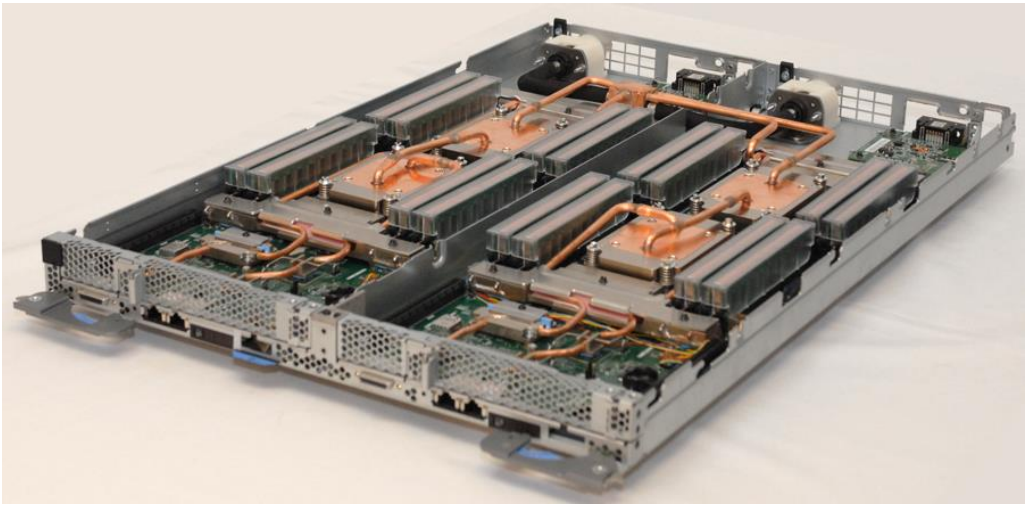

lrz
Inbetriebnahme
SuperMUC
Phase2

SuperMUC Phase 2

Foto: Torsten Bloth, Lenovo

- **Lenovo NeXtScale Water Cool (WCT)**
  - ✓ Cooling liquid temperatures 30° C – 45° C
  - ✓ Compressor-free cooling 365 d.p.a.
- **72 Racks**
  - ✓ 48 Compute
  - ✓ 9 Infiniband
  - ✓ 6 In-Row Cooler
  - ✓ 9 Management + Storage

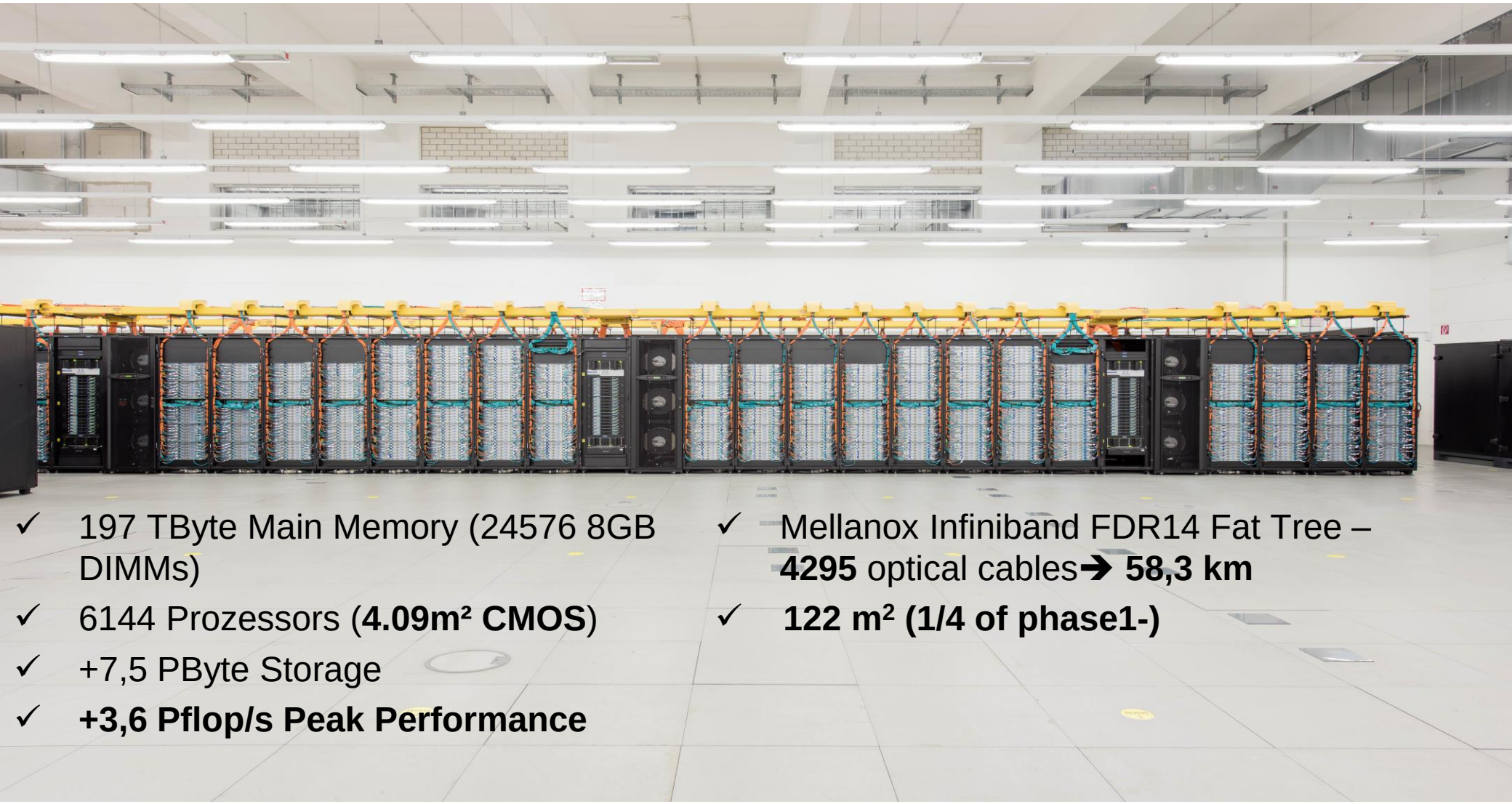
SuperMUC Phase 2



Fotos: Torsten Bloth, Lenovo

- **3072 Compute Nodes**
  - ✓ Lenovo NeXtScale nx360M5 WCT
  - ✓ 2 x Intel E5-2697v3 2.6GHz 14c
  - ✓ 64GB Main memory
  - ✓ Mellanox Connect-IB Single Port HCA
  - ✓ Diskless
  - ✓ Direct water cooling

SuperMUC Phase 2



- ✓ 197 TByte Main Memory (24576 8GB DIMMs)
- ✓ 6144 Prozessors (**4.09m² CMOS**)
- ✓ +7,5 PByte Storage
- ✓ **+3,6 Pflop/s Peak Performance**
- ✓ Mellanox Infiniband FDR14 Fat Tree – **4295** optical cables➔ **58,3 km**
- ✓ **122 m² (1/4 of phase1-)**

SuperMUC Phase 2

Fotos: Torsten Bloth, Lenovo

Savings in Energy:

- ✓ New Intel-Processor-Technology
- ✓ Direct Cooling
  - 10% Savings compared to air cooling
  - ~25% Savings by compressor-free cooling

- ✓ Energy-aware Scheduling
  - + 6% Savings

➡ ~40% better energy efficiency

„Extreme Scale-out“ 28 days later

Friendly-User Phase of the upcoming SuperMUC Phase 2 (3.6 PFlop/s peak, 2.8 Pflop/s Linpack, 86016 cores)

Available: **63.4** million core-h
Used: **43.8** million core-h

41 Scientists from **14** Institutes
14 Applications running on full system

Extreme Scale-out Phase2, lessons learned

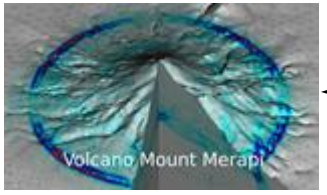
Ferdinand Jamitzky[1], Helmut Satzger[1], Nicolay Hammer[1], Momme Allalen[1], Alexander Block[1], Markus M. Müller[1], Anupam Karmakar[1], Matthias Brehm[1], Reinhold Bader[1], Luigi Iapichino[1], Antonio Ragagnin[1], Vasilios Karakasis[1], Dieter Kranzlmüller[1], Arndt Bode[1], Herbert Huber[1], Martin Kühn[2], Rui Machado[2], Daniel Grünewald[2], Philipp V. F. Edelmann[3], Friedrich K. Röpke[3], Markus Wittmann[4], Thomas Zeiser[4], Gerhard Wellein[5], Gerald Mathias[6], Magnus Schwörer[6], Konstantin Lorenzen[6], Christoph Federrath[7], Ralf Klessen[8], Karl-Ulrich Bamberg[9], Hartmut Ruhl[9], Florian Schornbaum[10], Martin Bauer[10], Anand Nikhil[11], Jiaxing Qi[11], Harald Klimach[11], Hinnerk Stüben[12], Abhishek Deshmukh[13], Tobias Falkenstein[13], Klaus Dolag[14], and Margarita Petkova[14]

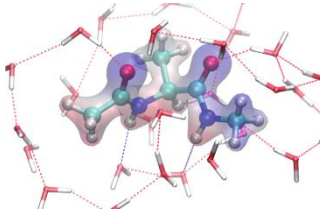
[1] LRZ, Boltzmannstrasse 1, 85748 Garching b. Muenchen http://www.lrz.de
[2] CCHPC - Fraunhofer ITWM, Fraunhofer Platz 1, 67663 Kaiserslautern http://www.gpi-site.com
[3] Heidelberger Institut für Theoretische Studien, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany
[4] Erlangen Regional Computer Center (RRZE), University of Erlangen-Nürnberg, Martensstr. 1, 91058 Erlangen, Germany
[5] Department of Computer Science, University of Erlangen-Nürnberg, Germany
[6] Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität München, Oettingenstr. 67, 80538 München, Germany
[7] Research School of Astronomy and Astrophysics, The Australian National University, Canberra, ACT 2611, Australia
[8] Universität Heidelberg, Zentrum für Astronomie, Institut für Theoretische Astrophysik, Albert-Ueberle-Strasse 2, D-69120 Heidelberg, Germany
[9] Chair for Computational and Plasma Physics at the LMU, Munich
[10] Chair for System Simulation, University of Erlangen-Nürnberg, Cauerstraße 11, 91058 Erlangen, Germany
[11] Chair of Simulation Techniques & Scientific Computing, University Siegen
[12] Universität Hamburg, Zentrale Dienste, Schlüterstraße 70, 20146 Hamburg
[13] Fakultät für Maschinenwesen, Institut für Technische Verbrennung, RWTH Aachen University, Templergraben 64, 52062 Aachen
[14] Universitäts-Sternwarte München, Scheinerstr. 1, D-81679 München, Germany

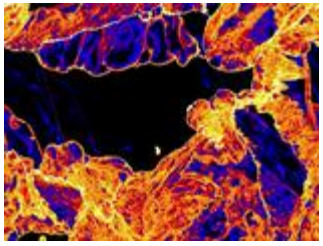
Abstract. During May and June 2015 LRZ conducted a friendly user operation block operation of the their upcoming new extension of SuperMUC called Phase 2 which consists of 86,016 Intel Haswell cores distributed to 6 islands resulting in a peak performance of 3.6 PFlop/s. Selected user groups had the opportunity to use the system for 28 days of continuous operation as so called "friendly users" and run jobs up to the whole system size. This work presents results obtained during this period and the lessons learned from the operational point of view.

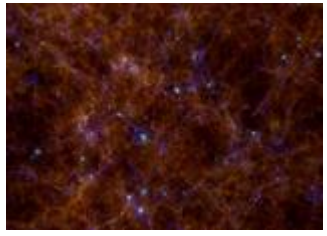
Keywords: Supercomputing, HPC

14 Applications 2015



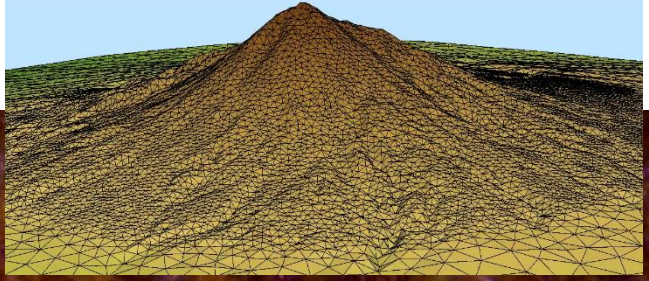


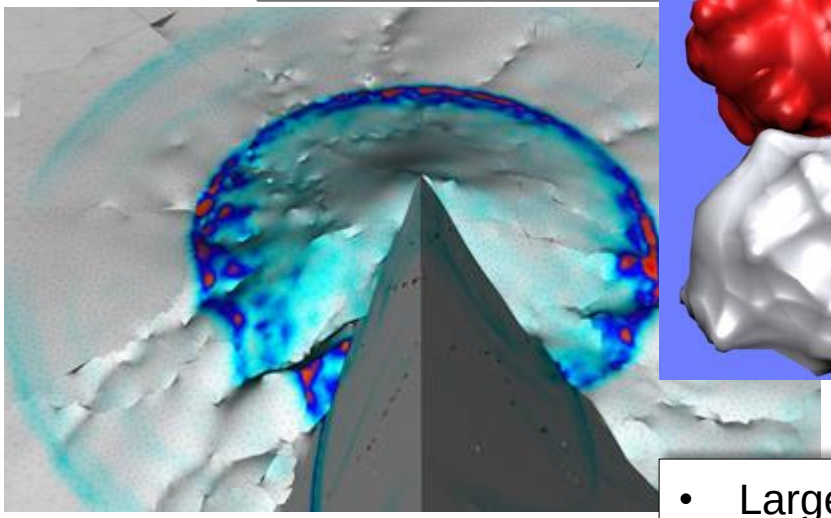


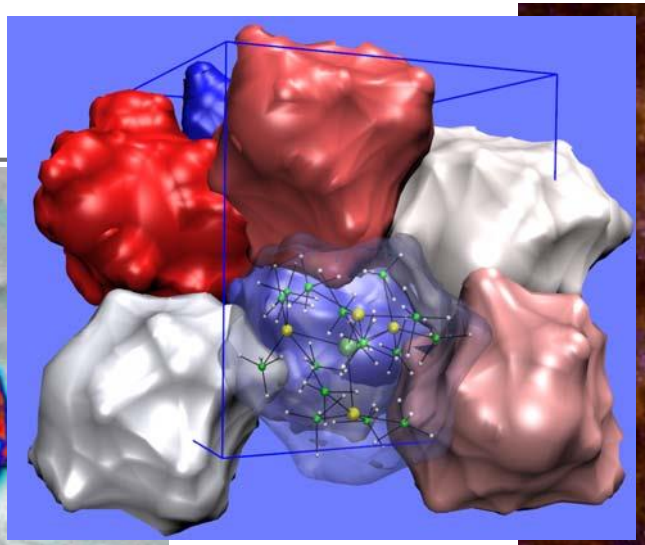


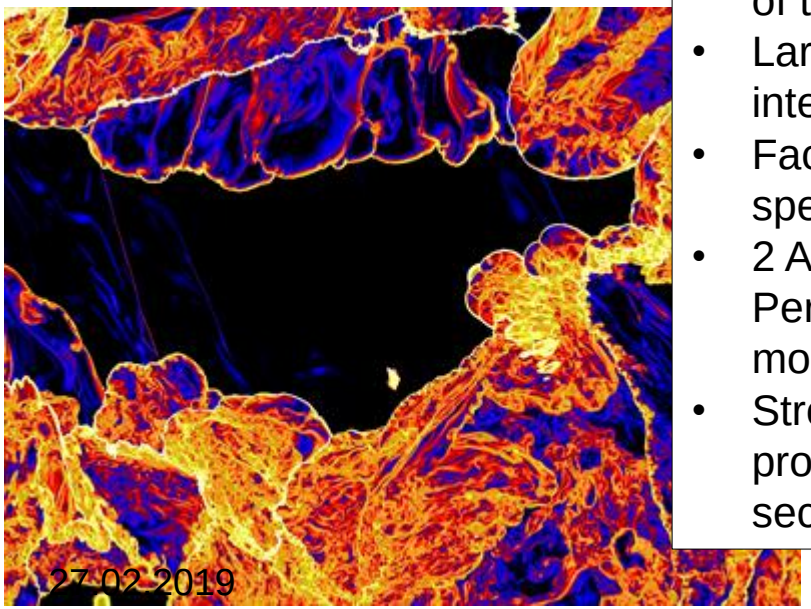
Software	Application
BQCD	Quantumchomodynamics
SeisSol	Seismology
GPI-2 / GASPI	Global Adress Space Library
Seven-League Hydro	Astropysics
ILBDC	Lattice Boltzmann
Iphigenie	Molekular Dynamics
FLASH	Astro CFD
Gadget	Cosmology
PSC	Plasmaphysics
waLBerla	Lattice Boltzmann
Musubi	Lattice Boltzmann
CIAO	CFD, Combustion
Vertex3D	Stellar Astrophysics
LS1-Mardyn	Material Science

Results









- Largest Cosmology Simulation so far (10% of the visible universe)
- Largest pseudo-spectral simulation of interstellar turbulence (10,000^3 Cells)
- Factor 100 better resolution for molecular spectra
- 2 Applications with sustained PFLOP/s Performance (SeisSol and LS-Mardyn) for more than 20 hours
- Strong scaling of a seismic reconstruction problem using GPI-2 (from 16 hours to 55 seconds)

Integration of all IT-Systems into LRZ Dual-Cube Dark Center

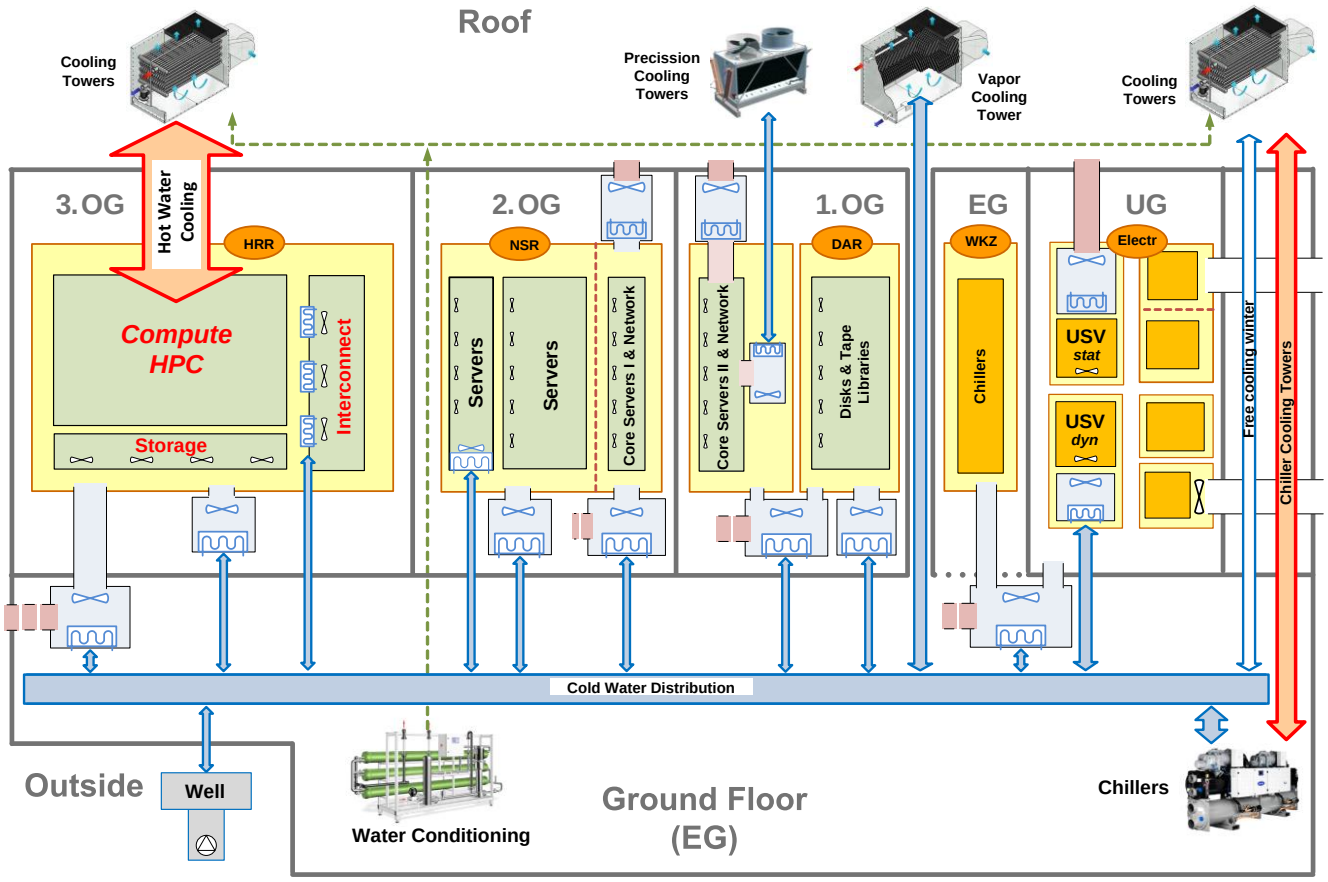
Infrastructure and Servers for HPC, Back-up and Archiving, Munich Network, Visualization and „General IT-Services“

Advantages of LRZ-Dual-Cube Dark Center:

- ✓ RAS through isolation, redundancy and IT-control
- ✓ Efficient fire extinction on the basis of Argon
- ✓ Reliable energy provision by layered USV concepts based on flying wheels (12) array of batteries, diesel engine
- ✓ Energy efficiency by layered cooling concept: cold air, chilled water at different temperatures, direct cooling based on warm water
- ✓ … and intelligent technologies: free cooling, reuse of waste heat (adsorption machines, building climate), use of ground water / geothermal heat, fine grained and intelligent monitoring, tools for optimization („automatic DVFS“) and user-information
- ✓ Flexibility supported by functional spezilization and separation of floors of the Dark Center
- ✓ Synergies for end-users by tight interaction of HPC-Big Data-Networking-Visualization

View of control part for direct water cooling infrastructure

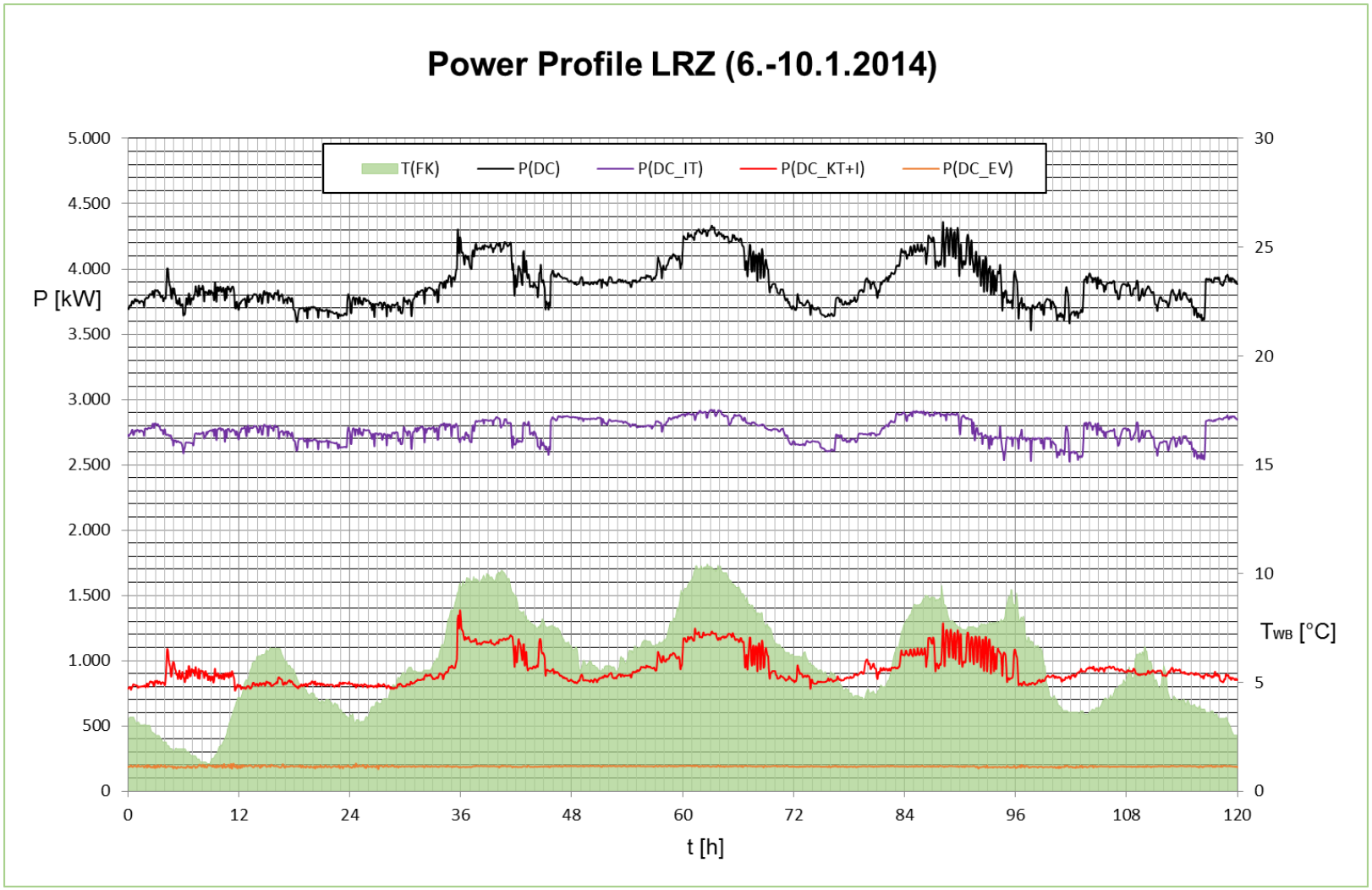
New Generation of HPC Data Centers Use a Mix of Different Cooling Technologies



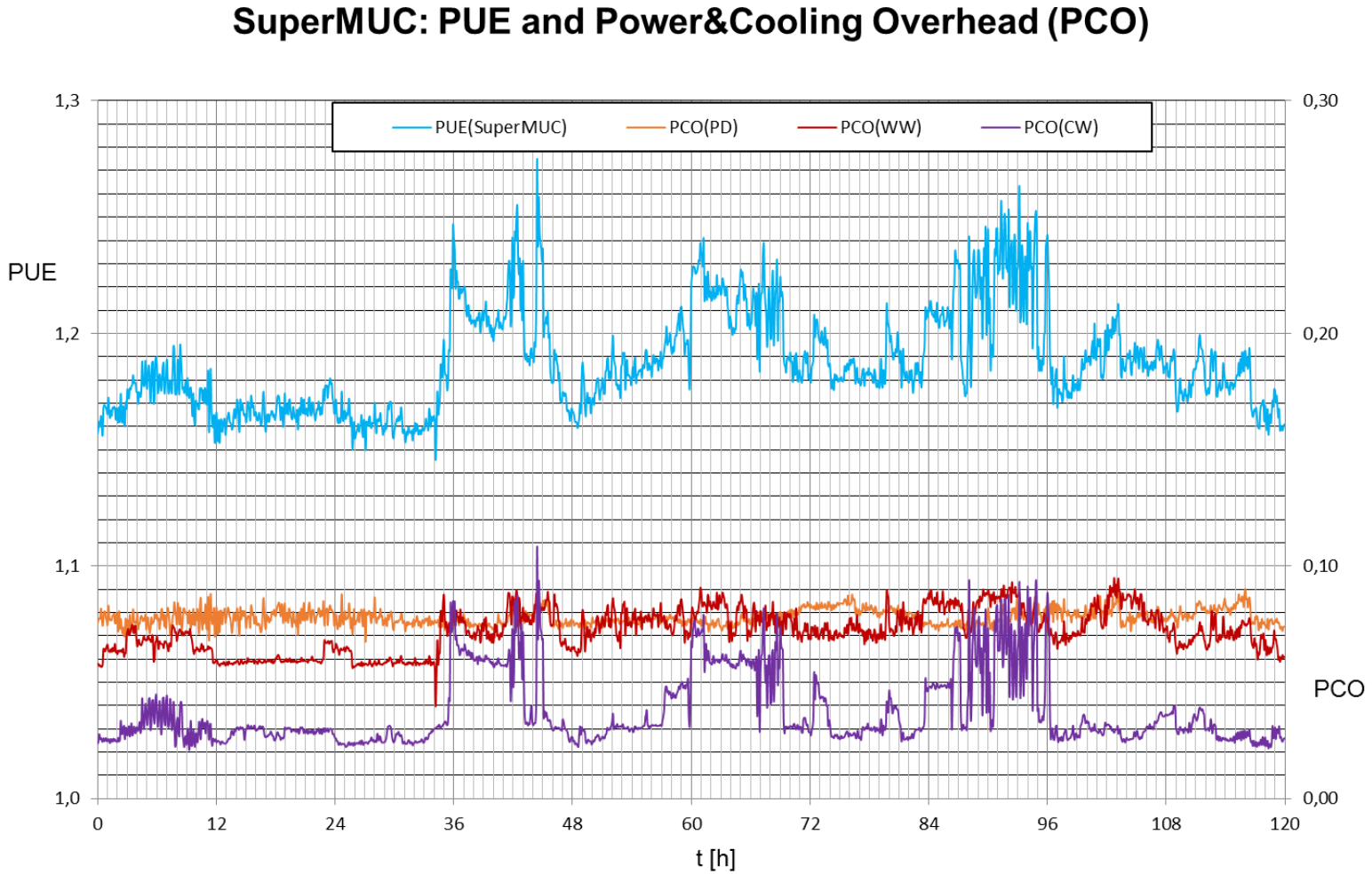
Cooling capacity LRZ (new construction):

- Vapor cooling: 2MW
- Well water: 600kW
- Chillers: 3.2MW
- Evaporative cooling towers: 8MW

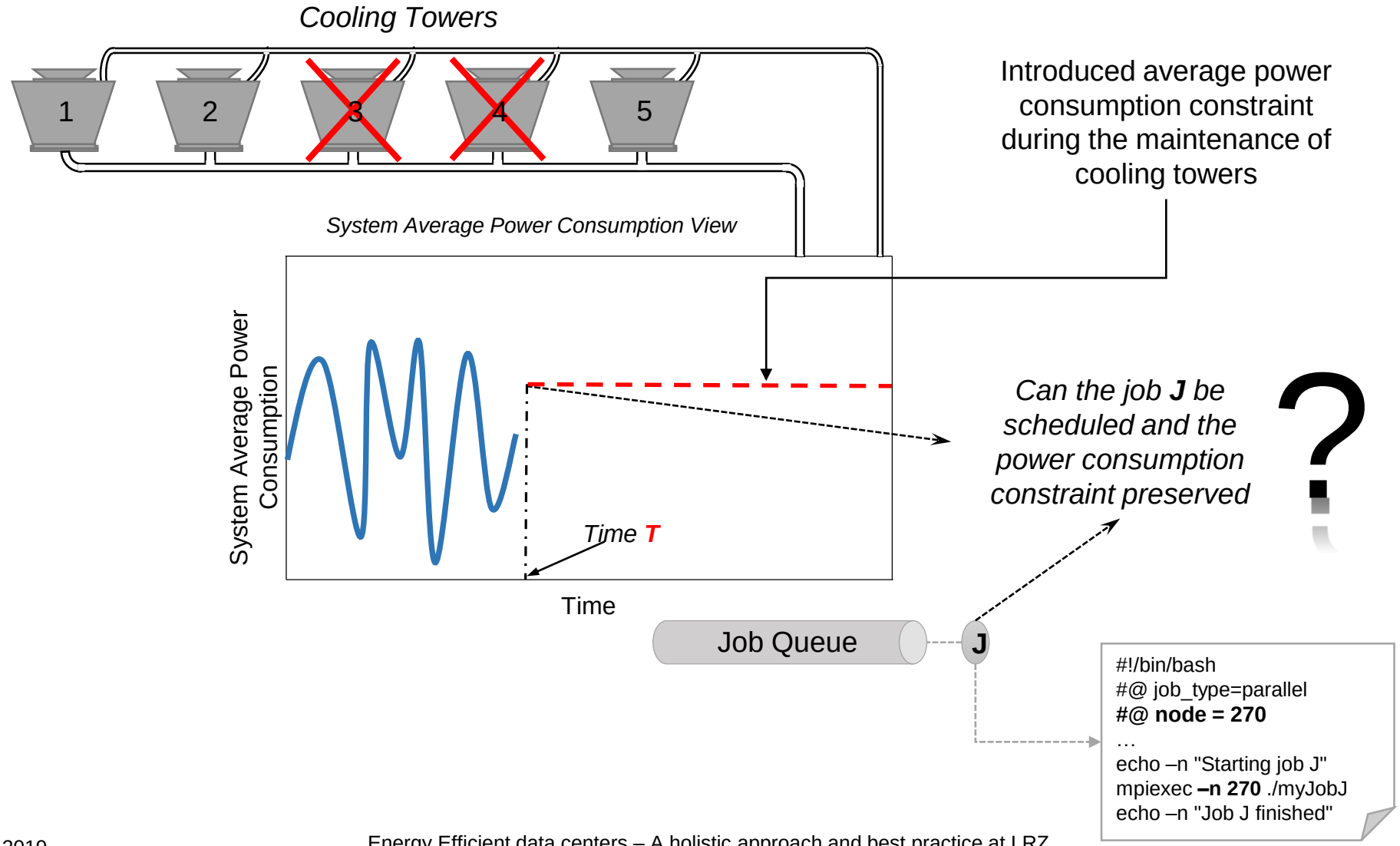
Power Profile of LRZ (6. – 10.1.2014)



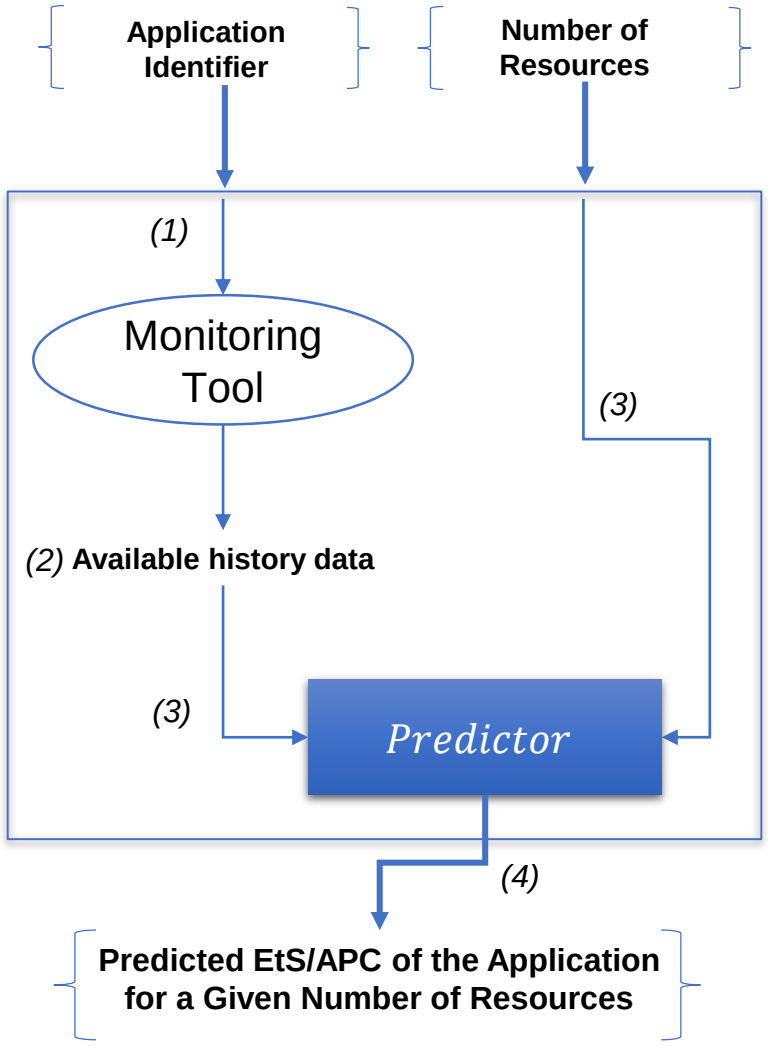
Energy Efficiency of SuperMUC (6. – 10.1.2014)

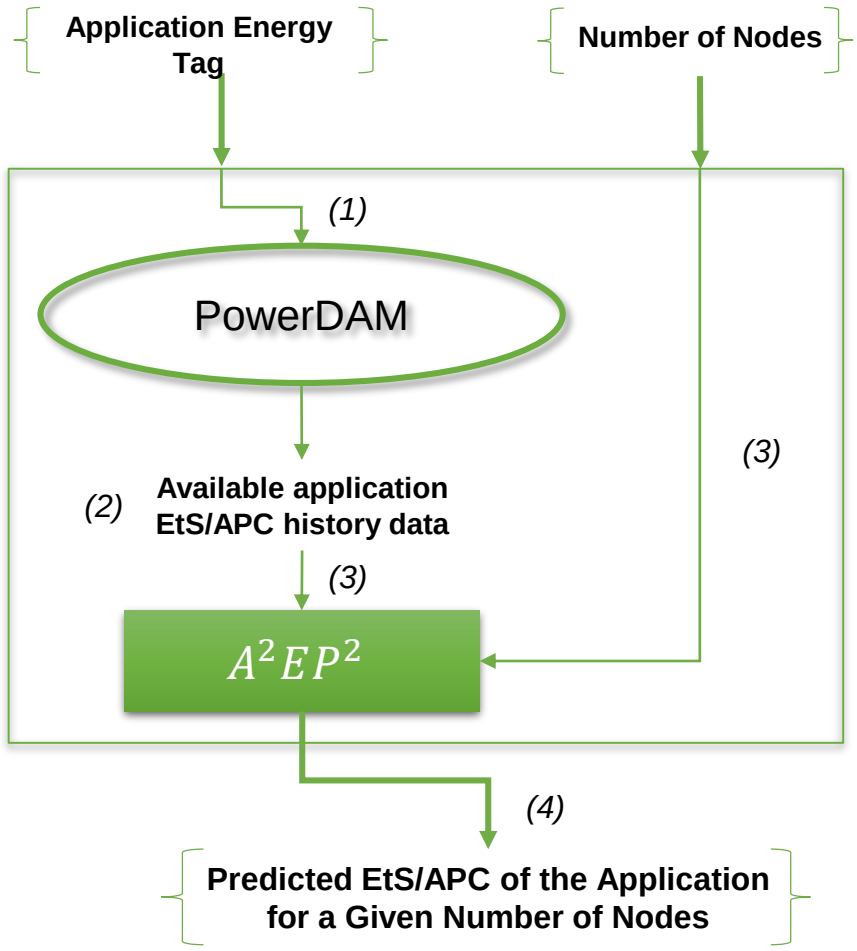


Predicting the Power Consumption of Strong and Weak Scaling HPC Applications, Hayk Shoukourian

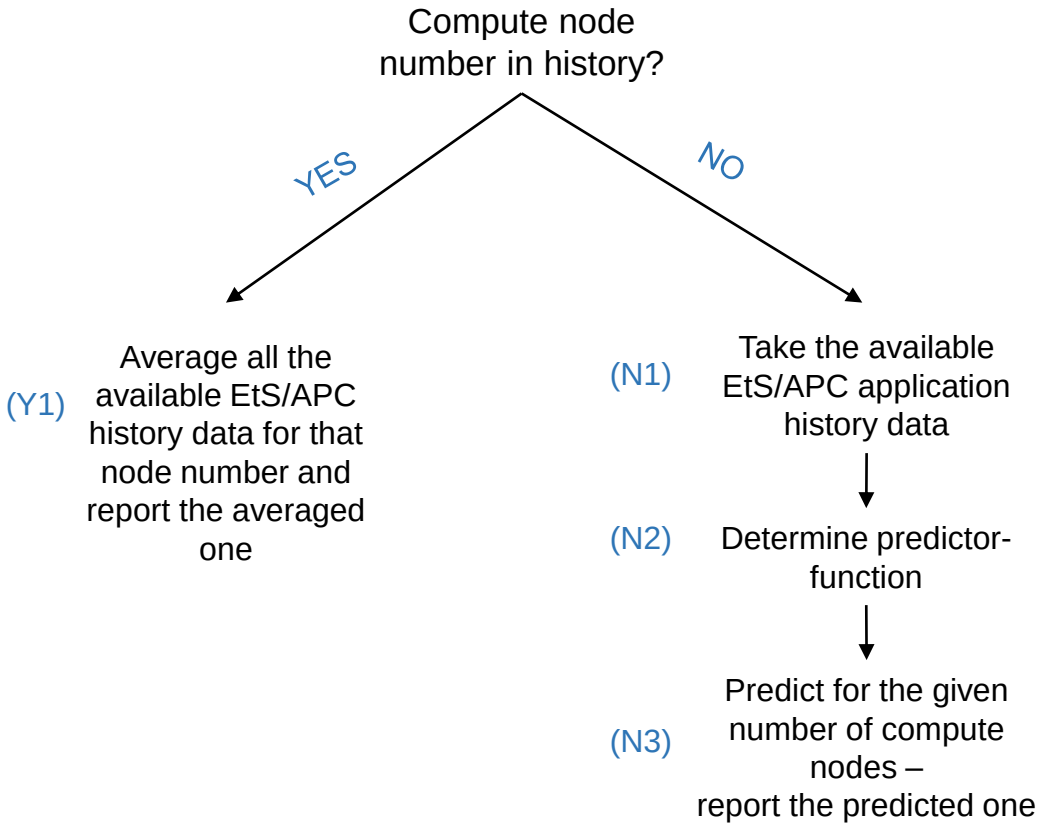


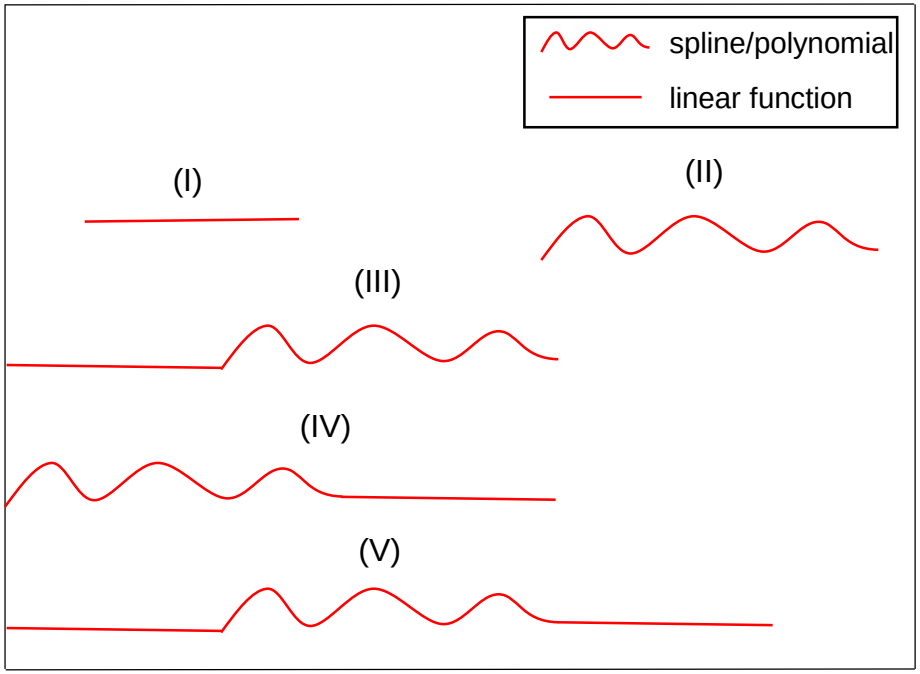
Adaptive Energy and Power Consumption Prediction (AEPCP) Process & Model





Adaptive Application Energy and Power Predictor (A^2EP^2)



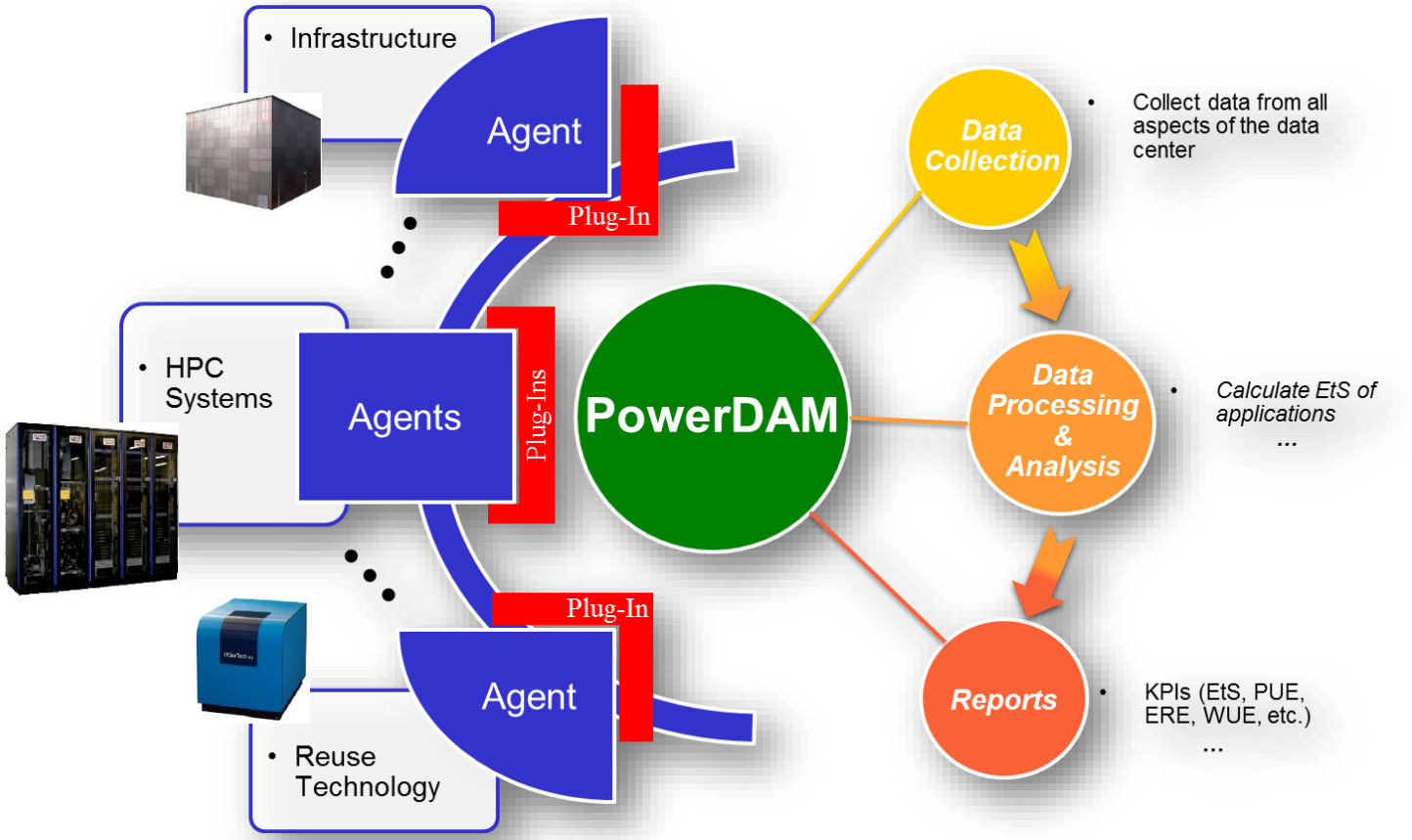


A^2EP^2 predictor-function estimation scenarious

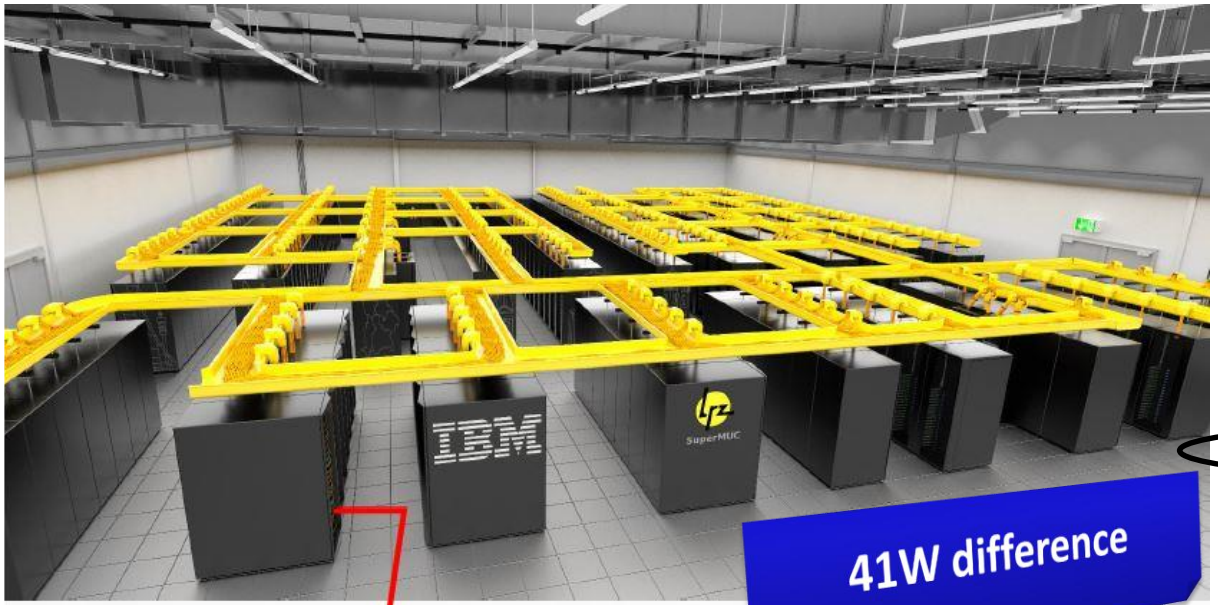
A^2EP^2 Workflow

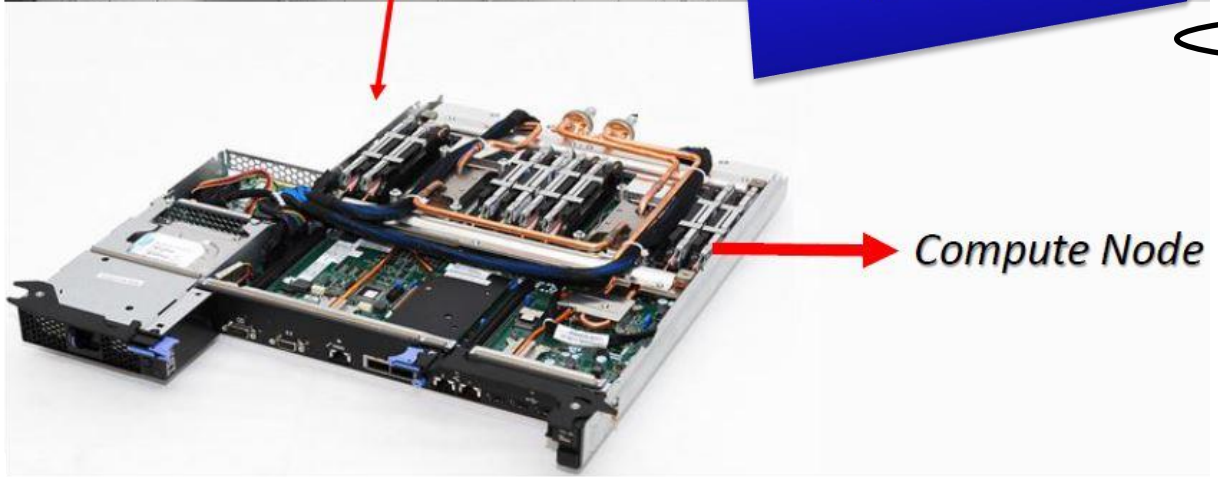
$$\%RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^{measured} - x_i^{predicted})^2}.\frac{100 \cdot n}{\sum_{i=1}^{n} x_i^{measured}}$$

Power Data Aggregation Monitor (PowerDAM)



Differences In Node Power Draw - SuperMUC





Node	Average Node Power (watt)
i05r01a03-ib	208
i05r01a04-ib	202
i05r01a05-ib	207
i05r01a06-ib	221
i05r01a07-ib	203
i05r01a08-ib	206
i05r01a09-ib	203
i05r01a10-ib	215
...	
i05r03c24-ib	210
i05r03c25-ib	212
i05r03c26-ib	221
i05r03c27-ib	211
i05r03c28-ib	229
i05r03c29-ib	216
...	
i05r05a16-ib	218
i05r05a17-ib	215
i05r05a18-ib	203
i05r05a19-ib	188
i05r05a20-ib	211
i05r05a21-ib	207
...	
i05r05a35-ib	205
i05r05a36-ib	207
i05r05a37-ib	198
i05r05a38-ib	219
i05r05c03-ib	215
i05r05c04-ib	218
i05r05c05-ib	216
i05r05c06-ib	215
i05r05c07-ib	200
i05r05c08-ib	206
...	

Energy efficiency in the (HPC-) Data Center

LRZ: „holistic“ approach to streamline the „four „pillars“: work in progress:

- building / infrastructure: (√)
- energy efficient system hardware: √
- system monitoring / analysis / control: √
- Efficient application algorithms: (√)