Energy to Solution:
A New Mission for Parallel Computing

By Ernst A. Graf

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

- Why is Energy to Solution relevant for Euro-Par
- Complexity of HPC Datacenters: LRZ as an example
- Energy to Solution analysis and Optimiziation: A Wholistic Approach
- First Steps at Leibniz Supercomputer Centre

Why „Energy to Solution“

Euro-Par Mission

„Euro-Par is an annual series of international conferences dedicated to the promotion and advancement of all aspects of parallel and distributed computing“

- Algorithms
- Theory
- Software Technology
- Hardware
- Applications from Scientific to Mobile

What about cost?

Cost for energy rises dramatically

Annual cost (in Germany prices) for TOP_10 of Top_500: 150 M€ p.a.

(66.758 MW [list], min. 100 MW including infrastructure)

Need to consider: „Energy to solution“ in a Wholistic Approach

Engineering approach: codesign Building-Infrastructure-System Hw/Sw-Application

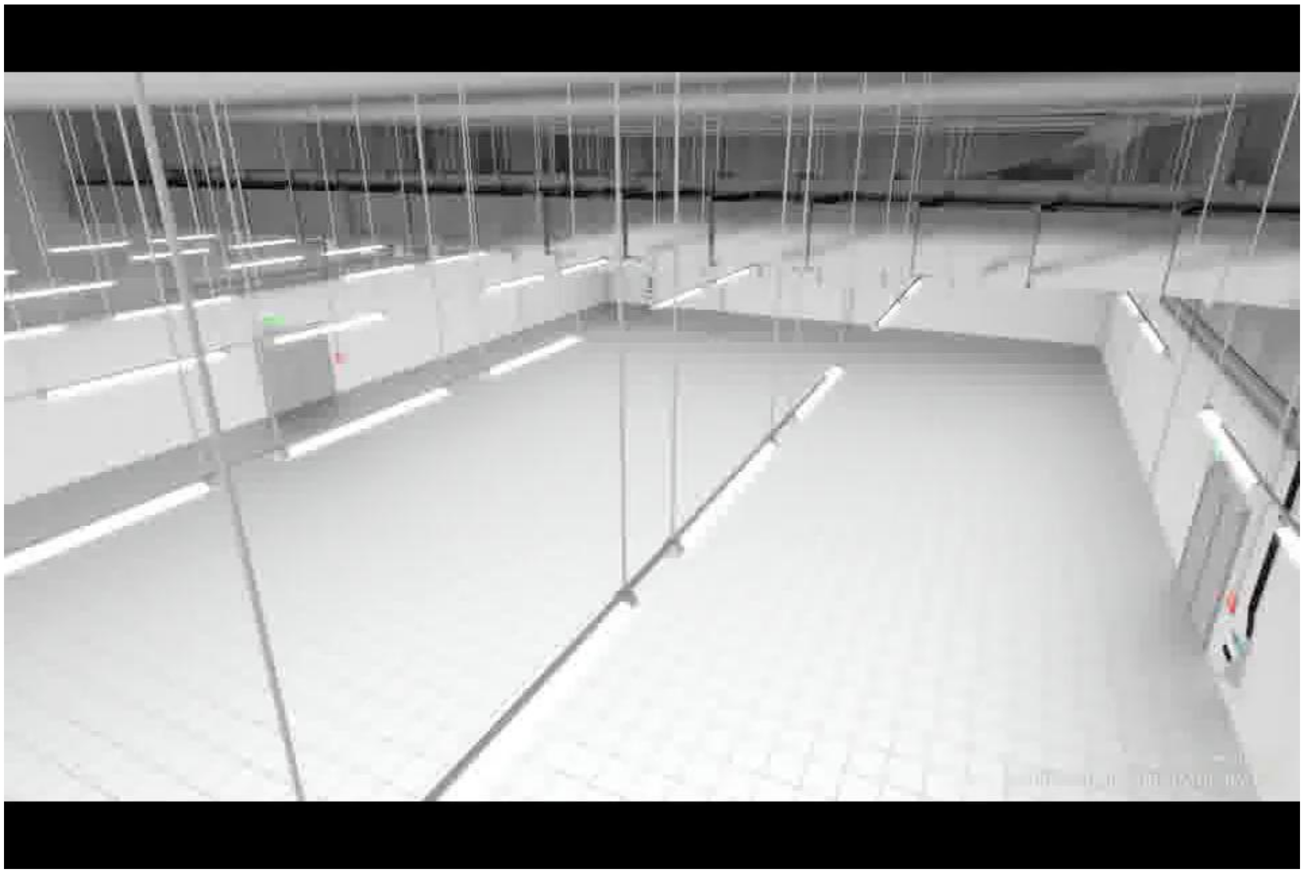
Calculation Basis: MW taken from TOP_500

Power	System	Processor	Performance (Linpack)
17.808	Tianhe-2	Xeon + PHi	33.862,7
8.209	Titan	Opteron + NVIDIA	17.590
7.890	Sequoia	Power BGQ	17.173,2
12.660	K	SPARC	10.510
3.945	Mira	Power BGQ	8.586,6
4.510	Stampede	Xeon + PHi	5.168,1
2.301	JUQUEEN	Power BGQ	5.008,9
1.972	Vulcan	Power BGQ	4.293,3
3.423	SuperMUC	Xeon	2.897
4.040	Tianhe-1	Xeon + NVIDIA	2.566
66.758 Accumulated MW (mostly without infrastructure: cooling, USV, cable losses, storage, interconnect, …..)			**107.655,8** Accumulated PFLOPs

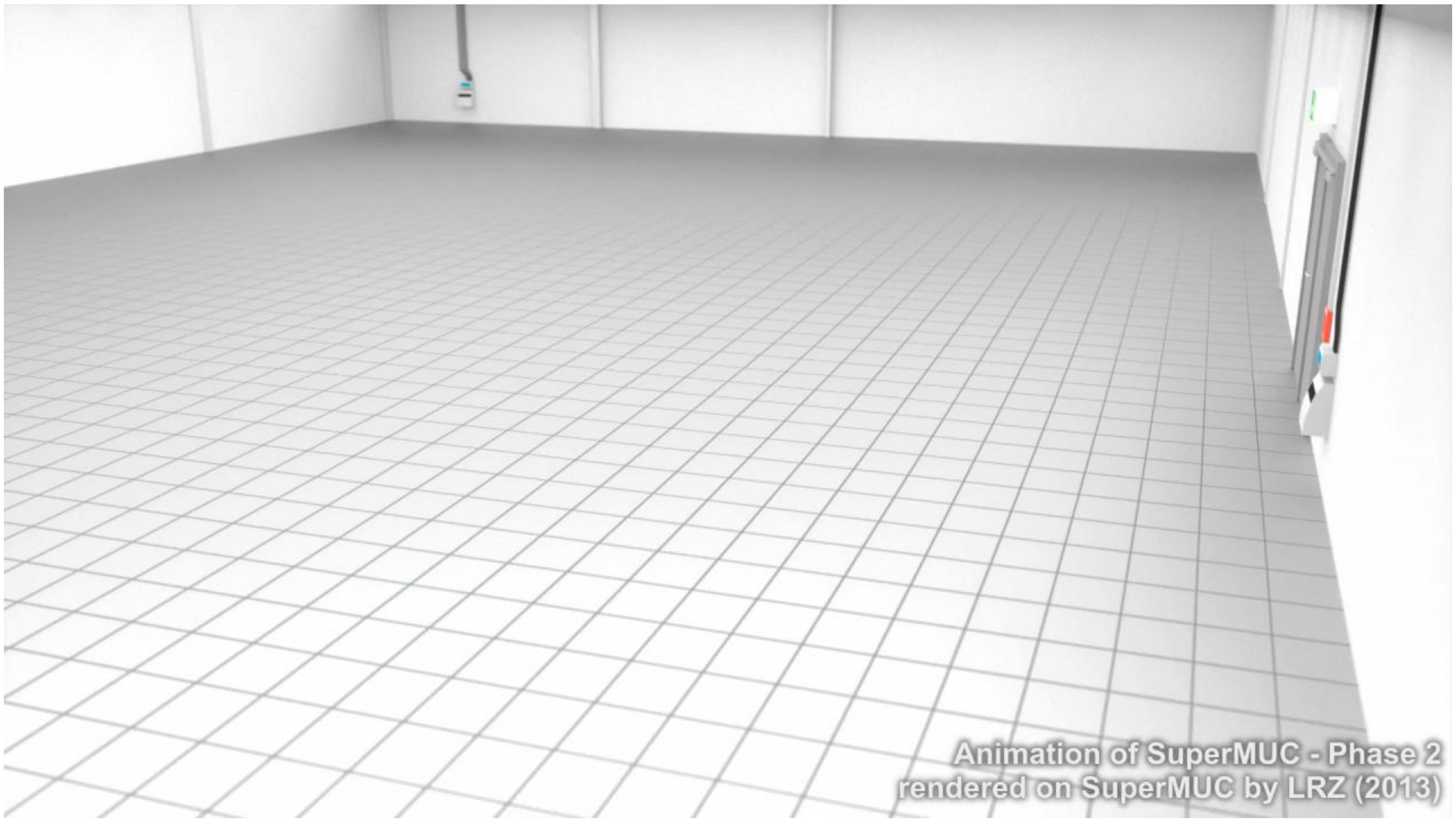
Energy cost 2012 (NUS consulting)

US-$-Cents per KWh	
Italy	20.23
Germany	15.15
Spain	13.52
UK	12.45
Belgium	11.92
Australia	11.68
Austria	11.05
Poland	9.30
US	8.89
France	8.76
Finnland	8.64
Sweden	7.95
Canada	7.58

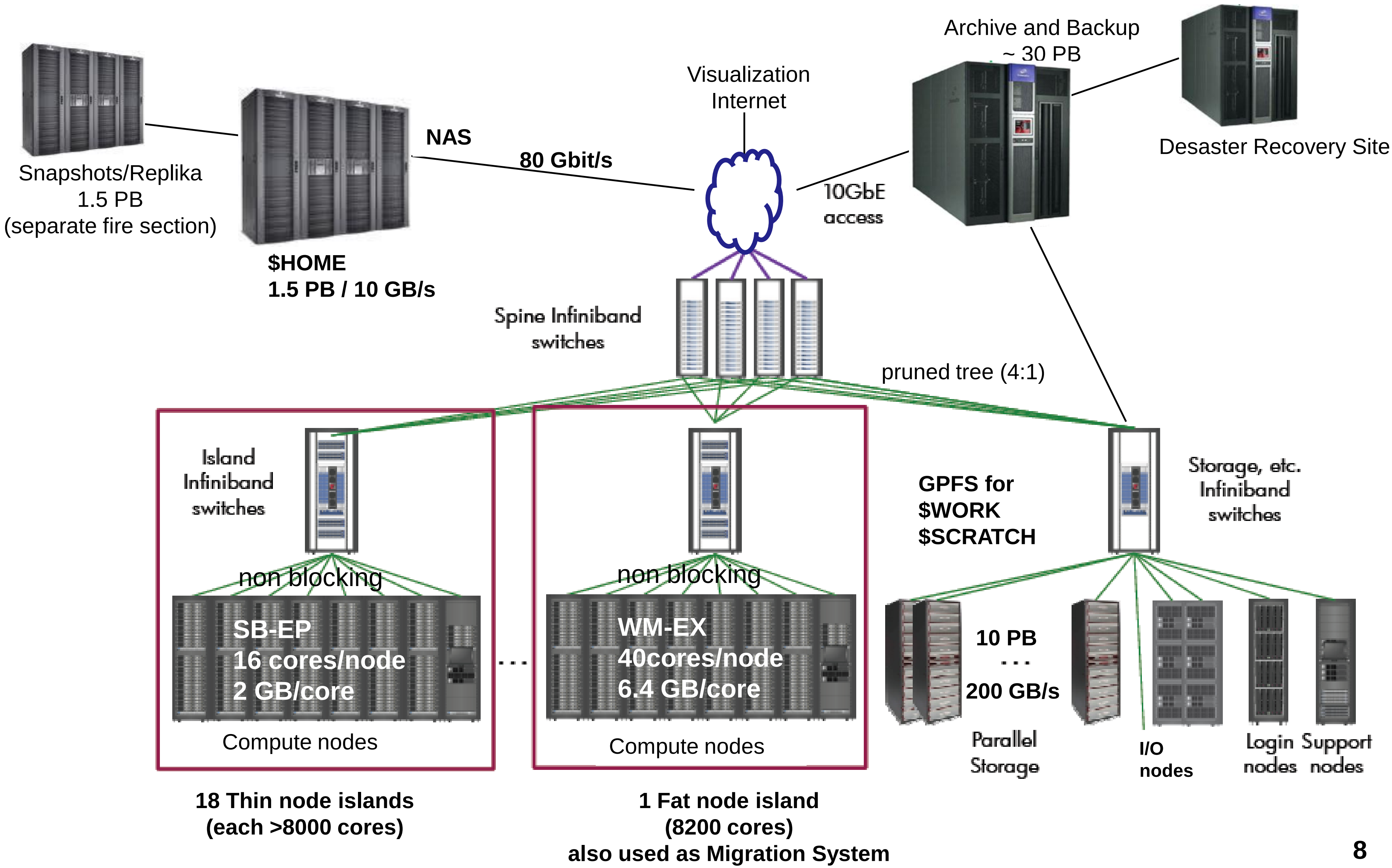
Basis:
1 MW for 450 h delivery



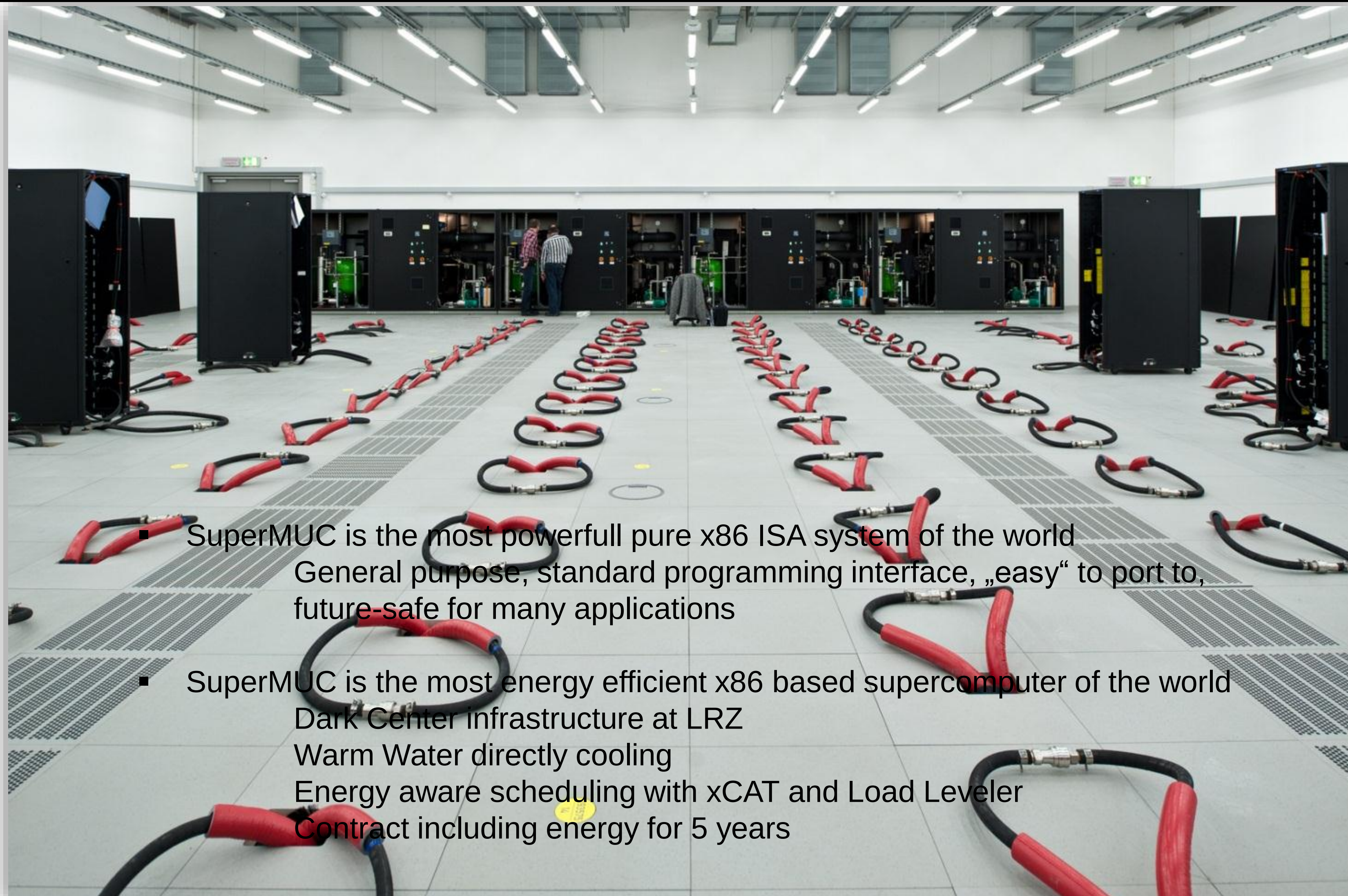
SuperMUC - Phase 2 (Animation)



SuperMUC General Configuration - the traditional view



What‘s special about SuperMUC

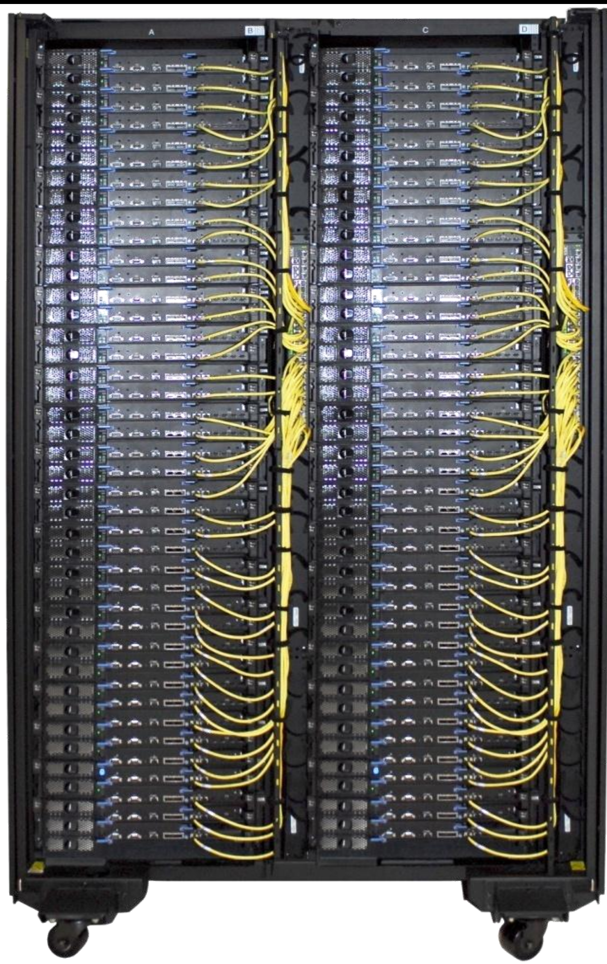


- SuperMUC is the most powerfull pure x86 ISA system of the world
  - General purpose, standard programming interface, „easy“ to port to, future-safe for many applications
- SuperMUC is the most energy efficient x86 based supercomputer of the world
  - Dark Center infrastructure at LRZ
  - Warm Water directly cooling
  - Energy aware scheduling with xCAT and Load Leveler
  - Contract including energy for 5 years

IBM iDataplex dx360 M4

- **Heat flux > 90% to water; very low chilled water requirement**
- **Power advantage over air-cooled node:**
  - Warm water cooled ~10%
    (cold water cooled ~15%)
  - due to lower $T_{components}$ and no fans
- **Typical operating conditions: T_{air} = 25 – 35°C, T_{water} = 18 – 45°C**

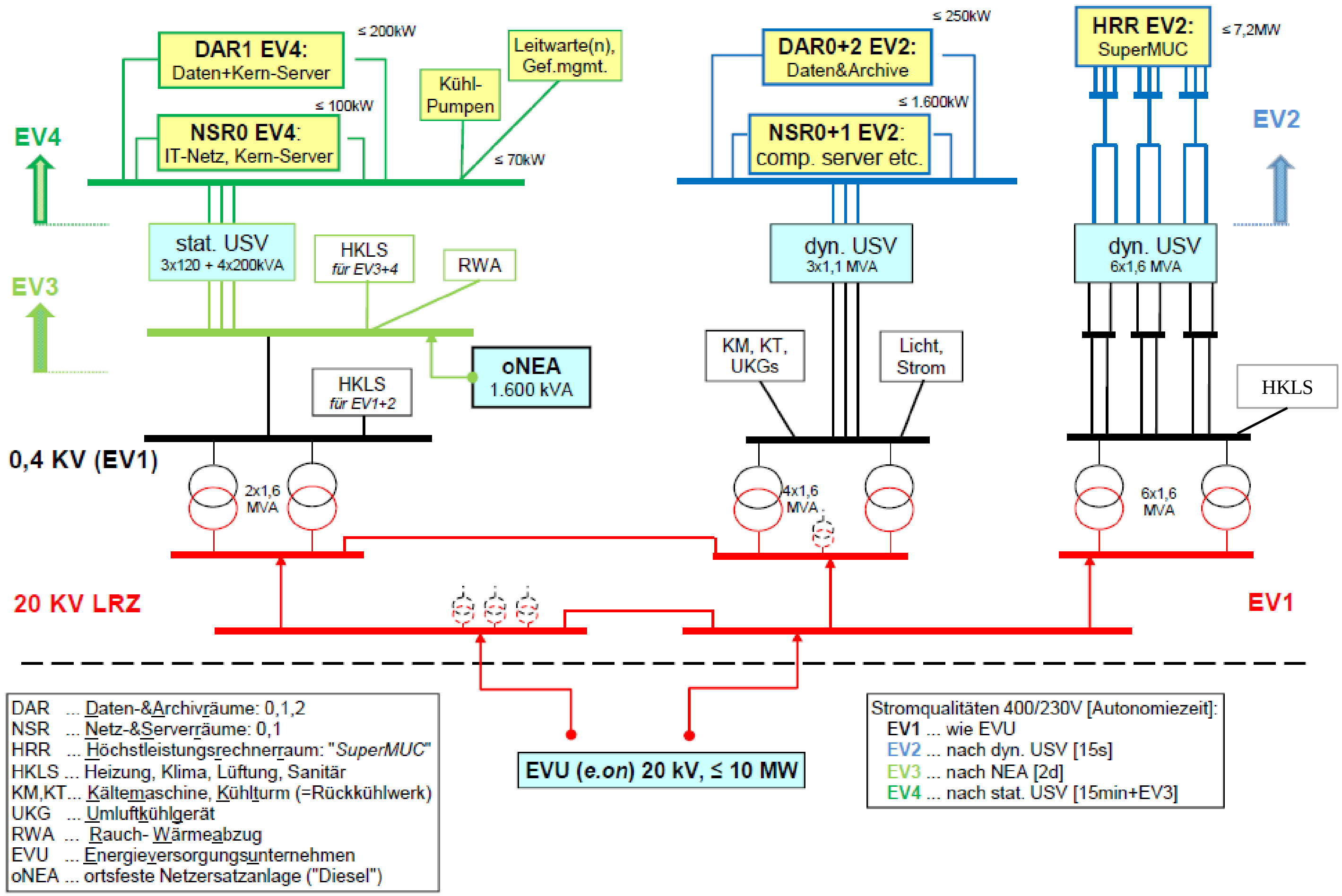
IBM System x iDataPlex Direct Water Cooled Rack



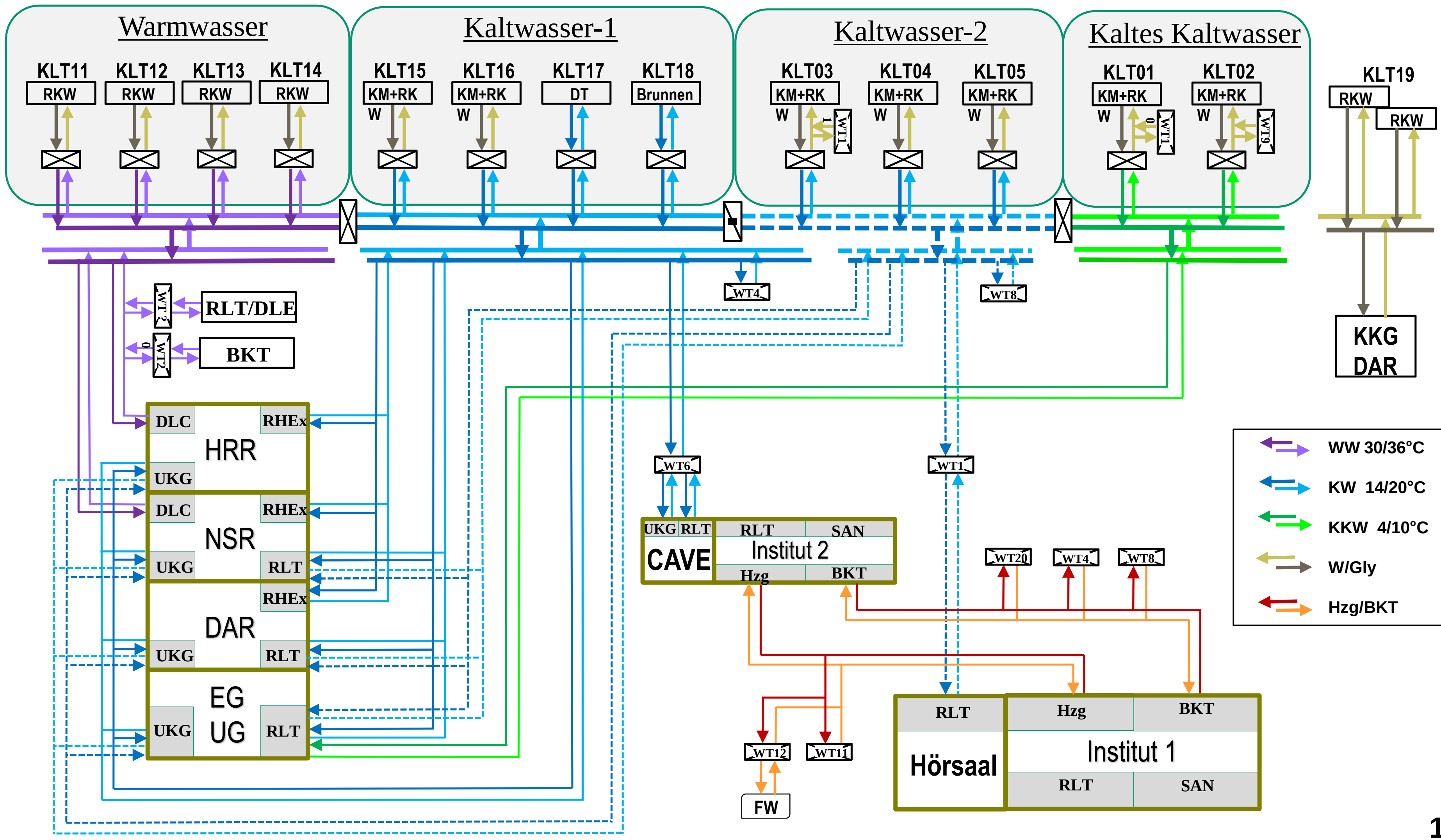
iDataplex DWC Rack w/ water cooled nodes
(front view)

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

LRZ Power Distribution



Layout of cooling Infrastructure



LRZ: Cold Water Distribution Infrastructure

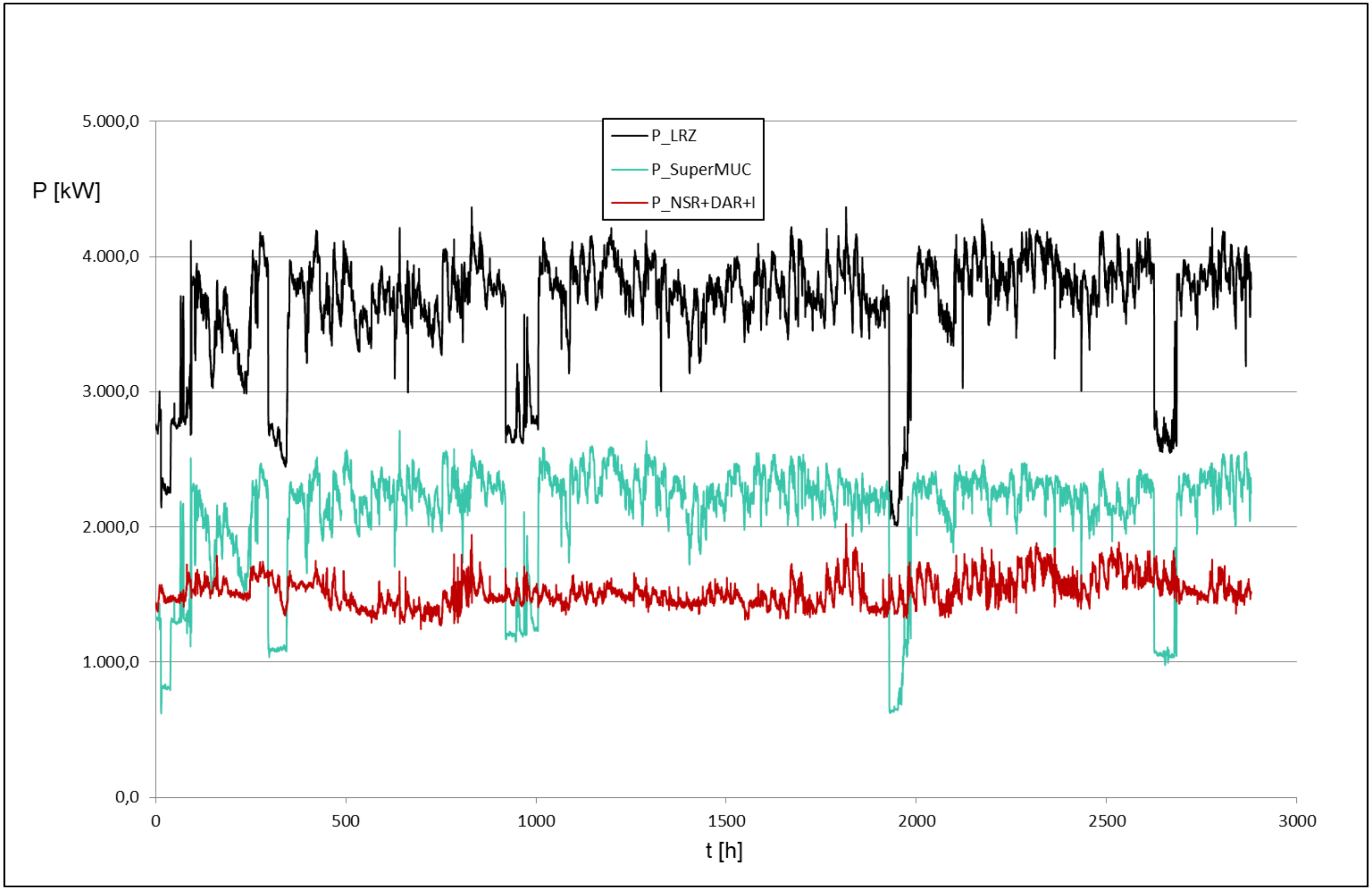
LRZ: Warm Water Distribution Infrastructure

LRZ: Cooling tower infrastructure roof

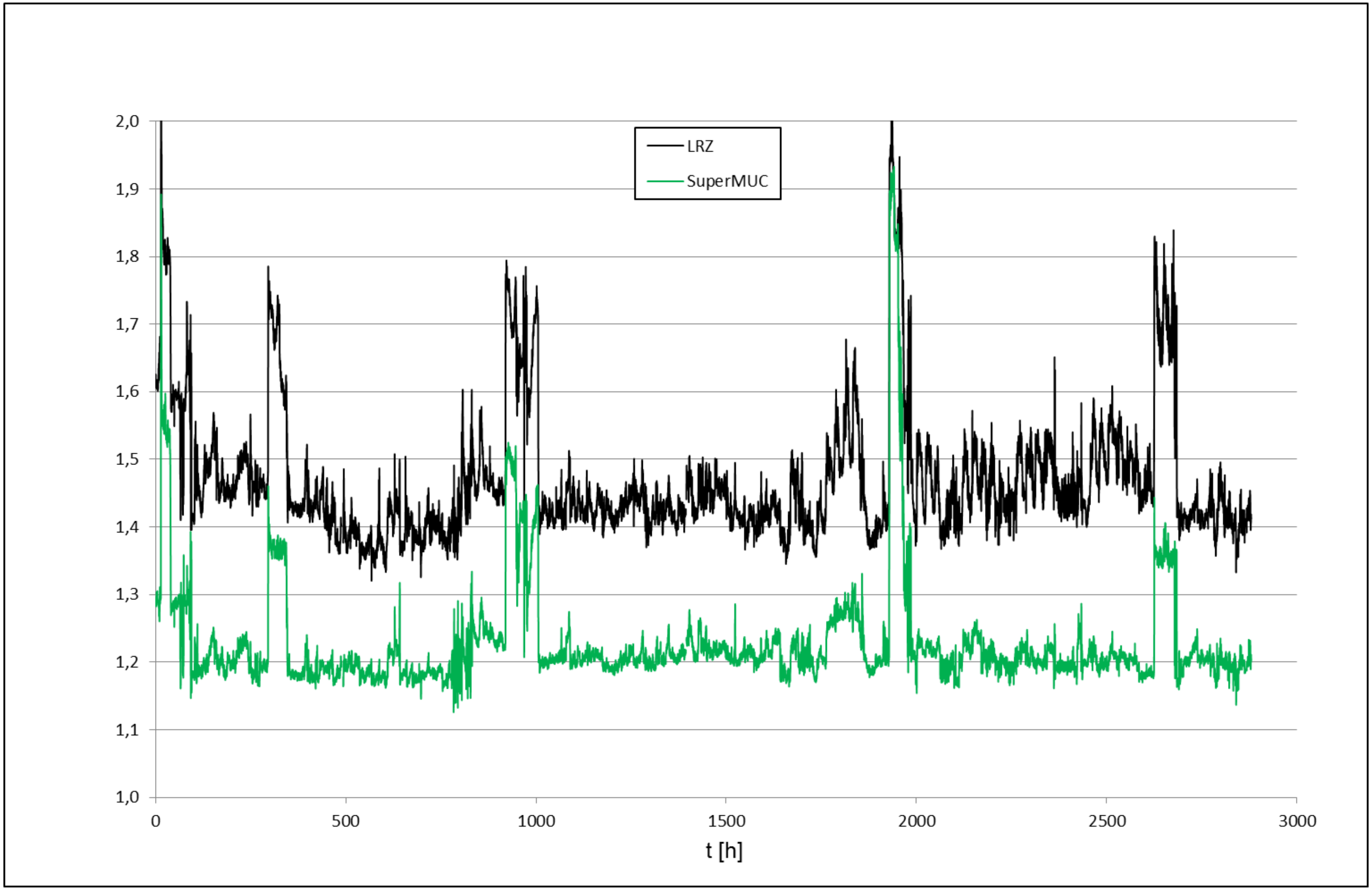
LRZ: Energy Consumption - Some Case Studies

- Three month analysis total power requirement LRZ: Feb to May 2013
- Corresponding Power Usage Effectiveness
- Considering the Cost of Non-Optimized Infrastructure
- Effects of Brown-Outs: Disturbation of the System
  - Infrastructure masters brown-outs
  - Brown-outs cost money for additional infrastructure power
- Coincidence of Brown-out and Failures in the Infrastructure
  - System remains in operation (fault tolerant)
  - Cost for Power and Personel

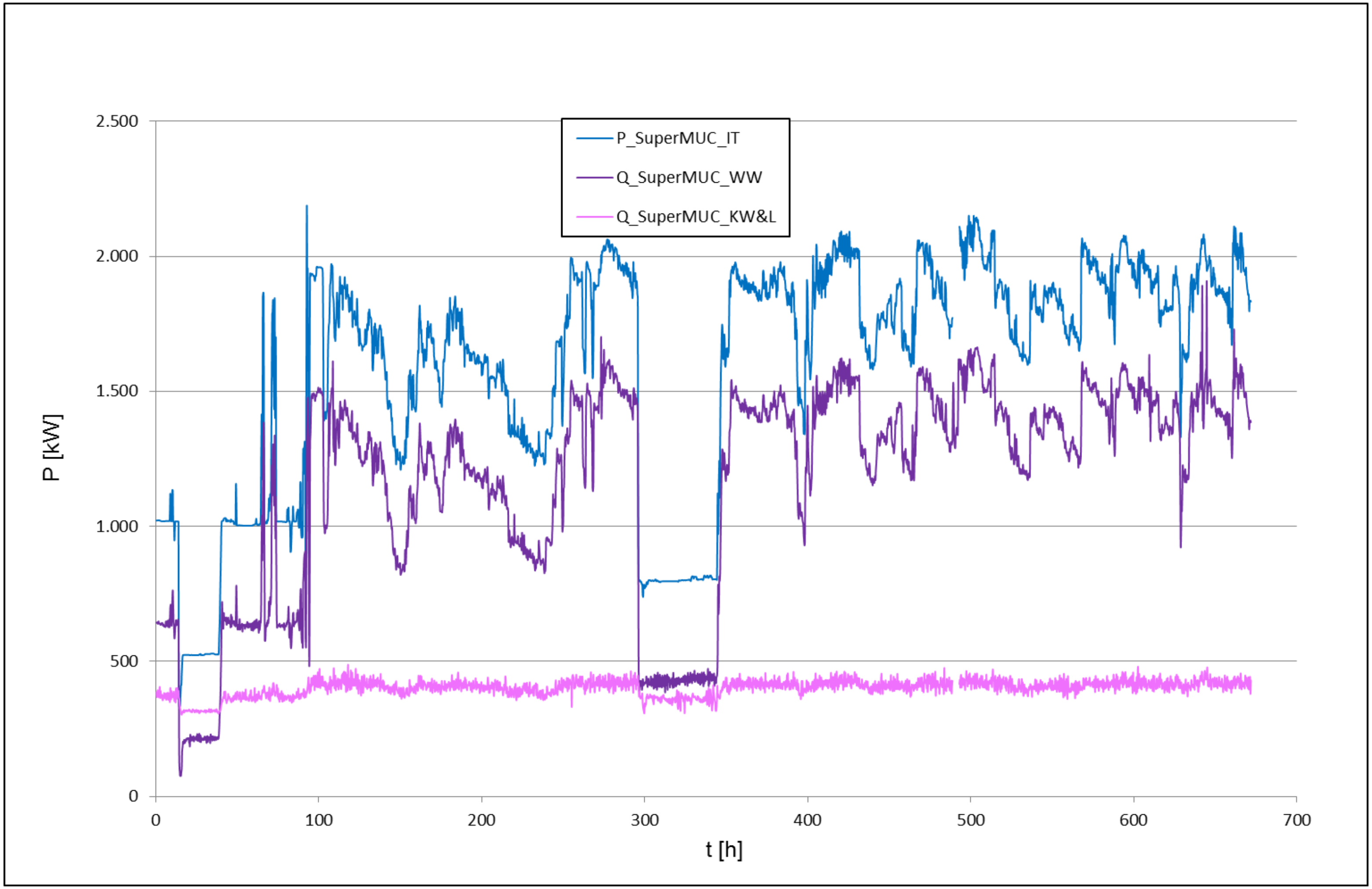
Power Requirements at LRZ (02-05/2013)



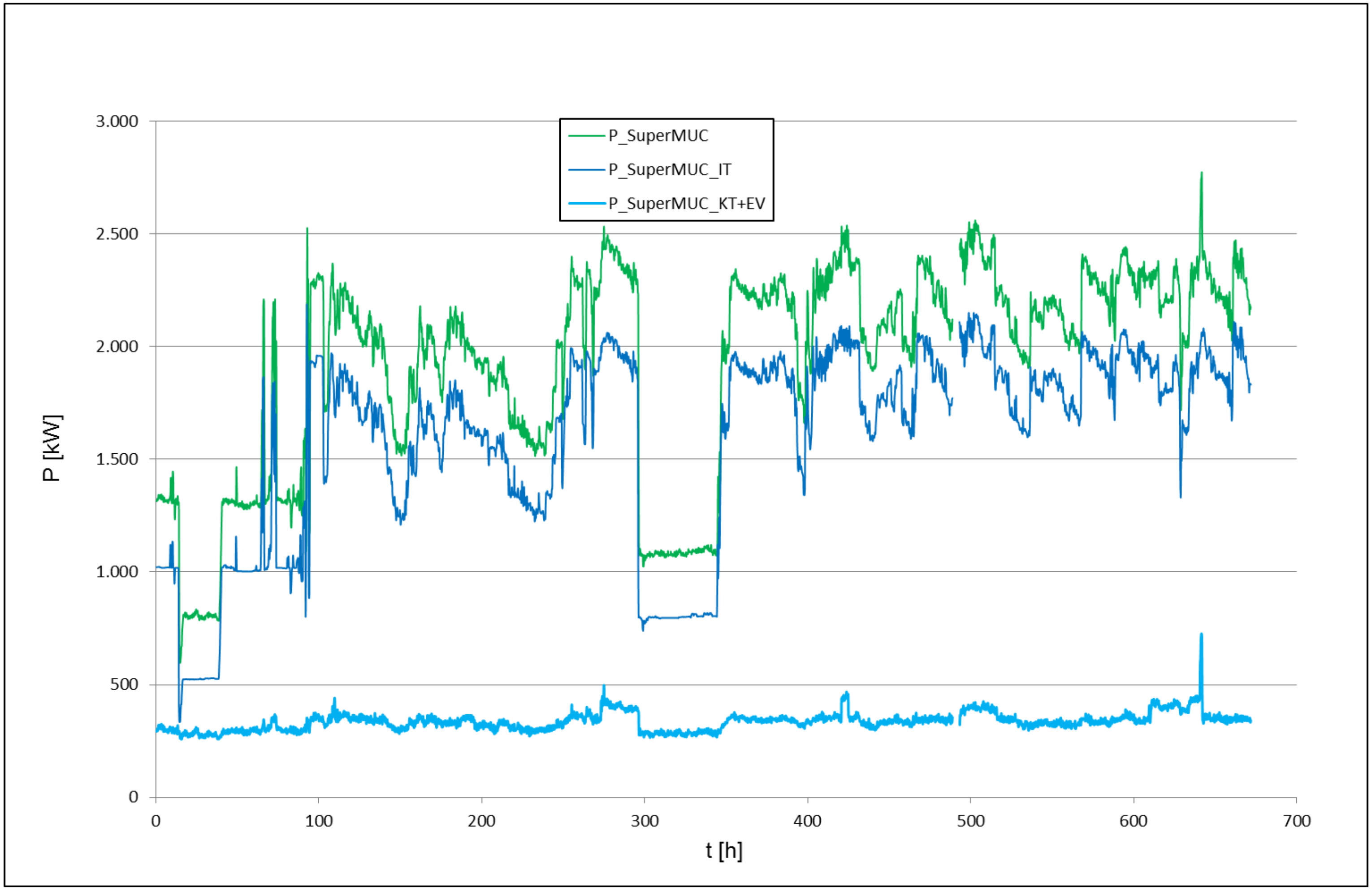
PUE at LRZ (02-05/2013)
Total Facility Power / IT Equipment Power



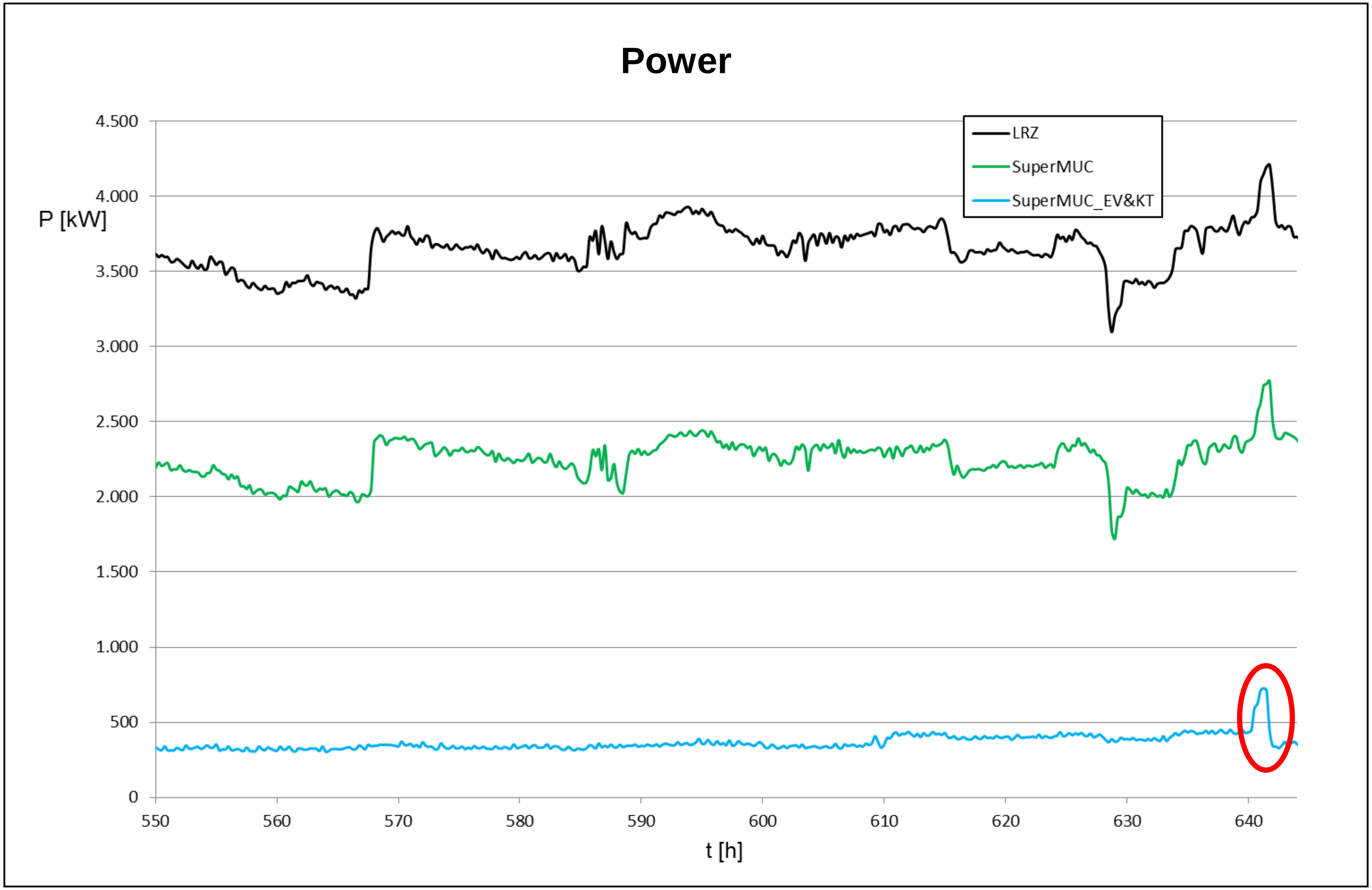
Waste Heat into Cooling Infrastructures by SuperMUC (2-2013)



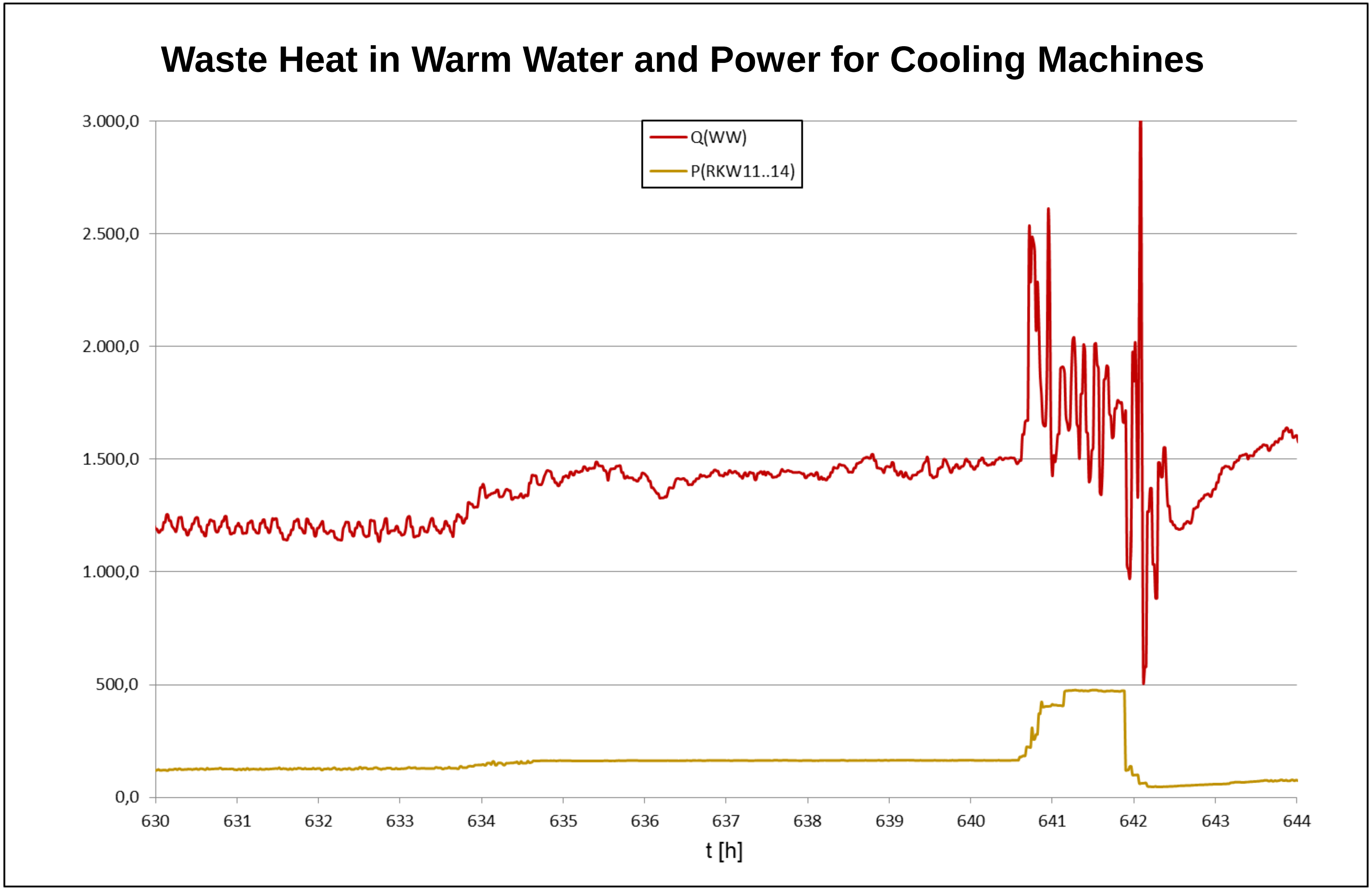
Power Requirement for SuperMUC (2/2013)



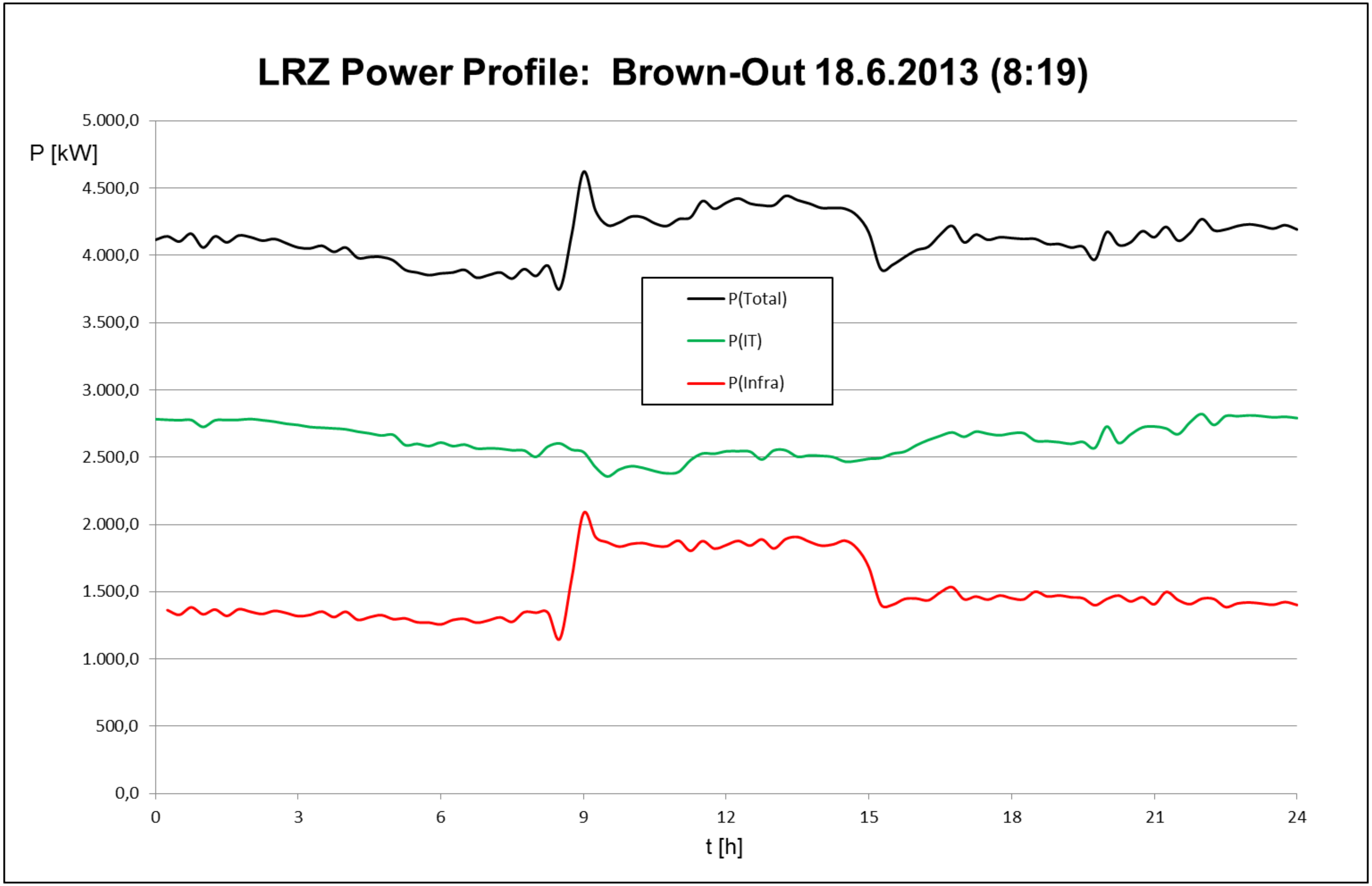
(1) Test Warm Water Cooling Infrastructure on Feb 27, 2013 (ΔT = -20 K, 10 °C instead of 30 °C)



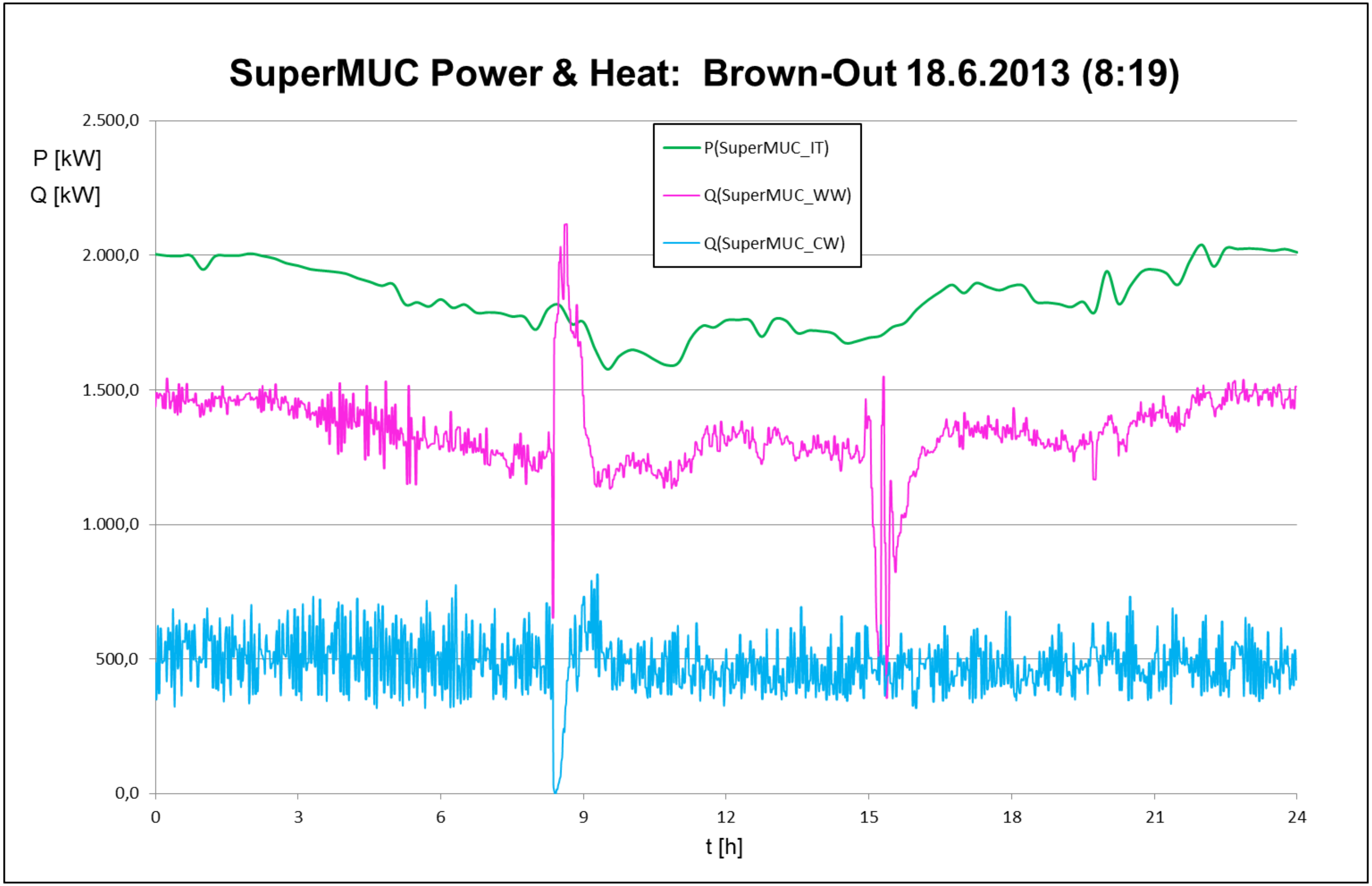
(2) Test Warm Water Cooling Infrastructure on Feb 27, 2013



(1) Brown-Out June 18, 2013



(2) Brown-Out June 18, 2013



Intermediate Summary: the Cooling Infrastructure

- An integrated, heterogenous solution (air-, chilled water -, warm water cooling)
- It works even in presence of external disturbation, internal failures or misregulation (large capacity, of cooling infrastructure, skilled personel)
- It is efficient in absence of disturbations
- Need for Integrated, Fine-Grained Monitoring and Control Tool
  - Today: 3+ databases
- Advantage for LRZ integrated cooling technology as compared to rack-based systems (would require more than 200 cooling towers)

But: We did not yet consider influence of system load

- varying load depending on algorithms, applications, usage strategies
- processor P and C-states
- influence of tools

Energy Efficient HPC, the Whole Picture

- Reduce the power losses in the power supply chain
- Exploit your possibilities for using compressor-less cooling und use energy-efficient cooling technologies (e.g. direct liquid cooling)
- Re-use waste heat of IT systems

Energy efficient infrastructure

- Use newest semiconductor technology
- Use of energy saving processor and memory technologies
- Consider using special hardware or accelerators tailored for solving specific scientific problems or numerical algorithms

Energy efficient hardware

- Monitor the energy consumption of the compute systems and the cooling infrastructure
- Use energy aware system software to exploit the energy saving features of your target platform
- Monitor and optimize the performance of your scientific applications

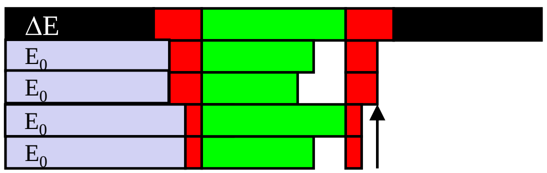
Energy aware software environment

- Use most efficient algorithms
- Use best libraries
- Use most efficient programming paradigm

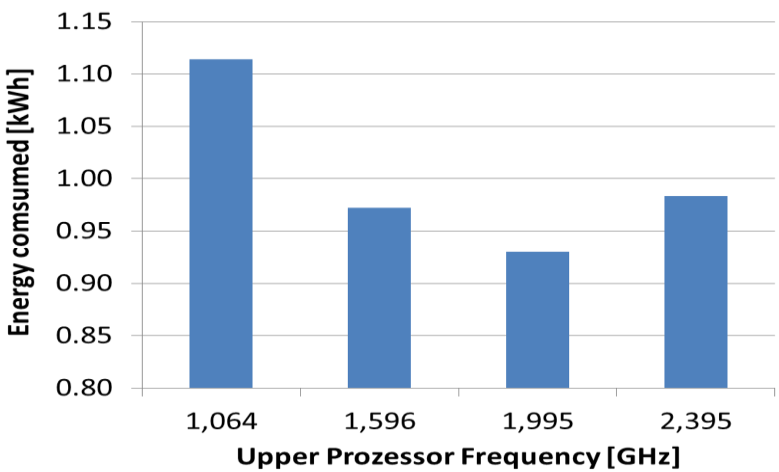
Energy efficient applications

Energy-aware System Software: Minimizing Energy to Solution for Parallel Applications

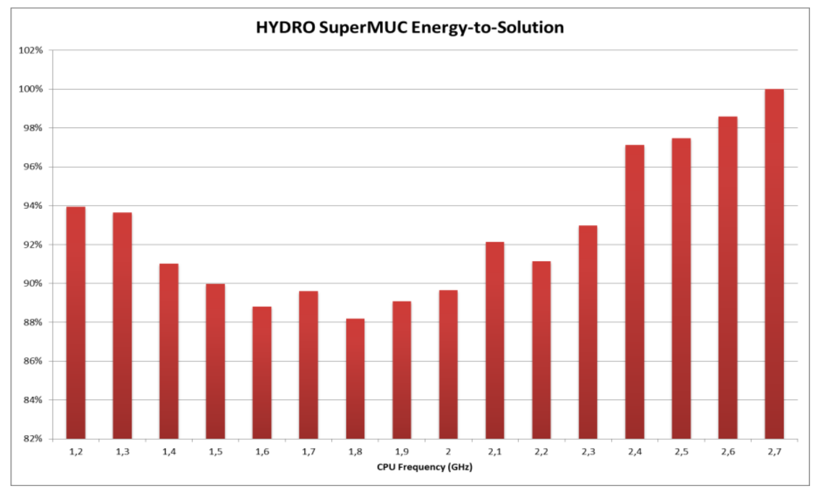
For minimum Energy to Solution: run serial application on low power platform



For minimum Energy to Solution:
Energy Saving due to frequency scaling must be greater than Energy consumed by unused processors in lowest energy state and un-core system components



Example 1: Geophysical Application SeisSol
- 40 E7-4870 cores (one node)
- MPI
- On demand Linux governor

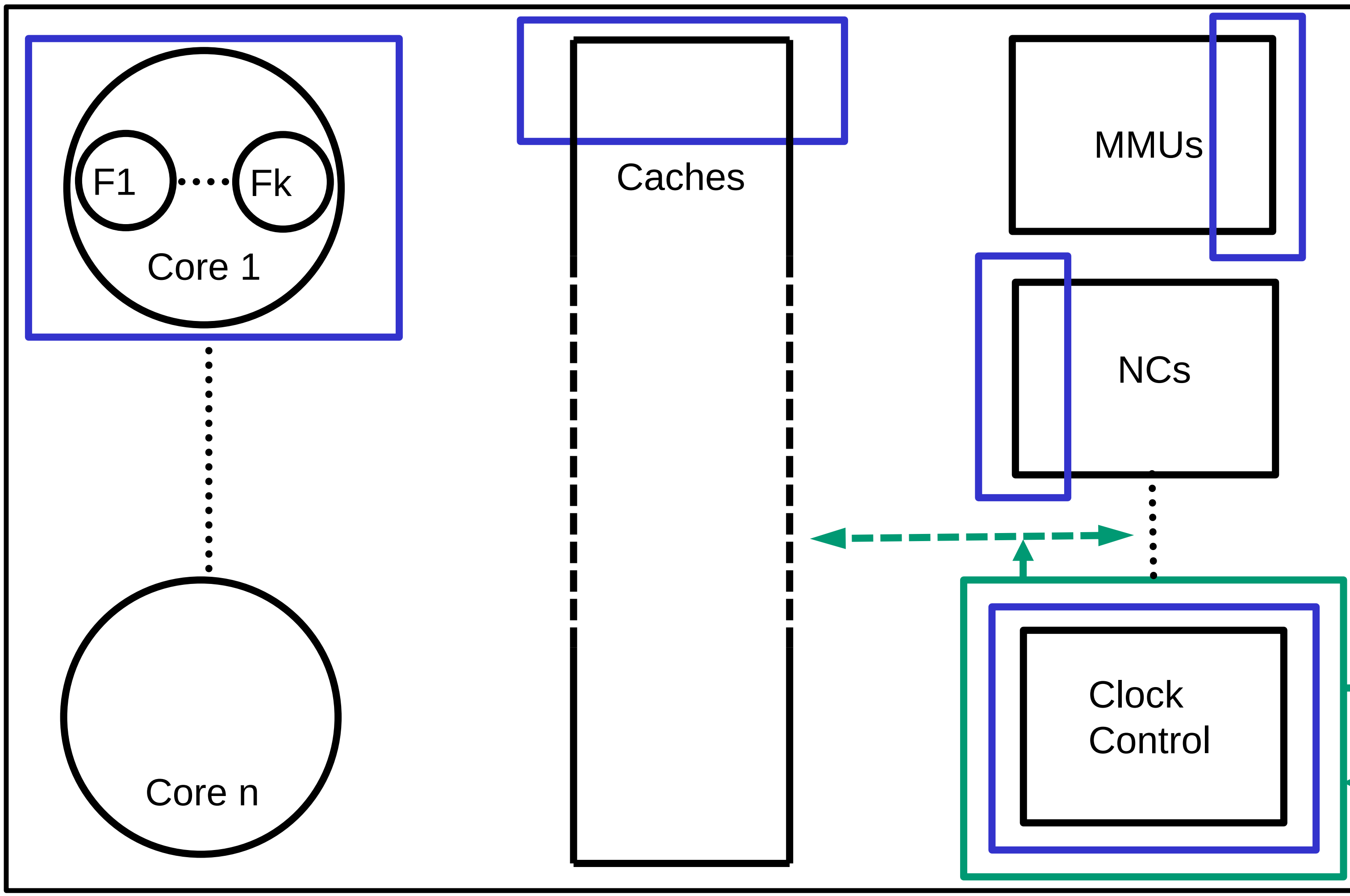


Example 2: CFD Application HYDRO
- 256 Intel E5-2680 cores (16 nodes)
- MPI
- On demand Linux governor

New Roles in Energy to Solution for HPC

- HPC Hw /Sw System Vendor(s):
  - Develop efficient system parts (C-states, P-states)
  - Use best cooling technologies
  - Develop energy to solution tools
- Data Center:
  - Procure appropriate building, infrastructure, HPC-system, energy contracts
  - Optimize „the whole thing“
  - Define usage strategies
- Algorithm and Library Designer:
  - Design best algorithm / library
  - Evaluate energy to solution for different architectures
  - Cooperate with data center usage strategy
- Application End User:
  - Make right choice for target architecture, ISA and algorithm
  - Make use of „energy to solution tools for your program

C- and P-States

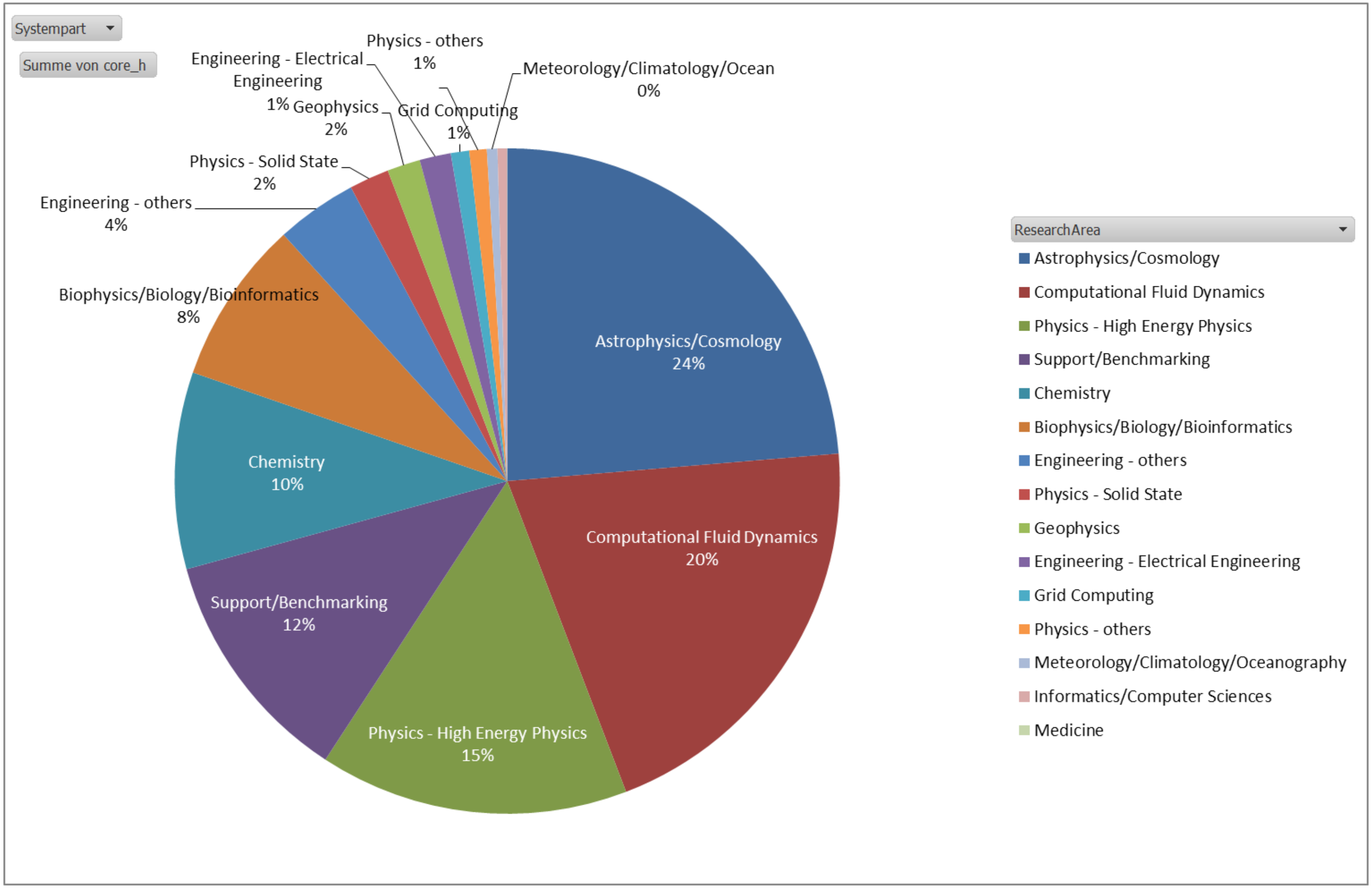


„Microprocessor“

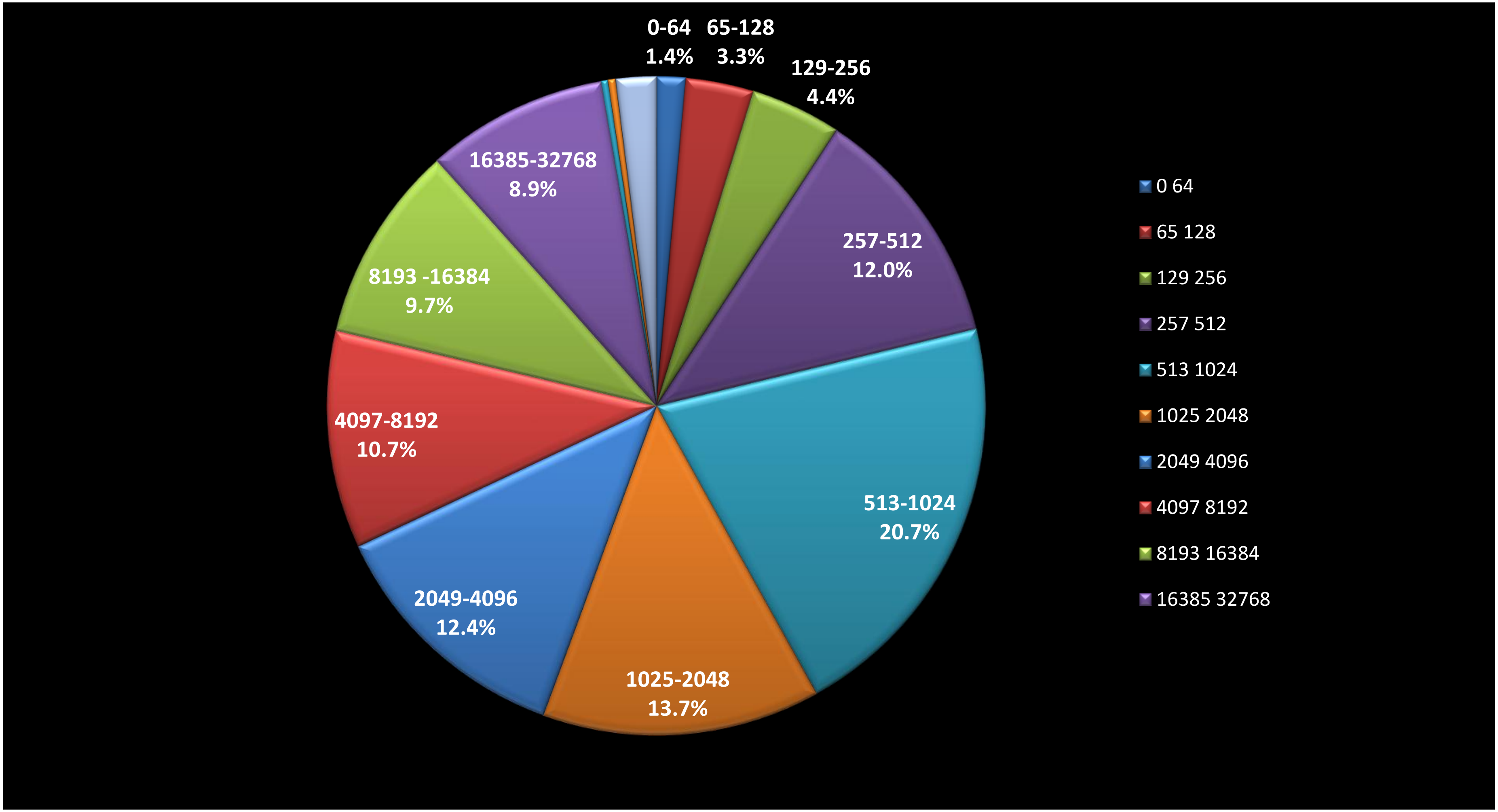
C-State: Describes active part of micro-processor
CO: all active

P-State: Describes clock frequency for actitve parts

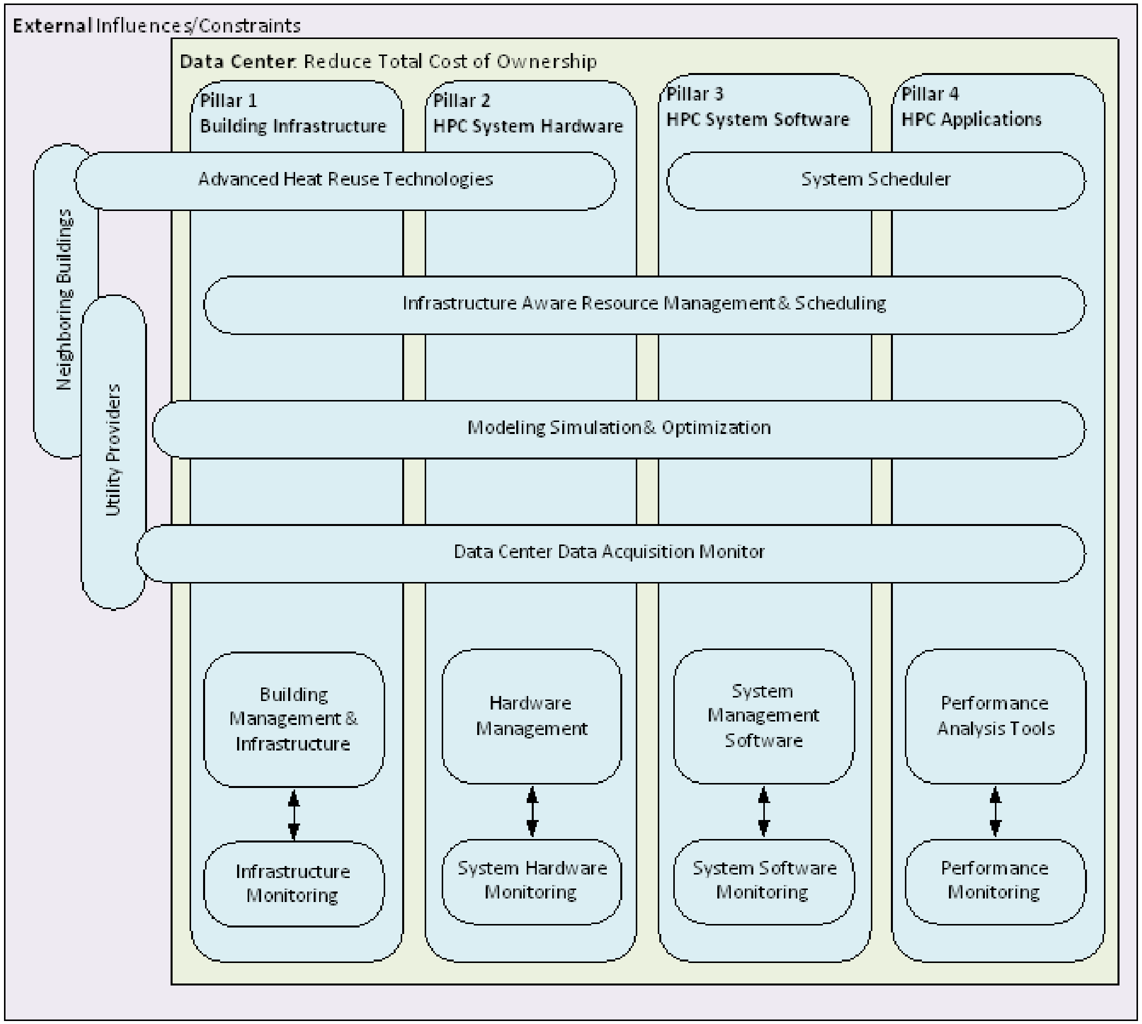
SuperMUC Usage by Research Areas



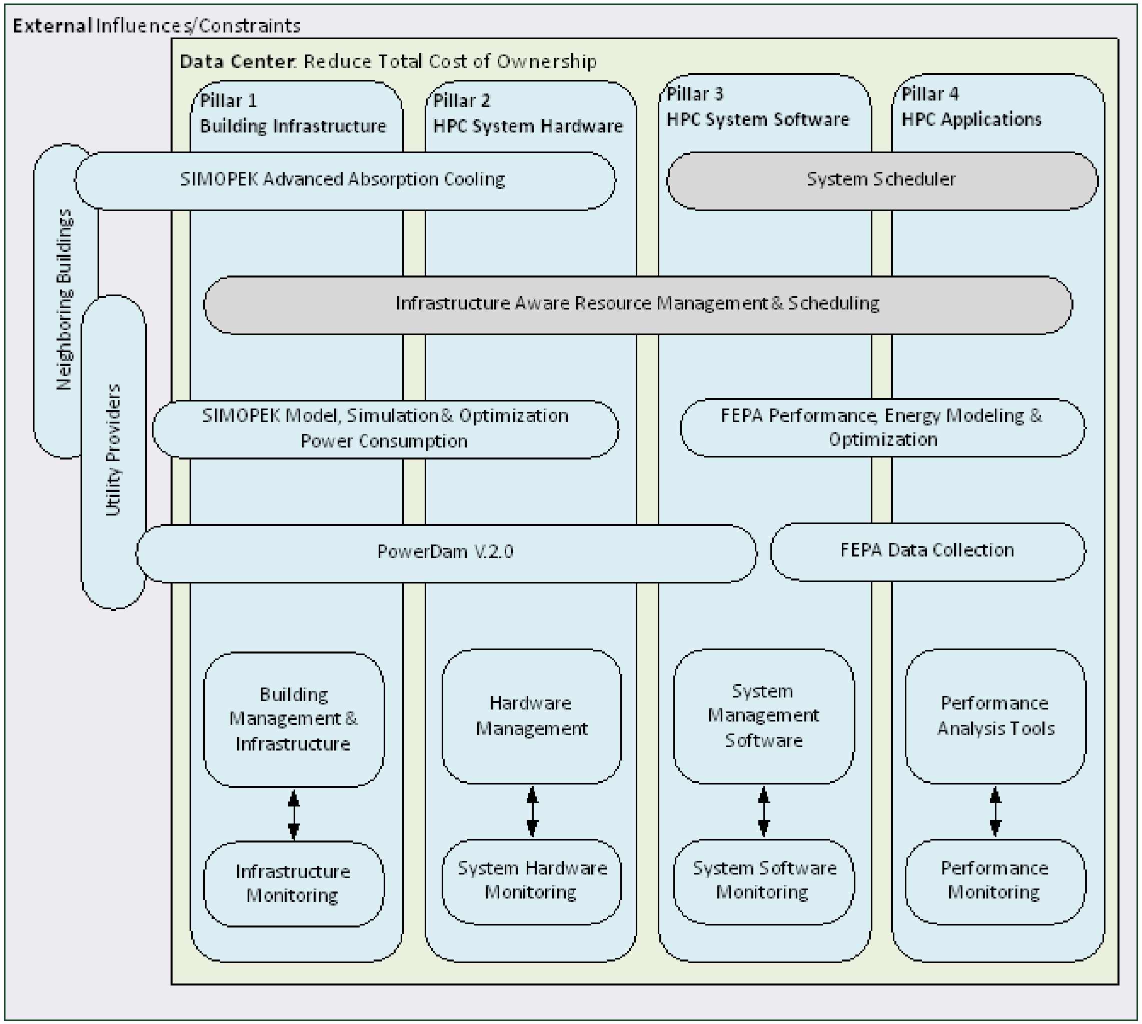
SuperMUC Usage by Jobsize



(1) LRZ: Simopek and PowerDAM



(2) LRZ: Simopek and PowerDAM



SuperMUC next steps

- Universe SuperMUC subcluster: Installation in April / Mai 2013
- SuperMUC Manycore: Autumn 2013, Based on Intel PHI
- SuperMUC Phase 2014/2015: Contract signed, Full system 6.46 PFLOPs
- LRZ in Exascale projects: DEEP/DEEPER (Intel PHI and Xtoll)
Mont-Blanc 1 and 2 (ARM technology)
EESI
- Successor to SuperMUC needs strong support for users:
Scalability issues for the „Mega-core-system“
- New „HPC styles“: Big Data
Realtime HPC
Integrated Visualization
Steering

Energy to Solution - Summary

- TCO for HPC needs to include „Engineering Approach“
- We need
  - Cooperation of all stakeholders (building to algorithm)
  - Codesign (prediction of requirements and parameters)
  - Tools for „optimal compromise“ between application performance and cost
  - Experiments and experience with all sorts of new technologies
- On the basis of today‘s technology, EXASCALE is not affordable:
  ø of TOP_10 systems June 2013 extrapolated to EXASCALE: energy cost in German price around 1.5 B€ p.a.!!!

In cooperation with:

Helmut Breinlinger, Detlef Labrenz, Axel Auweter, Herber Huber, Albert Kirnberger, Torsten Wilde, Jeanette Wilde, Hayk Shoukourian, Reinhold Bader, Matthias Brehm, Werner Baur, Victor Apostolescu, IBM-Team, Intel-Team, YIT-Team, Bauamt München-Team, Herzog und Partner-Team and many other building, infrastructure, system providers and cooperation partners