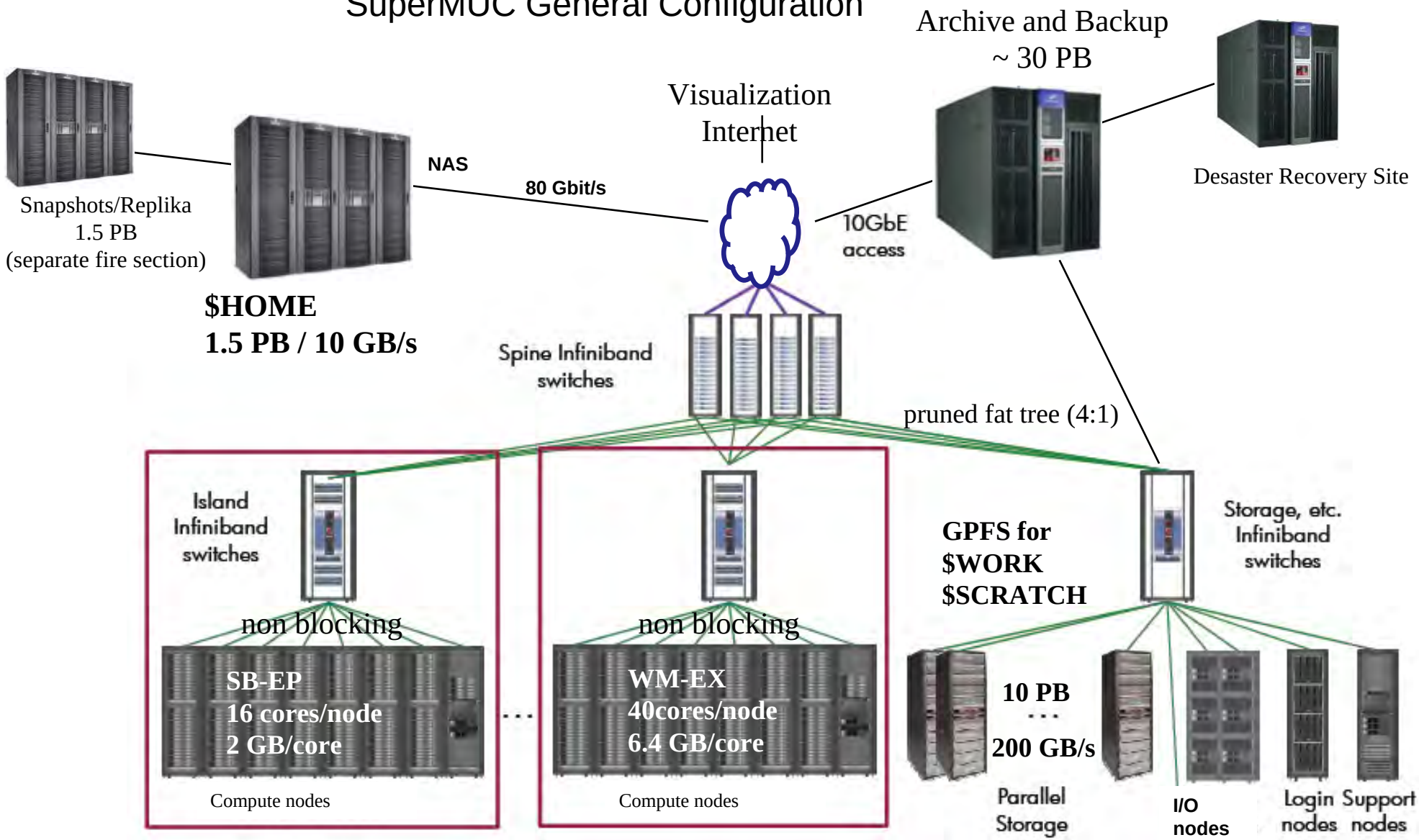
Case Study: LRZ Liquid Cooling, Energy Management, Contract Specialities

Herbert Huber, Axel Auweter, Torsten Wilde, High Performance Computing Group, Leibniz Supercomputing Centre

Ingmar Meijer, Charles Archer, Torsten Bloth, Achim Bömelburg, Steffen Waitz, IBM

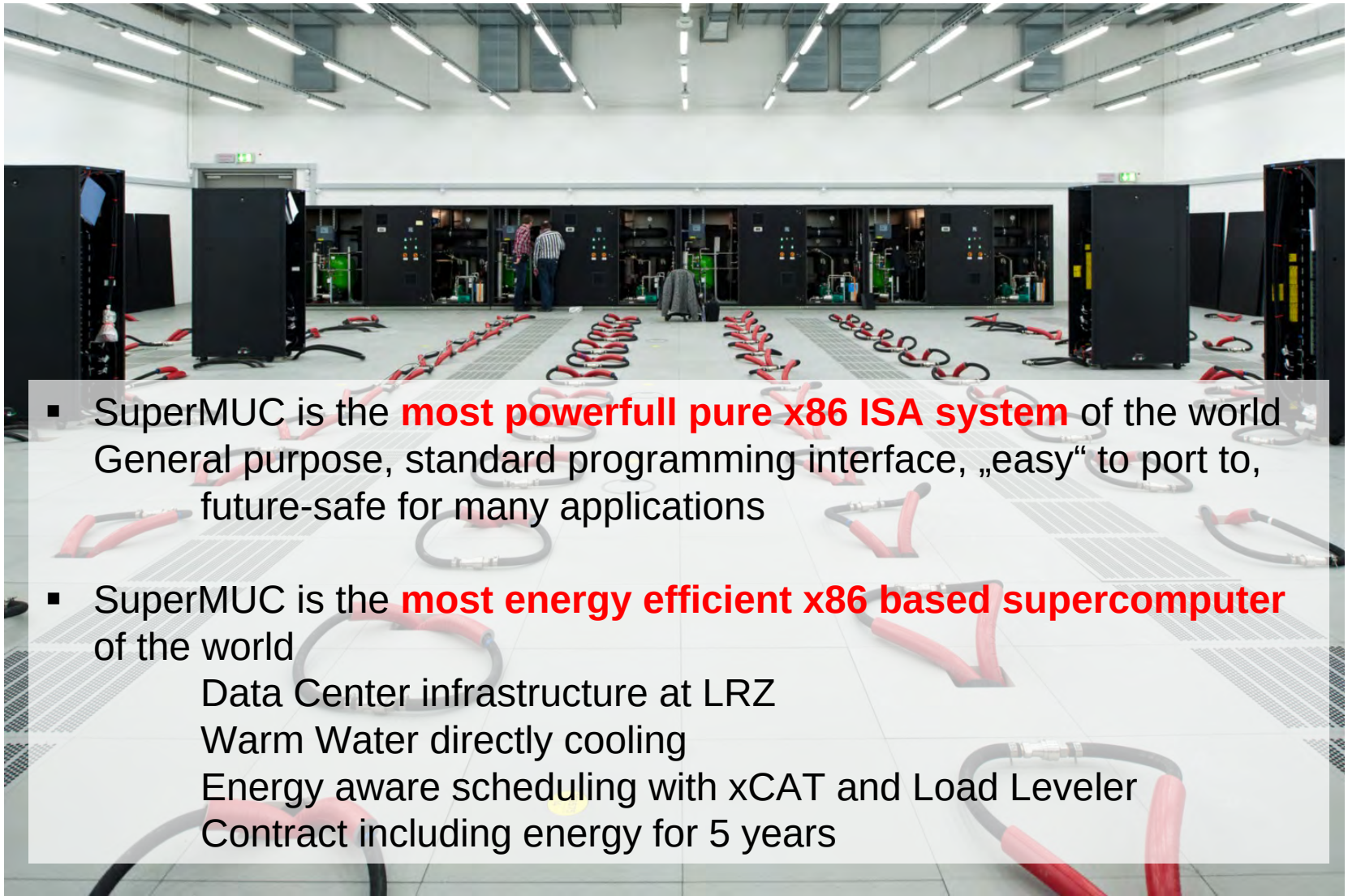
SuperMUC General Configuration



18 Thin node islands
(each 8192 cores)

1 Fat node island
(8200 cores)
also used as Migration System

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
  Data Center infrastructure at LRZ
  Warm Water directly cooling
  Energy aware scheduling with xCAT and Load Leveler
  Contract including energy for 5 years

The LRZ Petascale Power Challenge

Energy Efficiency and SuperMUC

Motivation:

•Academic and governmental institutions in Bavaria use electrical energy from **renewable sources**

•We currently pay 15.8 Cents per KWh

•We already know that we will have to pay at least 17.8 Cents per KWh in 2013

The 4 Pillars of Energy Efficient HPC

Energy Efficient HPC

- 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

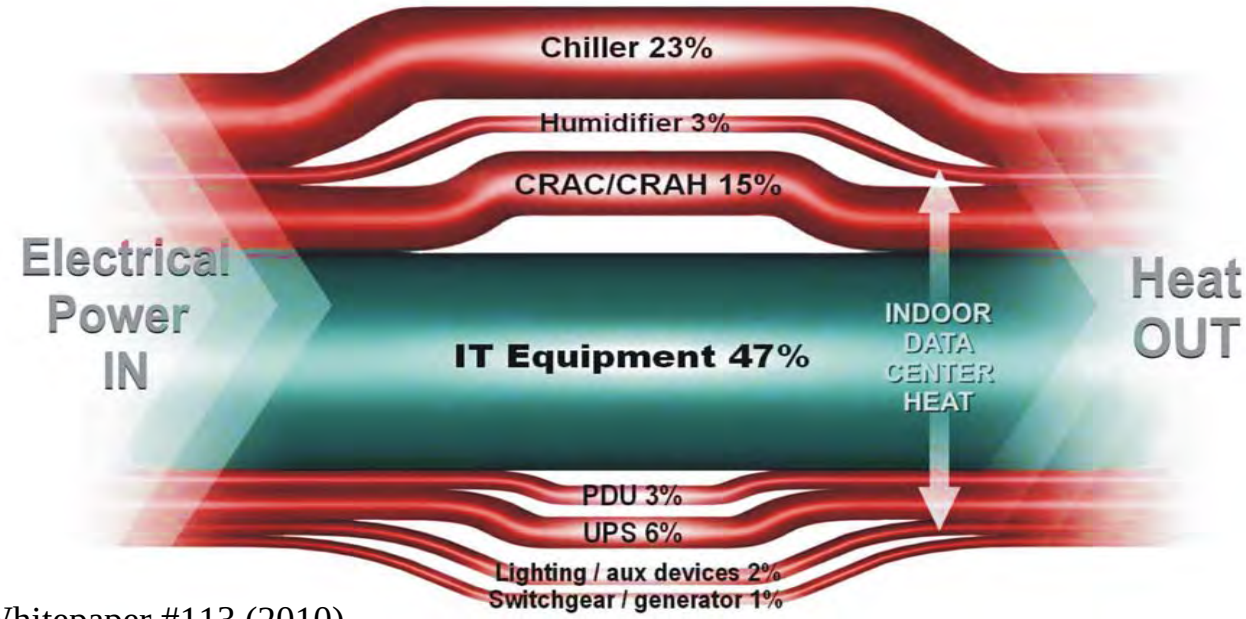
- Monitor the energy consumption of the compute systems and the cooling infrastructure
- Use energy aware system software to exploit the energy saving features of your target platform
- Monitor and optimize the performance of your scientific applications

Energy aware software environment

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

Energy efficient applications

Energy Consumption of Datacenters



APC, Whitepaper #113 (2010)

- **Air-cooled datacenters are inefficient.**
  Typical cooling needs as much energy as IT equipment and both are thrown-away.
- **Provocative: datacenter is a huge "heater with integrated logic."**

A few Words about PUE (1)

- **PUE definition**

$$PUE = \frac{Total\ facility\ power}{IT\ equipment\ power}$$

- **Common ways to fake the masses with very low PUE values**
  - Measure the IT power load at the UPS output
  - → Impact of losses associated with electrical distribution components and non-IT related devices, e.g., rack mounted fans is neglected
  - Use peak IT power load values for PUE calculation
  - → Efficiency values of transformers and UPS systems are optimized
- **Improving the energy efficiency only on the IT side leads to higher PUE values**

A few Words about PUE (2)

- **PUE values**
  - Chiller-assisted cooling:
    - measured annual LRZ PUE of 1.4 (no cold isle enclosures)
    - PUE of 1.3 very common for data centres with strict cold isle or hot isle enclosures
    - PUE of 1.22 in principle possible but very hard to reach in Germany!
  - Free cooling
    - PUE value close to 1.1
    - PUE values below 1.1 in principle possible but very hard to reach because of energy losses in power transformers, UPS systems, power bars, etc.
- **Total annual LRZ operation costs including cooling**
  - Chiller-assisted cooling (PUE of 1.3): 1799,30 €/kWy
  - Free cooling (PUE of 1.1): 1522,50 €/kWy

  → Annual savings of 276,80 € per kW of IT load

Direct Water Cooling

Air Cooling versus Direct Liquid Cooling

	Air	Water
Thermal conductivity [J/(m*K*s)	0,026	0,598
Volumetric heat capacity [J/(m3 * K)]	1213	4174472
Thermal inertia [J/(m^2 * K * $s^{1/2}$)]	5,09	1579,98

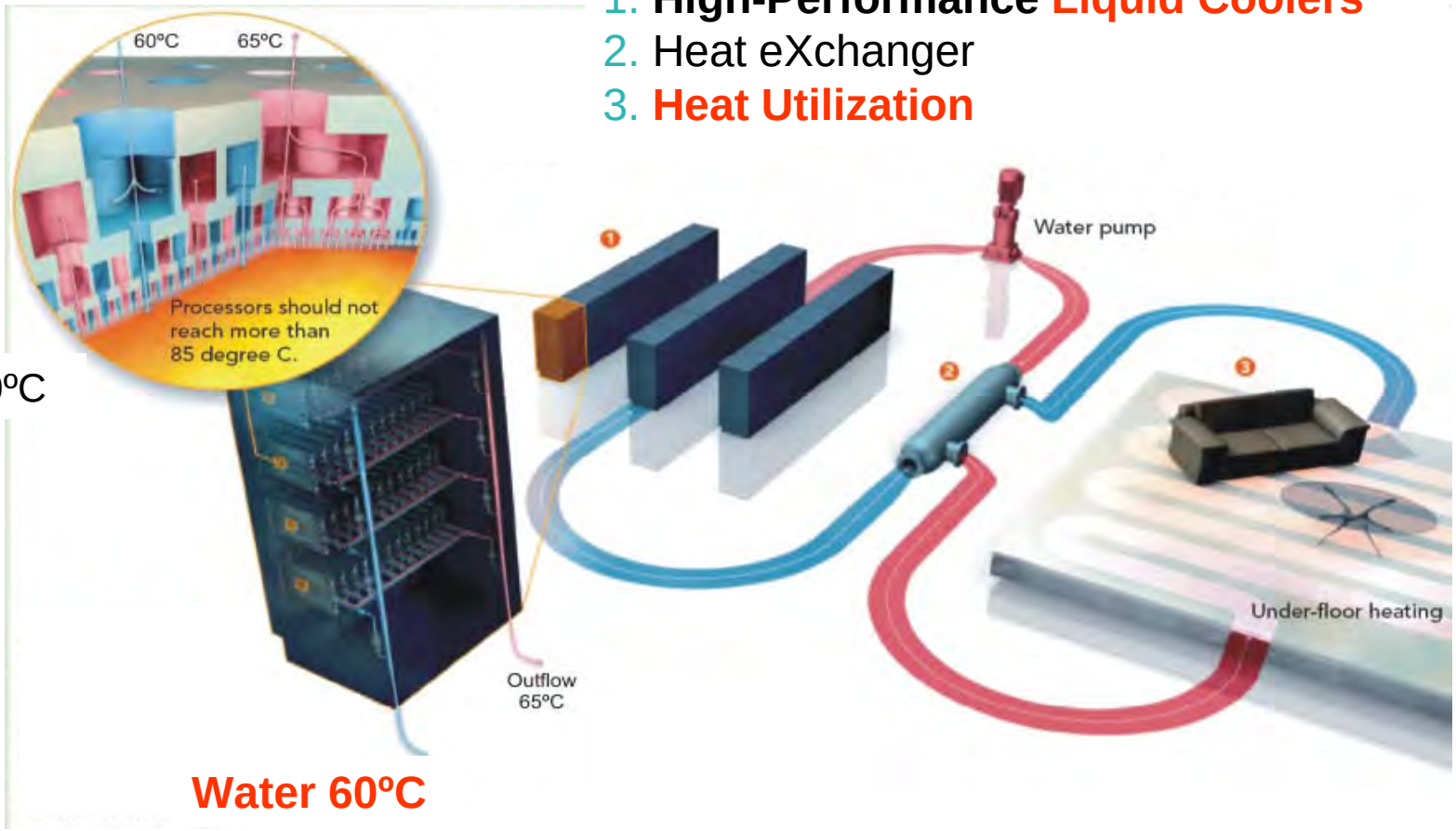
The thermal inertia or cooling efficiency of water is more than 300 times higher than the cooling efficiency of air

Automotive industry uses water cooling since decades for high power engines

Hotwater Cooling Concept

1. **High-Performance Liquid Coolers**
2. Heat eXchanger
3. **Heat Utilization**

CMOS 80ºC

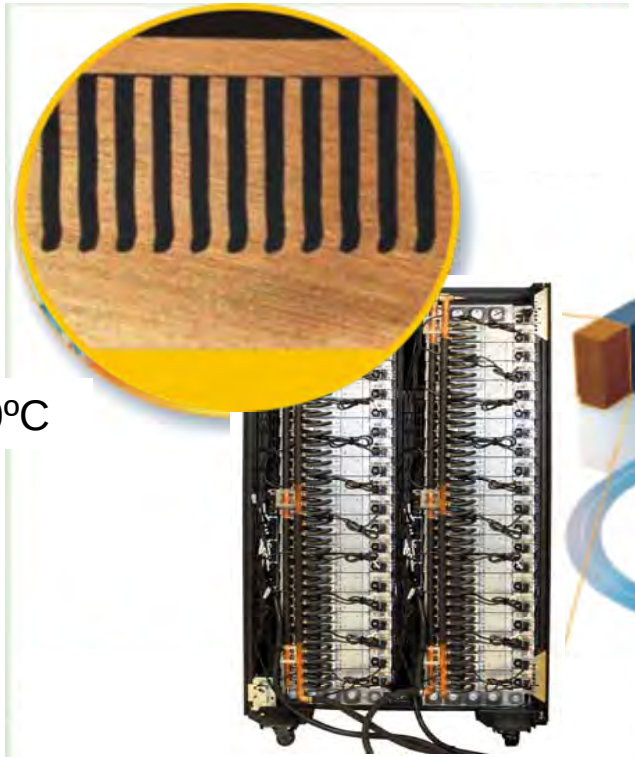


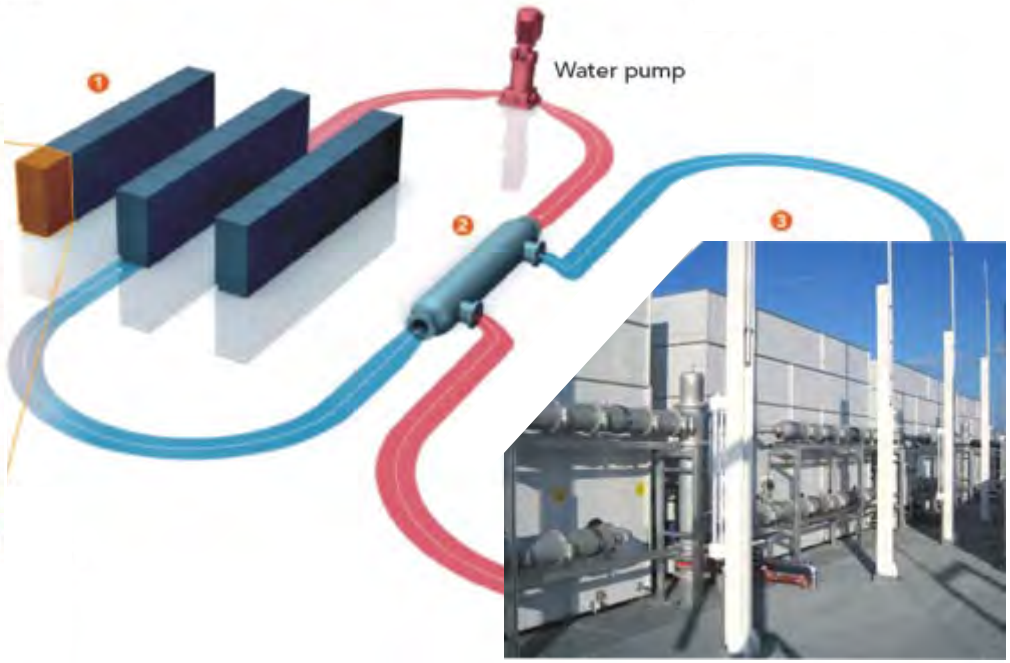
Water 60ºC

Hotwater Cooling Concept SuperMUC

1. **Cost-Performance Liquid Coolers**
2. Heat eXchanger
3. **Free Cooling**

CMOS 50ºC



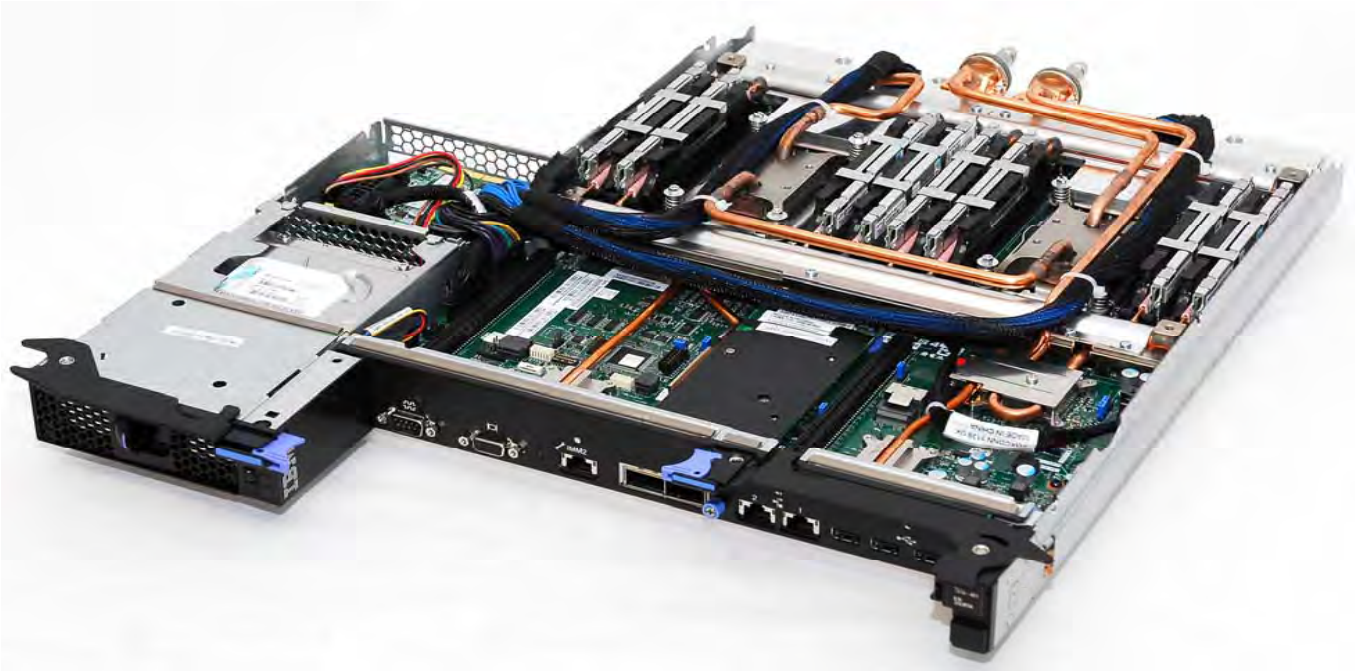


Water 30ºC

Hybrid Datacenter w/ Direct Water Cooled Nodes

- Highly energy-efficient **hybrid-cooling** solution:
  - Compute racks
    - 90% Heat flux to warm water
    - 10% Heat flux to CRAH
  - Switch / Storage racks
    - Rear door heat exchangers
- Compute node **power consumption reduced ~ 10%** due to lower component temperatures and no fans.
- Power Usage Effectiveness P_{Total} / P_{IT}: **PUE ~ 1.1**
- **Heat recovery** is enabled by the compute node design:
  Energy Reuse Effectiveness $(P_{Total} – P_{Reuse}) / P_{IT}$: **ERE ~ 0.3**

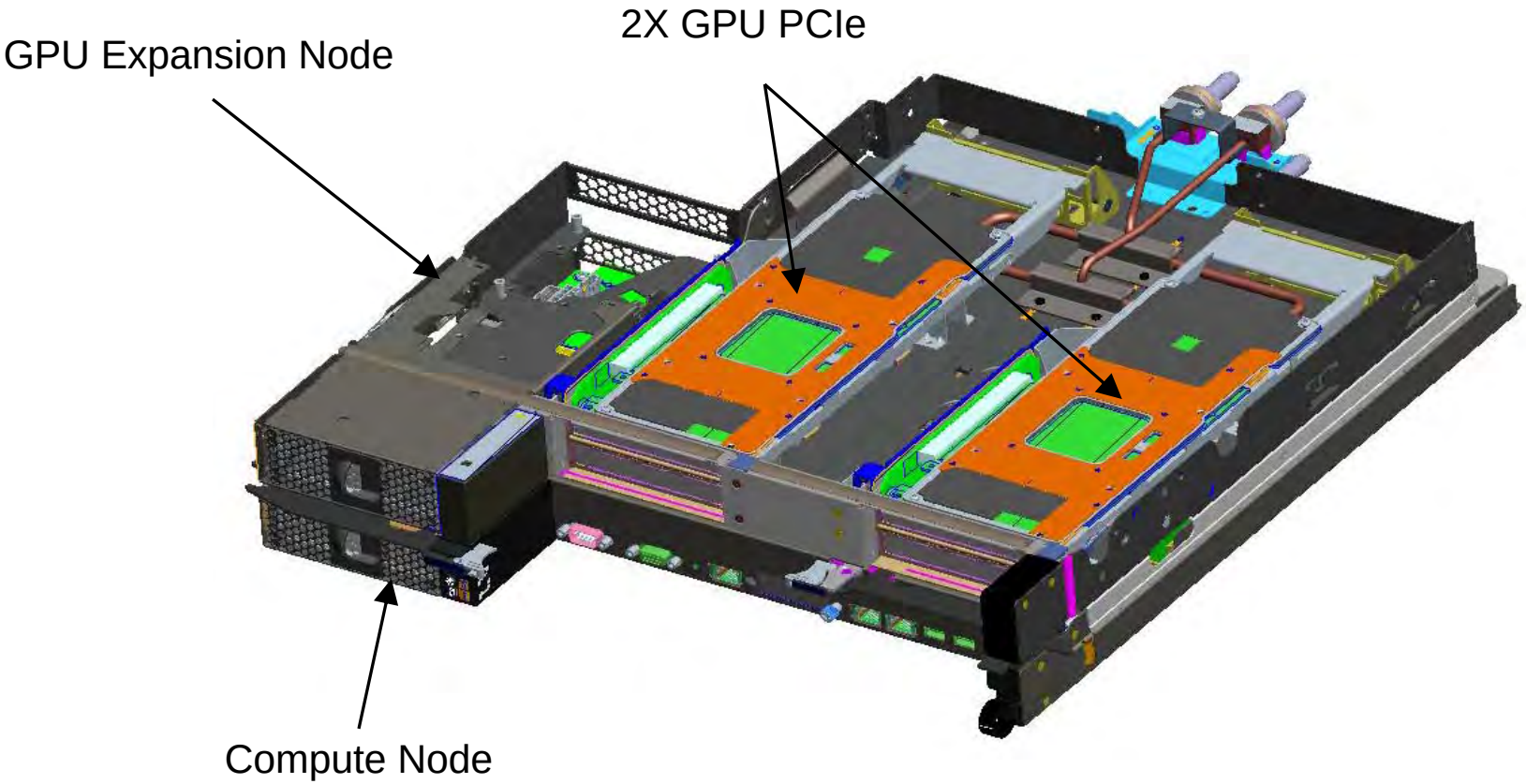
IBM System x iDataPlex Direct Water Cooled dx360 M4



- 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.

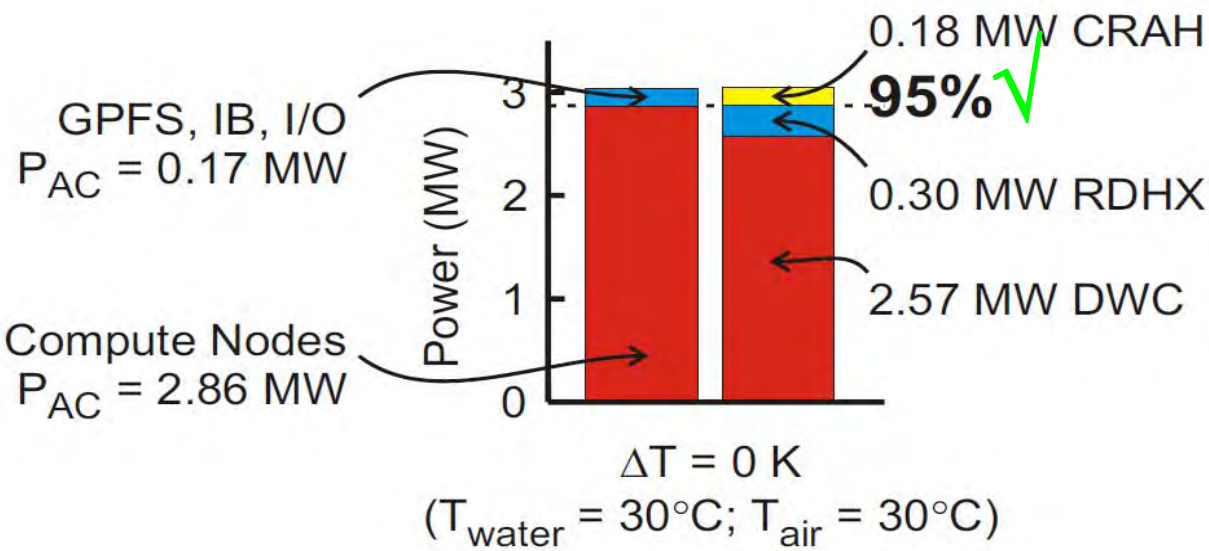
Intel MIC Expansion Node

Cooling warm water (**efficiency**) or
chilled water (**performance**)



SuperMUC: Requirement 90% Heat to Water

IBM System x iDataPlex Direct Water Cooled dx360 M4
2x Intel SB-EP 2.7 GHz (Turbo OFF). 8x Samsung 4 GB.



LRZ Warm Water Cooling Infrastructure 3rd Floor

Warm water distribution pipes

Hydraulic gates

LRZ Warm Water Cooling Infrastructure 3rd Floor

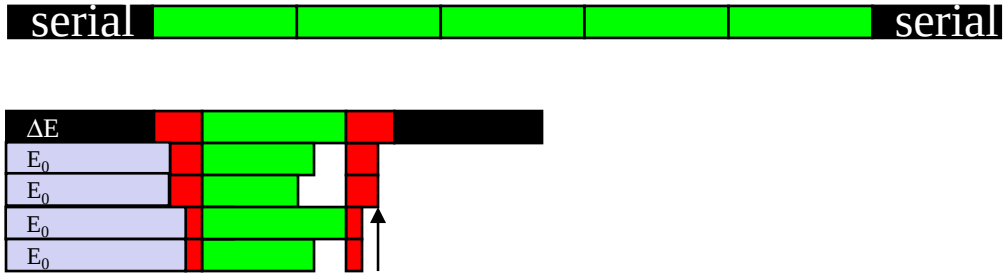
Warm Water Cooling Infrastructure on the Roof: Four 2 MW Cooling Towers

SuperMUC: Energy-aware System Software

Energy- and Topology-aware Scheduling

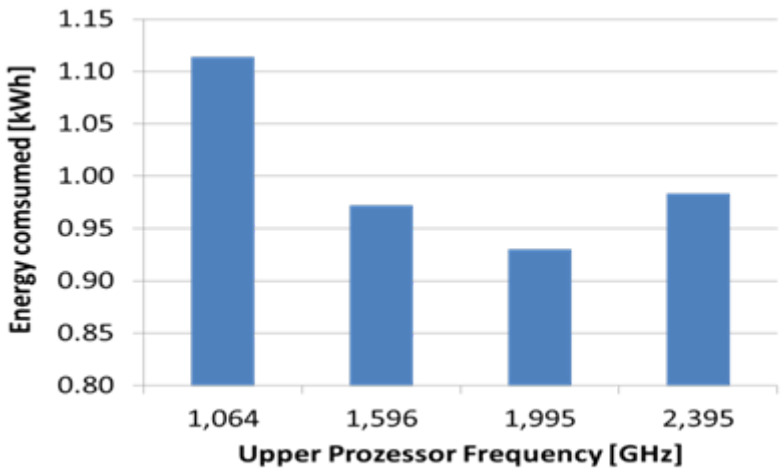
- **Energy- and topology-aware resource manager**
- ✓ Turn unused nodes into deep sleep mode (S3)
  - ❖ Enhancement to IBM LoadLeveler
- ✓ Optimize job energy efficiency by automated frequency scaling
  - ❖ Enhancement to LoadLeveler
  - ❖ Subject of collaboration between IBM and LRZ
- ✓ Optimal placement of MPI tasks on IB network topology

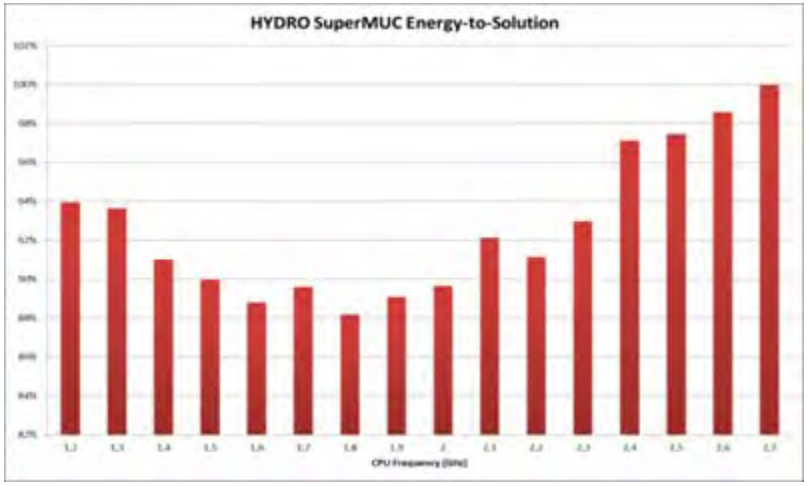
Some first results in "Minimizing Energy to Solution for Parallel Applications"



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 govenor

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

SuperMUC Contract Specialities

SuperMUC Budget

SuperMUC:	2012-2017 Phase 1
Peak Performance	3 PF
Investment Costs (Hardware and Software)	~48 Mio €
Operating Costs (Electricity costs and maintenance for hardware und software, some additional personnel)	~35 Mio €
SUM (not one penny more!)	83 Mio €
Extension Buildings (construction and Infrastructure)	49 Mio €

Procurement Method: Competitive Dialog (1)

Activity	Start	End
First Phase of Dialogues with five competitors	**Wed 10.03.10**	**Wed 28.07.10**
First round of negotiations (clarifications, questions)	**Wed 10.03.10**	**Fri 19.03.10**
Second round of negotiations (economical and technical concept)	**Mon 12.04.10**	**Fri 23.04.10**
Third round of negotiations (economically and technically perfected proposal)	**Mon 10.05.10**	**Fri 21.05.10**
Preparation of procurement documents by LRZ	**Fri 28.05.10**	**Mon 14.06.10**
Posting of procurement documents to competitors	**Mon 14.06.10**	**Mon 14.06.10**
Preparation of proposals by the competitors	**Wed 16.06.10**	**Wed 14.07.10**
Evaluation of the proposals by LRZ	**Thu 15.07.10**	**Tue 27.07.10**
Shortlist: selection of two competitors for the second phase of dialogues	**Wed 28.07.10**	**Wed 28.07.10**

Procurement Method: Competitive Dialog (2)

Activity	Start	End
Second Phase of Dialogues	Thu 29.07.10	Fri 10.09.10
Preparation of the final procurement documents by LRZ	Mon 13.09.10	Thu 23.09.10
Preparation of the final proposals by the competitors	Fri 24.09.10	Wed 27.10.10
Delivery deadline for final proposals	Thu 28.10.10	13:00 h
Evaluation of the final proposals by LRZ	Thu 28.10.10	Fri 12.11.10
Preparation of Contract	Mon 15.11.10	Mon 13.12.10

Summary

- 9 months duration
- 16 full day meetings with vendors

SuperMUC Procurement & Incentives for Energy and Cooling Efficiency

- **Incentives for Energy and Cooling Efficiency**
  - Supercomputer procurement on basis of Total Cost of Ownership; budget includes
    - Investment costs
    - Maintenance costs
    - Operation costs over 5 years duration (including cooling!)
  - Cooling costs (electrical energy needed to generate 1 m^3 of cooling water)
    - 0,36 kWh per m^3 chiller-free cooled water (20°C – 40°C)
    - 1,46 kWh per m^3 chilled water (14°C)
- **Up-evaluate highly cooling effective solutions in the evaluation of the bids (maximum up-evaluation of 10%)**
- **Tell the vendors that you are interested in such kind of solutions**

Summary

Summary / Leasons learned

- **Tell the vendors that you are really interested in buying a highly energy and cooling efficient system**
- **In order to really improve energy efficiency of your total data centre you need an integrated approach!**
- **Huge potential for energy savings (~ 40%) by means of**
  - **Direct Warm Water cooling allowing for Fee Cooling all year around**
  - **Energy-aware Job Scheduling**
    - Turn unused nodes into deep sleep mode
    - Run Application at optimal clock rate
  - **Energy-optimized data centre infrastructure and building automation software**
- **Direct liquid cooled IT hardware is very stable**