MEGWARE HPC Cluster am LRZ – eine mehr als 12-jährige Zusammenarbeit

Prof. Dieter Kranzlmüller (LRZ)

lrz Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

LRZ HPC-Systems at the End of the UNIX-Era (Years 2000-2002)

German national supercomputer Hitachi SR800 pseudo vector system with

- 168 SMP nodes
- 8 +1 CPUs per node
- 1376 GB memory
- 5000 GB disk
- 2016 GF peak performance

Bavarian vector computer Fujtsu VPP vector system with

- 52 vector CPUs
- 104 GB memory
- 1214 GB disk
- 114.4 GF peak performance

LRZ HPC-Systems at the End of the UNIX-Era (Years 1999-2002) #2

Bavarian large shared memory HPC system IBM p690 with

- 8 Power 4 CPUs
- 32 GB memory
- 936 GB disk
- 42 GF peak performance

Bavarian MPP system IBM SP2 with

- 77 nodes
- 16.7 GB memory
- 334 GB disk
- 20.7 GF peak performance

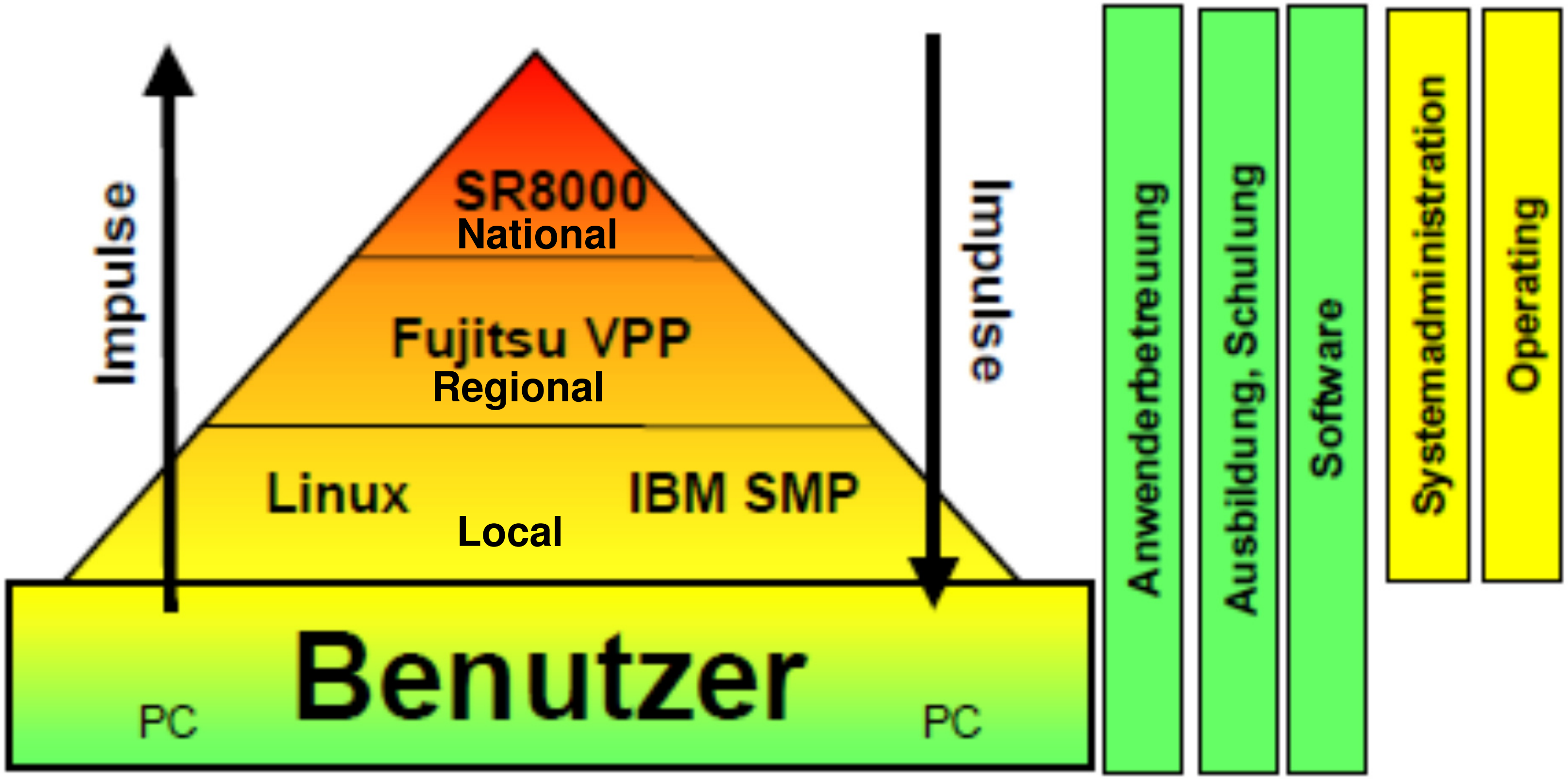
Bavarian vector computer CRAY T90 with

- 4 vector CPUs
- 1.0 GB memory
- 145 GB disk
- 7.2 GF peak performance
- abandonment in 2001

LRZ HPC-Systems at the End of the UNIX-Era (Years 1999-2002) #3

- LRZ home-made Linux Cluster for Munich Universities:
  - 2 dual Pentium II nodes
  - 17 dual Pentium III nodes (9 nodes with Myrinet communication network)
  - 2 quad Pentium III-Xeon nodes
  - 6 Pentium IV nodes
  - 56 GB memory
  - 70 GB disk
  - 62 GF peak performance
  - Vendors: FMS, DELL and Synchron

The LRZ HPC Pyramid as HPC Service Concept



2003: Replacement of IBM SP2 by MEGWARE IA32 and IA64 Linux Cluster

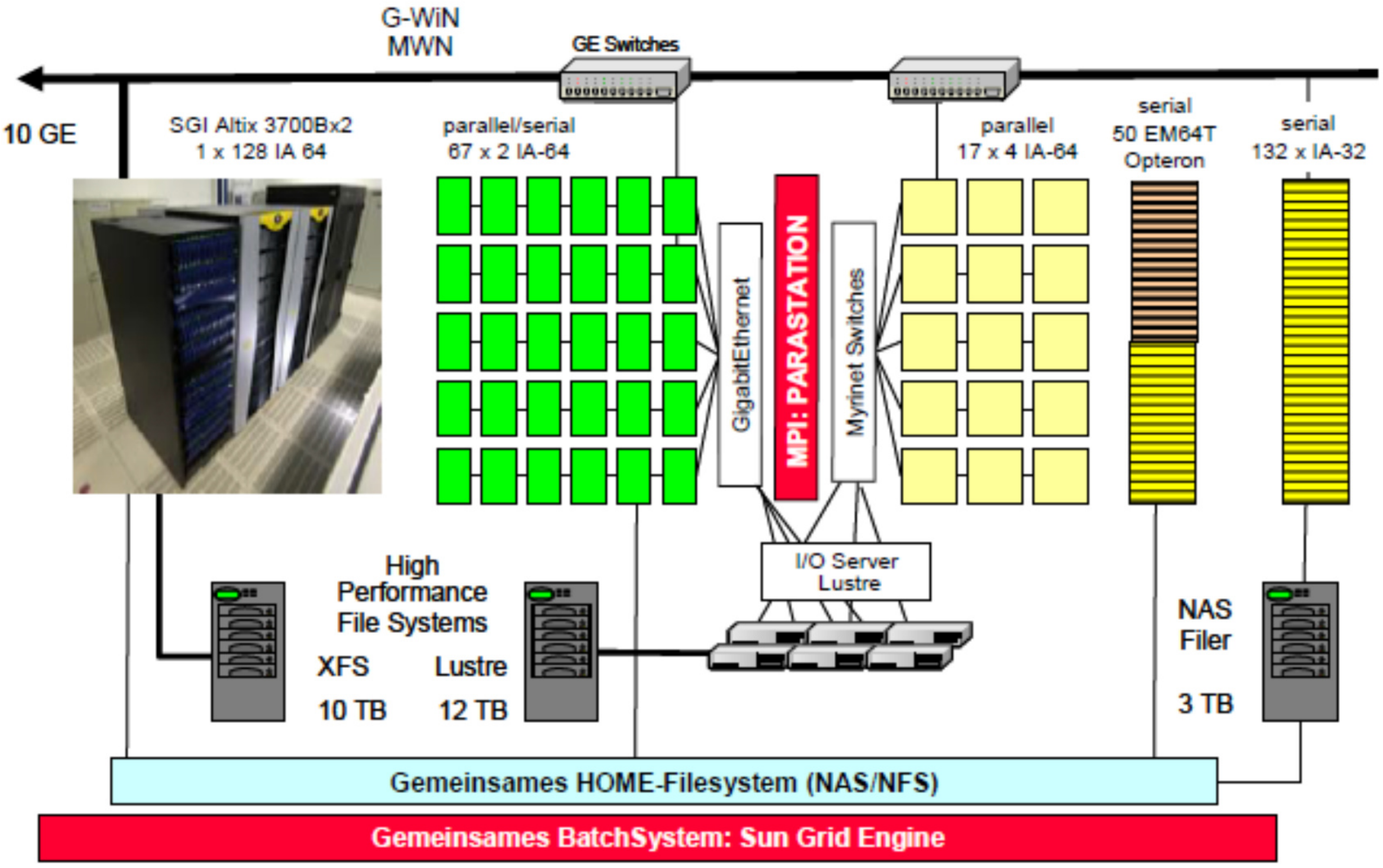
- MEGWARE IA32 cluster
  - 105 nodes with Intel 3,06 GHz Pentium4 processor, 2 GB memory
  - Gb Ethernet network
  - 643 GF peak performance
  - #341 in June 2013 Top500 list

- MEGWARE IA64 cluster
  - 17 quad Itanium2 (Madison) nodes with 8 GB memory
  - Myrinet 2000 communication network
  - 354 GF peak performance
  - 1,5 TB disk space (PVFS)

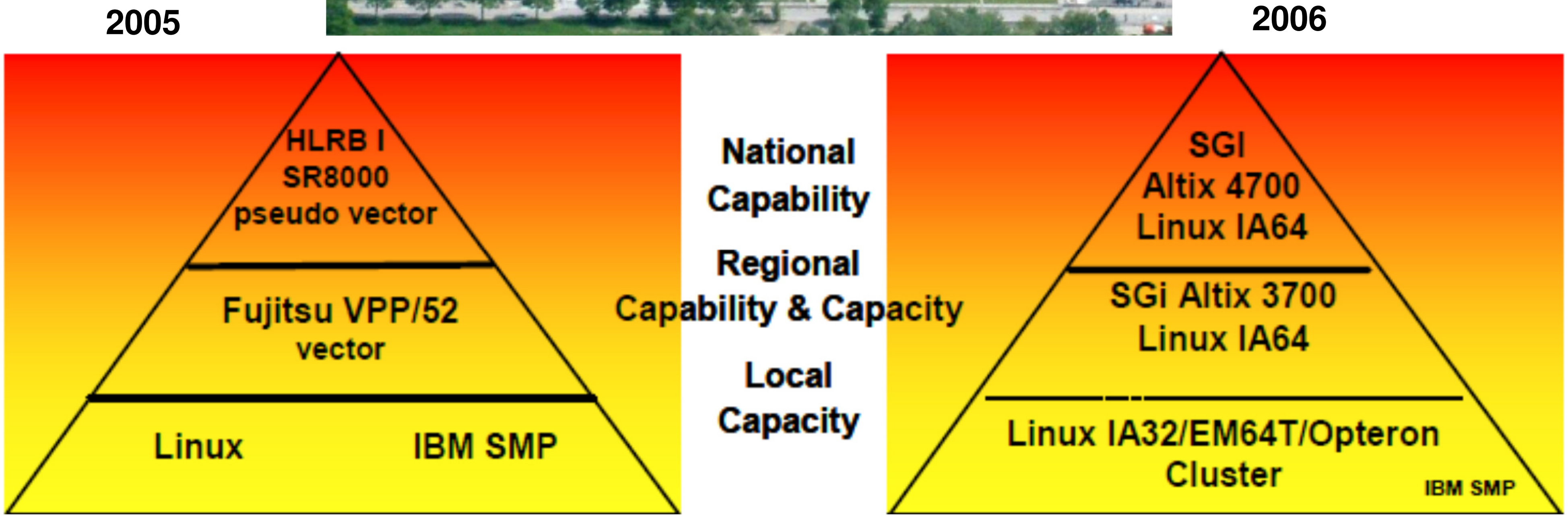
2004: Replacement of Fujitsu VPP by IA64 Linux Cluster and 128-way sgi Altix 3700Bx2

- Sgi Altix 3700Bx2
  - 128 Itanium2 (Madison) processors
  - 512 GB memory
  - NUMALink3 network
  - 819 GF peak performance
  - 10 TB disk space
- MEGWARE IA64 cluster
  - 17 quad Itanium2 (Madison) nodes with 8 GB memory and Myrinet 2000 communication network
  - 67 dual Itanium2 (Madison) nodes with 8 GB memory and Gb Ethernet communication network
  - 1677 GF peak performance
  - 12 TB disk space (Lustre)

The LRZ Linux Cluster in the Year 2005



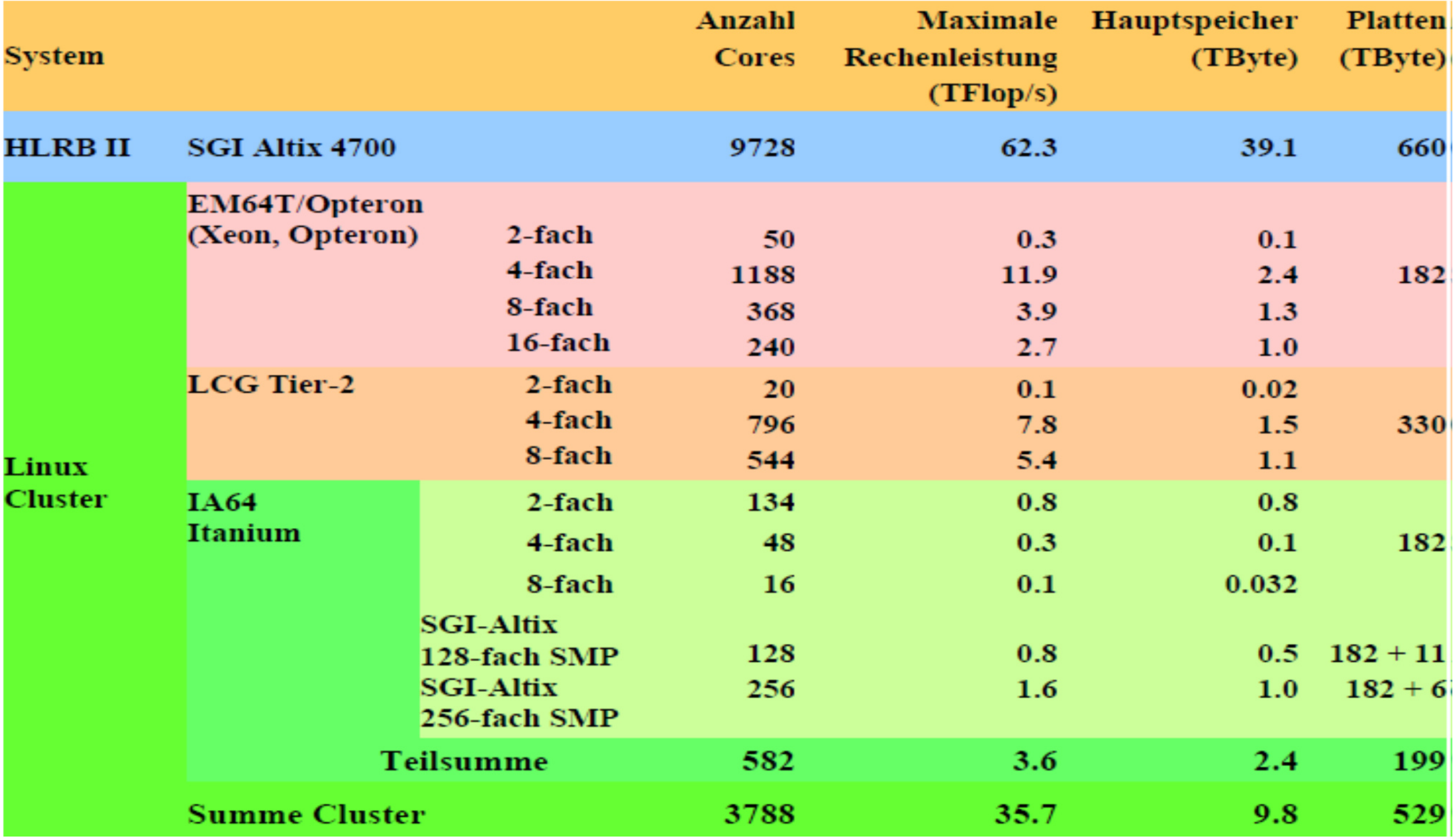
2006: Move from Munich to Garching and Consolidation of HPC Operating Systems and Platforms



2007: Further Extension of the Linux Cluster

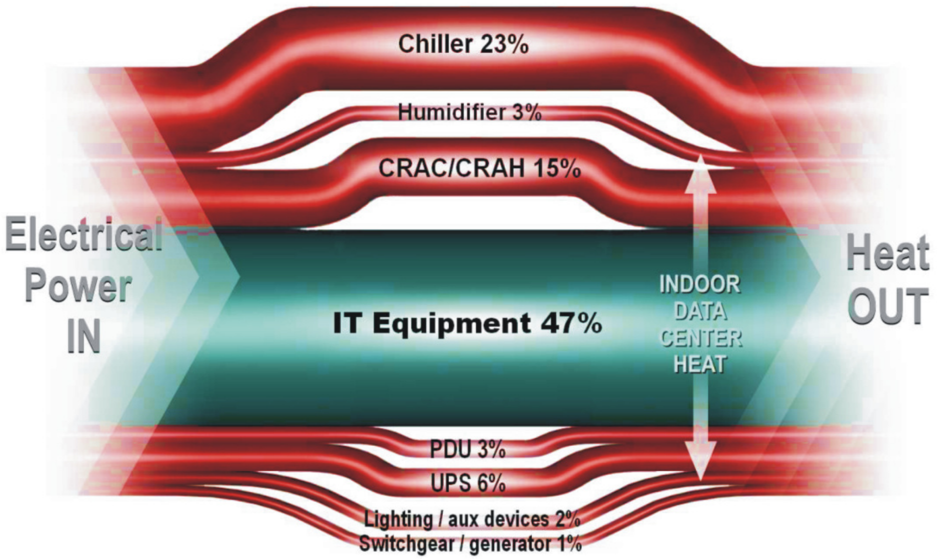
- Sgi Altix 4700
  - 256 Itanium Montecito processors
  - 1024 GB memory
  - NUMALink4 network
  - 1638 GF peak performance
  - 6 TB disk space
- x86-64 cluster
  - 232 MEGWARE AMD x86-64 dual core nodes
  - 99 MEGWARE Intel x86-64 quad core nodes
  - 38 MEGWARE AMD dual core quad socket nodes
  - 15 MEGWARE dCache server with 150 TB total disk space
  - 15 Sun X4600 dual core eight socket systems
- 182 TB of sgi disk storage for Lustre

LRZ HPC systems in the Year 2008



System			Anzahl Cores	Maximale Rechenleistung (TFlop/s)	Hauptspeicher (TByte)	Platten (TByte)
HLRB II	SGI Altix 4700		9728	62.3	39.1	660
Linux Cluster	EM64T/Opteron (Xeon, Opteron)	2-fach	50	0.3	0.1	
		4-fach	1188	11.9	2.4	182
		8-fach	368	3.9	1.3	
		16-fach	240	2.7	1.0	
	LCG Tier-2	2-fach	20	0.1	0.02	
		4-fach	796	7.8	1.5	330
		8-fach	544	5.4	1.1	
	IA64 Itanium	2-fach	134	0.8	0.8	
		4-fach	48	0.3	0.1	182
		8-fach	16	0.1	0.032	
		SGI-Altix 128-fach SMP	128	0.8	0.5	182 + 11
		SGI-Altix 256-fach SMP	256	1.6	1.0	182 + 6
	Teilsumme		582	3.6	2.4	199
	Summe Cluster		3788	35.7	9.8	529

Power Usage Effectiveness (PUE)



- Most **air-cooled datacenters** are **inefficient**. Cooling needs as much energy as IT equipment and both are thrown-away.
- Provocative: datacenter is a huge **“heater with integrated logic.”**
- **PUE of new LRZ data center ~ 1.5**

LRZ Activities to enhance the Power and Cooling Effectiveness of its Data Centre #1

- Use Total Cost of Ownership (TCO) as an important evaluation criteria in procurements
  - Invest and maintenance
  - Power bill (incl. cooling)
  - Total power cooling of components for the calculation of total IT operation costs

- Use of virtualization techniques (VMware)

- Improve PUE

LRZ Activities to further enhance the Cooling Effectiveness of its Data Centre #2

Implementation of a cold and hot aisle containment which is compatible with the argon fire extinguishing concept

Use of additional cold air ducts at power intensive racks (10 kW)

Installation of a room neutral and direct liquid cooled rack solution for very high power densities > 15 kW per rack

Indirect Liquid Cooled Rack Solutions

- Room neutral
- Better cooling efficiencies due to reduced air throw distances
- Optimal cold/hot aisle confinement

Rear Door Heat Exchanger

Closed Racks with Integrated Heat Exchangers

Air versus Water Cooling

Air cooling is the de-facto standard

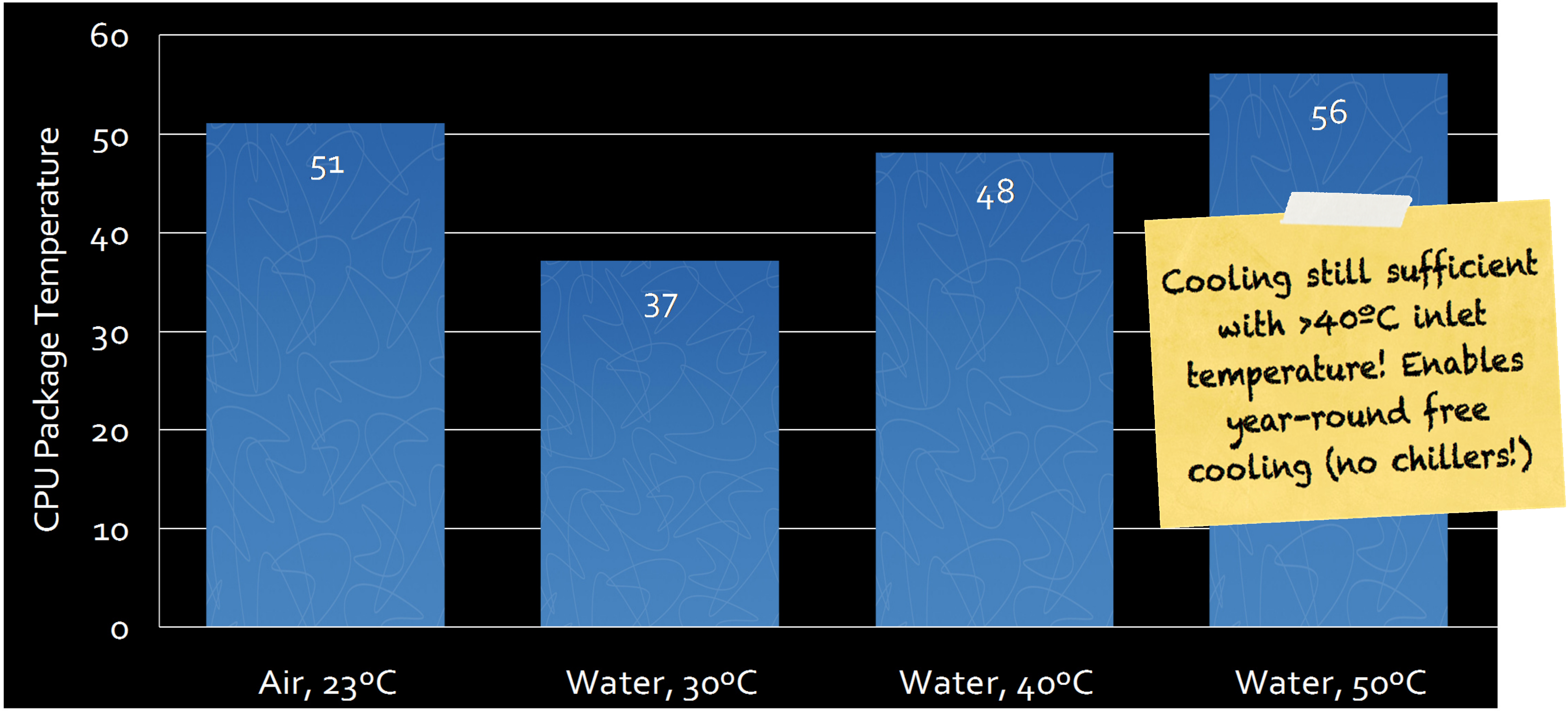
But:

	Air	Water	Factor
Thermal Conductivity	**0.026** W/(m*K)	**0.56** W(m*K)	21.5 x
Thermal Capacity	**1.00** J/(g*K)	**4.18** J/(g*K)	4.18 x

⇨ Water as coolant allows higher inlet temperatures (free cooling!)

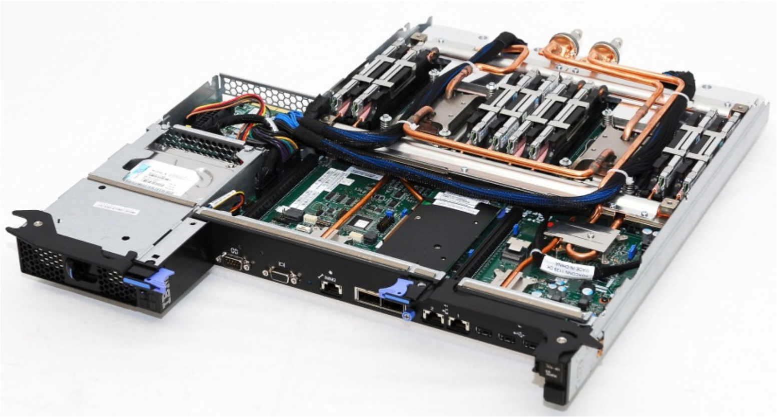
Water enables better heat reuse

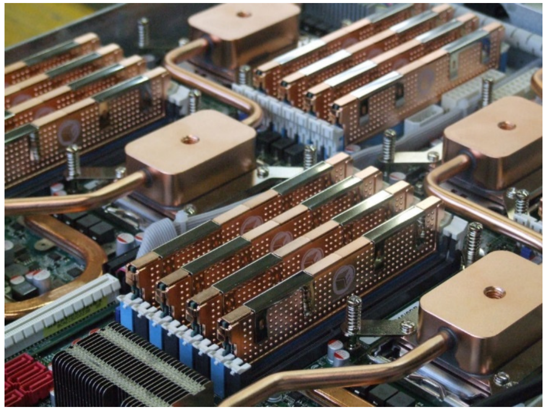
Air Cooling versus Direct Liquid Cooling



Cooling still sufficient with >40°C inlet temperature! Enables year-round free cooling (no chillers!)

2009-2011: Construction of New Building with Warm Water Cooling Loops & Procurement of Direct Warm Water Cooled HPC Systems





- 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 – 30° C, T_{water} = 18 – 45° C

2011: Delivery and Installation of CooLMUC

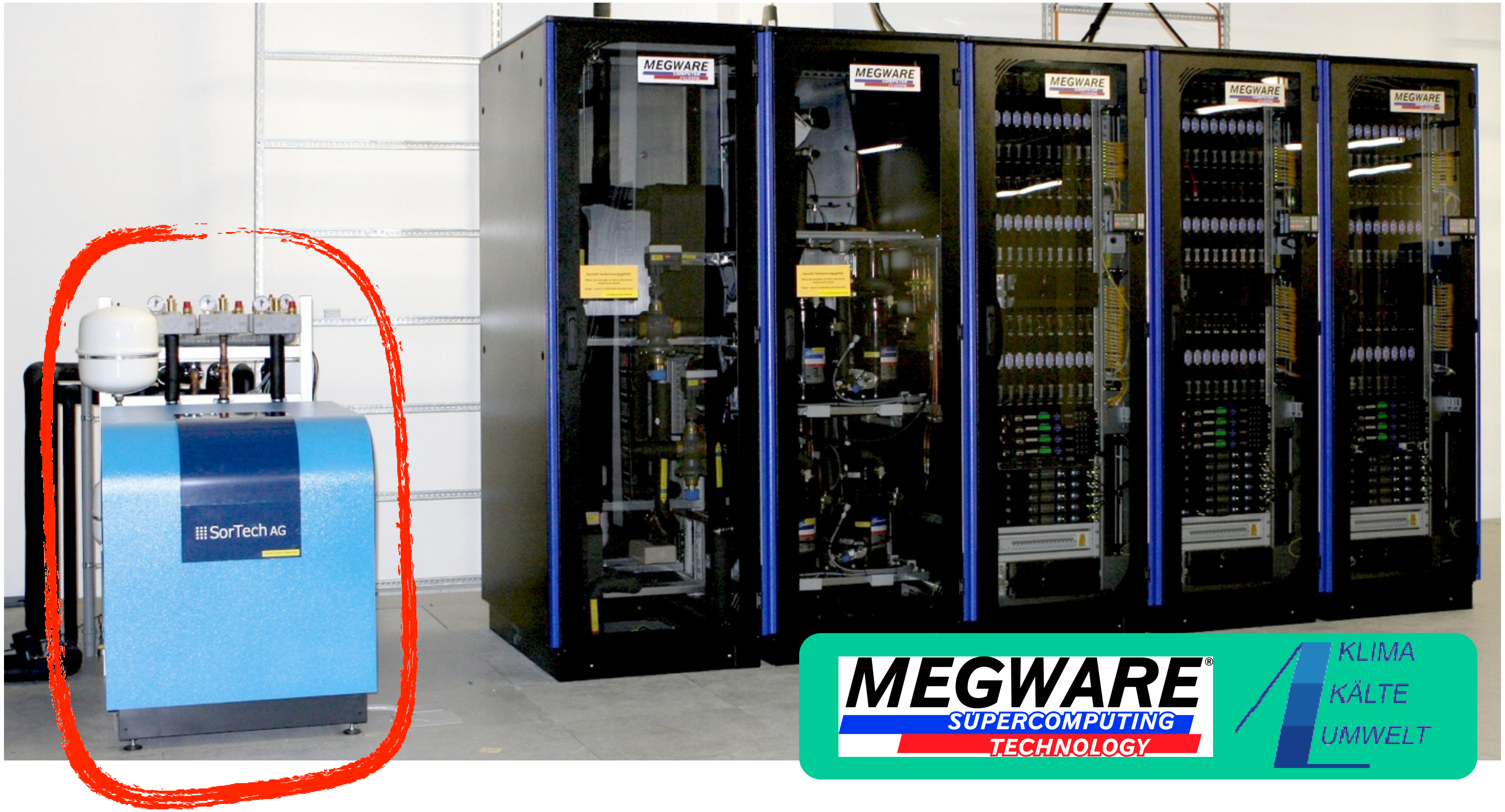
- The worlds first AMD-based direct water-cooled cluster with
  - 178 nodes (2x8 core AMD Magny Cour 2.0 GHz CPUs and 16 GByte RAM per node)
  - IB QDR network
  - Thorough power monitoring for compute & cooling hardware
  - Completely closed racks (no dependence on room air conditioning)
  - Reuse of waste-heat for cooling through a SorTech adsorption chiller

lrz
MEGWARE COMPUTER CLUSTER
direct server-cooling system
"CooLMUC"
World's First AMD Cluster System with Direct Hot Water Cooling
• 178 MEGWARE ClustCool® Dualnode AMD Opteron 6128HE
• ColdCon® - direct water cooling of nodes by means of highly effective heat exchangers
• Re-use of waste heat for air conditioning

MEGWARE COMPUTER CLUSTER

KLIMA KÄLTE UMWELT

Weitere Details zu CooLMUC → Vortrag von Herrn Wilde



MEGWARE HPC Cluster am LRZ – eine mehr als 12-jährige Zusammenarbeit

- Fazit
  - MEGWARE geht auf Kundenwünsche ein und ist in der Lage auch sehr innovative HPC-Lösungen anzubieten
  - LRZ ist mit den HPC-Lösungen von MEGWARE und dem MEGWARE-Support sehr zufrieden
    - Gute partnerschaftliche Arbeitsatmosphäre
    - Schnelle Reaktionszeiten
    - Hohe HPC-Expertise