Prototyping in PRACE

PRACE Energy to Solution prototype at LRZ

Torsten Wilde 1IP-WP9 co-lead and 2IP-WP11 lead (GSC-LRZ)

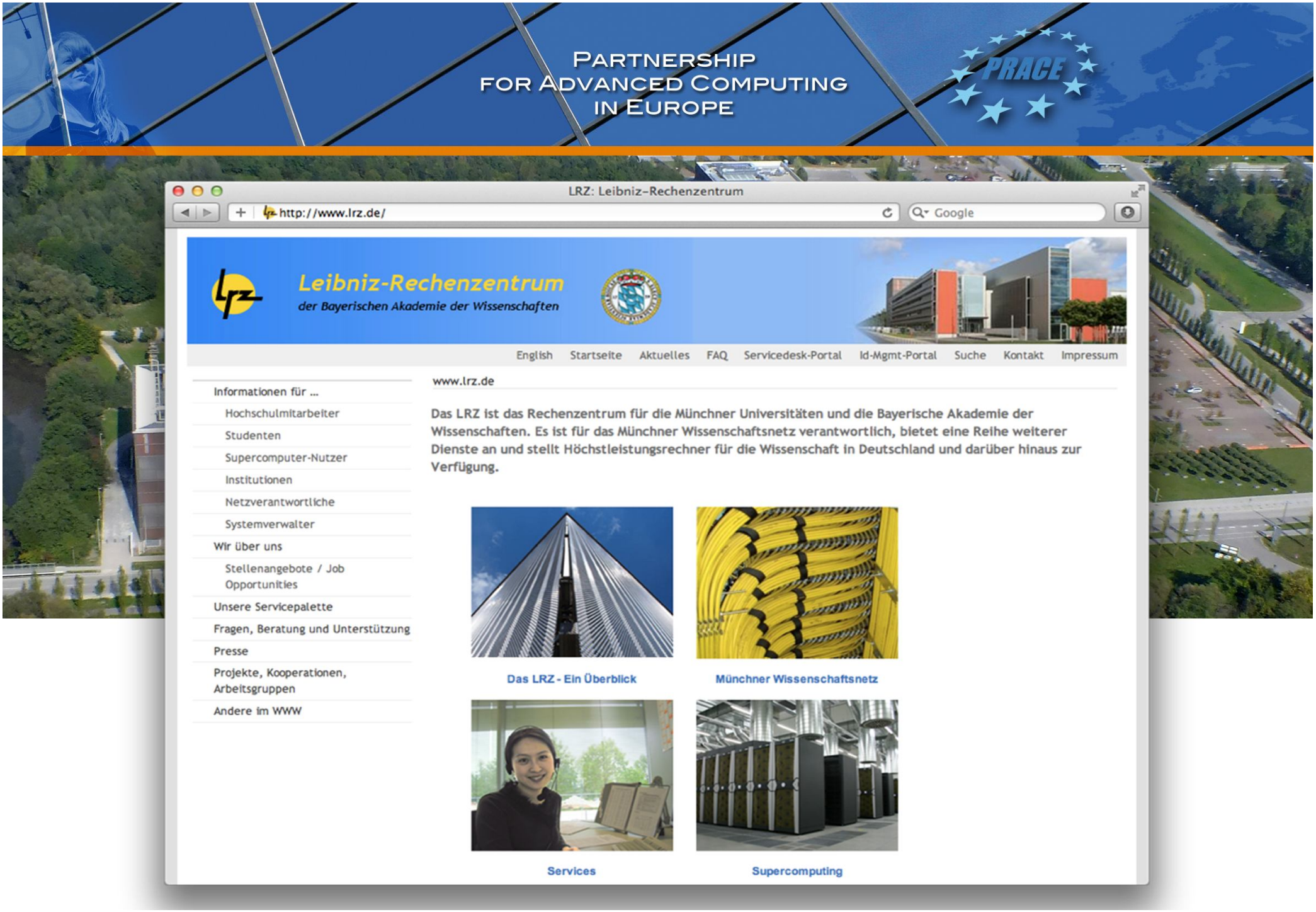
PRACE Industy Seminar, Bologna, April 16, 2012

CAPACITIES

e-infrastructure

Leibniz Supercomputing Center

Partnership for Advanced Computing in Europe



Outline

- PRACE
  - PRACE and Prototyping
  - Challenges, Energy efficiency
- Energy to Solution (EtS) Prototype at LRZ
  - Motivation
  - Technology
  - First results

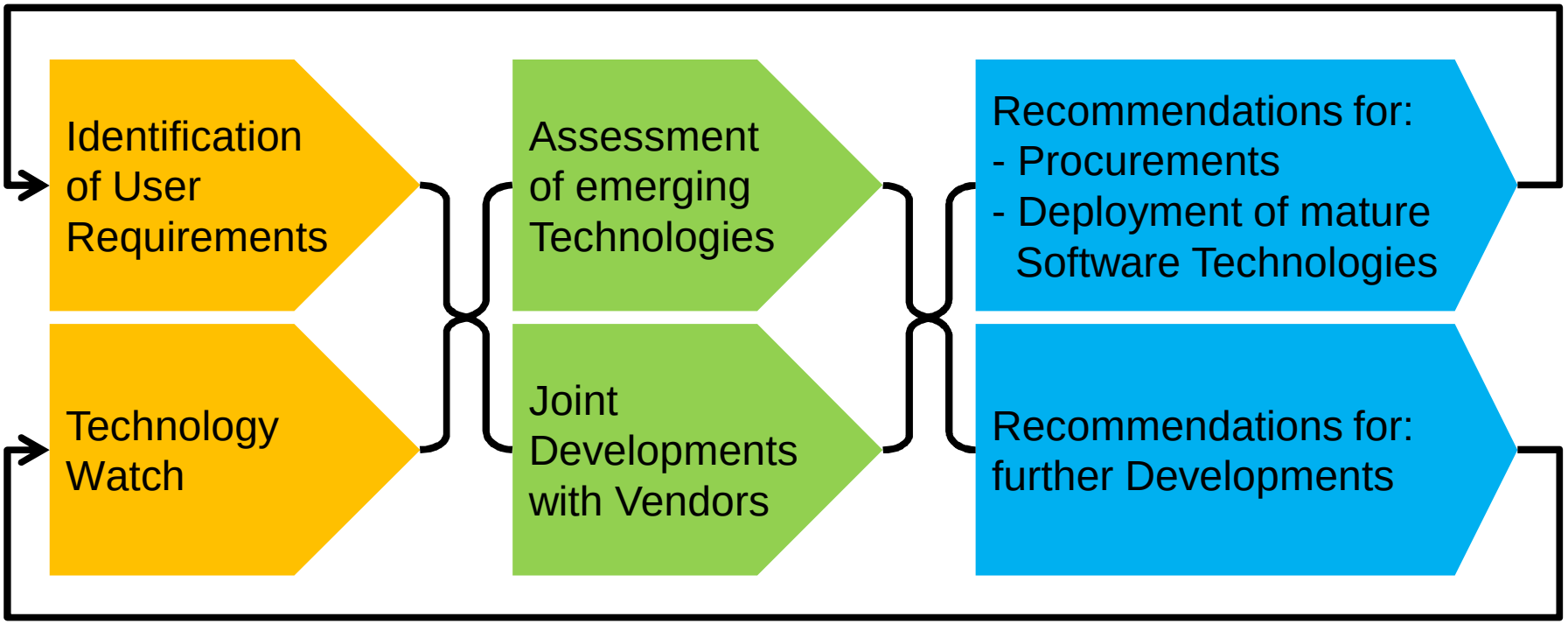
PRACE – A Partnership with a Vision

- Provide world-class HPC systems for world-class science
- Support Europe in attaining global leadership in public and private research and development

… and a Mission

- Create a world-leading persistent high-end HPC infrastructure managed as a single legal entity
  - **Deploy systems of the highest performance level (Tier-0)**
  - **Ensure a diversity of architectures to meet the needs of European user communities**
  - **Collaborate with vendors and ISVs on strategic HPC technologies**
  - Provide support and training

Prototyping is Mandatory for PRACE



- Prototyping is a mandatory step in the selection and deployment of new technologies
- Prototyping is a vehicle for cooperation with technology providers

Future System Architecture Projections/Targets

Energy Efficiency

Systems	2010	2012	2015	2018	Difference Today & 2018
System Peak [PF/s]	2	25	200	1000	O(1000)
Power [MW]	6	6-20	15-50	20-80	O(10)
System Memory [PB]	0.3	0.3-0.5	5	32-64	O(100)
GB RAM/Core	0.5-4	0.5-2	0.2-1	0.1-0.5	-O(10)
Node Performance [GF/s]	125	160-1000	500-7000	1000-10000	O(10)-O(100)
Cores/Node	12	16-32	100-1000	1000-10000	O(100)-O(1000)
Node memory BW [GB/s]	40	70	100-1000	400-4000	O(100)
Number of nodes	~20.000	10.000-100.000	5000-50.000	100.000-1.000.000	O(10)-O(50)
Total concurrency	~240.000	$O(10^6)$	$O(10^7)$	$O(10^9)$	O(50.000)
MTTI	days	days	O(1 day)	O(1 day)	-O(10)

Source: Rick Stevens and Andy White, IESP Meeting, Oxford 2010

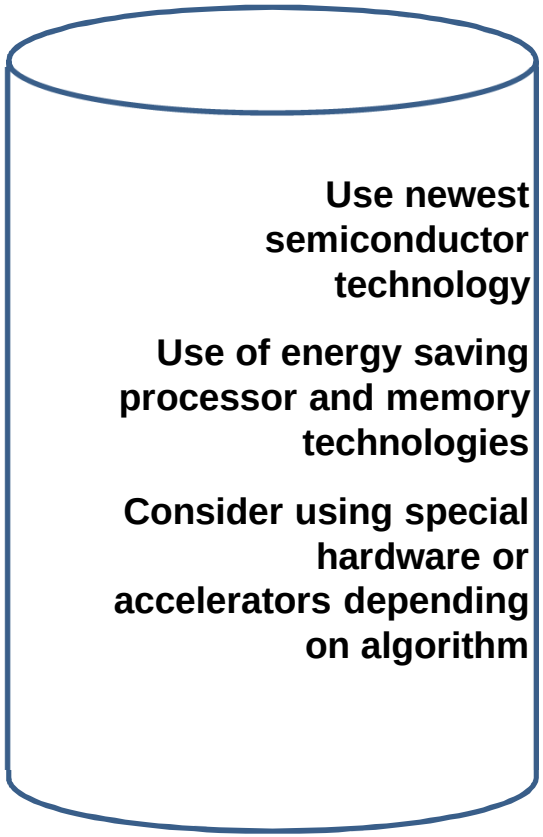
Todays “Energy Efficiency”

Nr.1 Top500
K computer, RIKEN, Japan
10.51 PetaFlop/s @
12.659 MW
(830.18 MFlop/s per W)

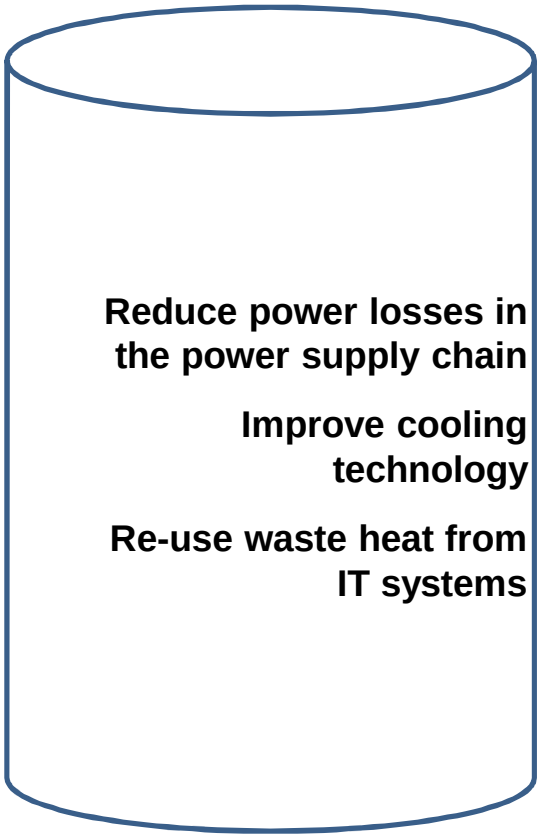
~ 1 Exaflop/s
*@ **1.2GW***

Nr. 1 Green500
Blue Gene/Q, IBM –
Rochester, MI, US
(2026.48 MFlop/s per W)

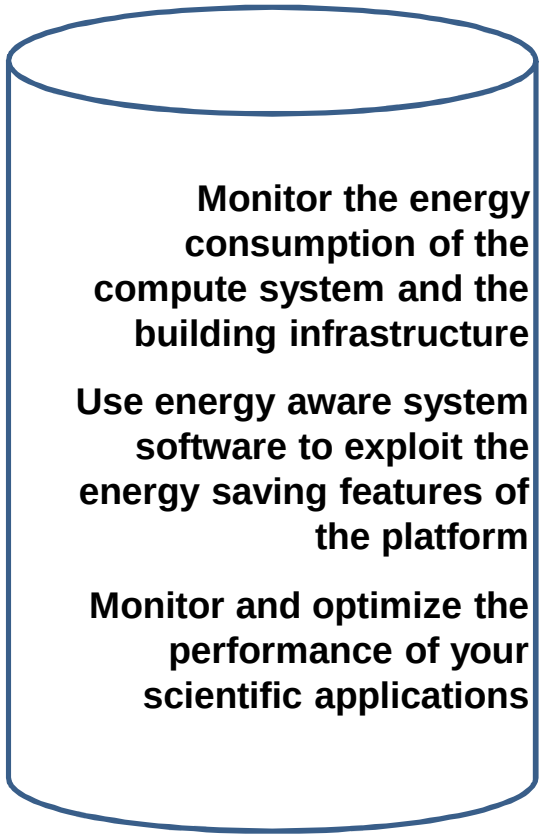
~ 1 Exaflop/s
*@ **500MW***



Energy efficient hardware



Energy efficient infrastructure



Energy aware management software

- **PRACE ARM and ARM+GPU**
- **Mont-Blanc**
- **Deep**

Use newest semiconductor technology

Use of energy saving processor and memory technologies

Consider using special hardware or accelerators depending on algorithm

Energy efficient hardware

Reduce power losses in the power supply chain

Improve cooling technology

Re-use waste heat from IT systems

Energy efficient infrastructure

Monitor the energy consumption of the compute system and the building infrastructure

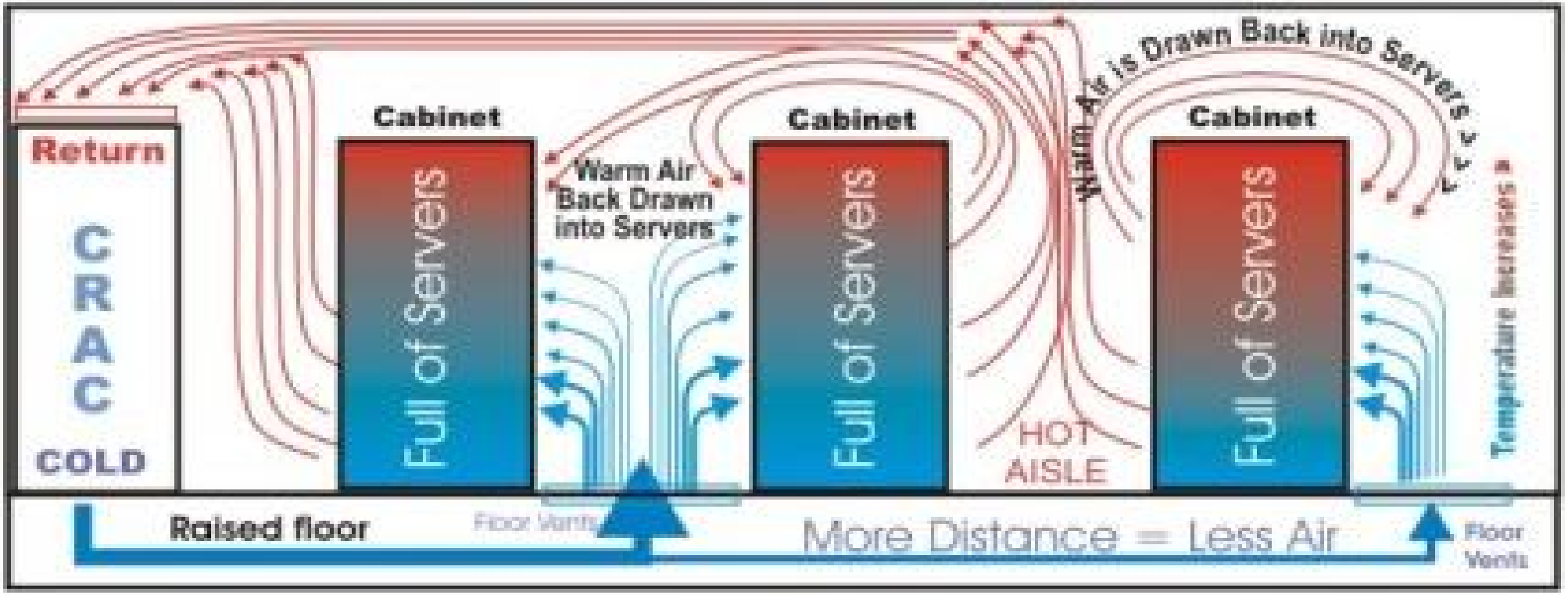
Use energy aware system software to exploit the energy saving features of the platform

Monitor and optimize the performance of your scientific applications

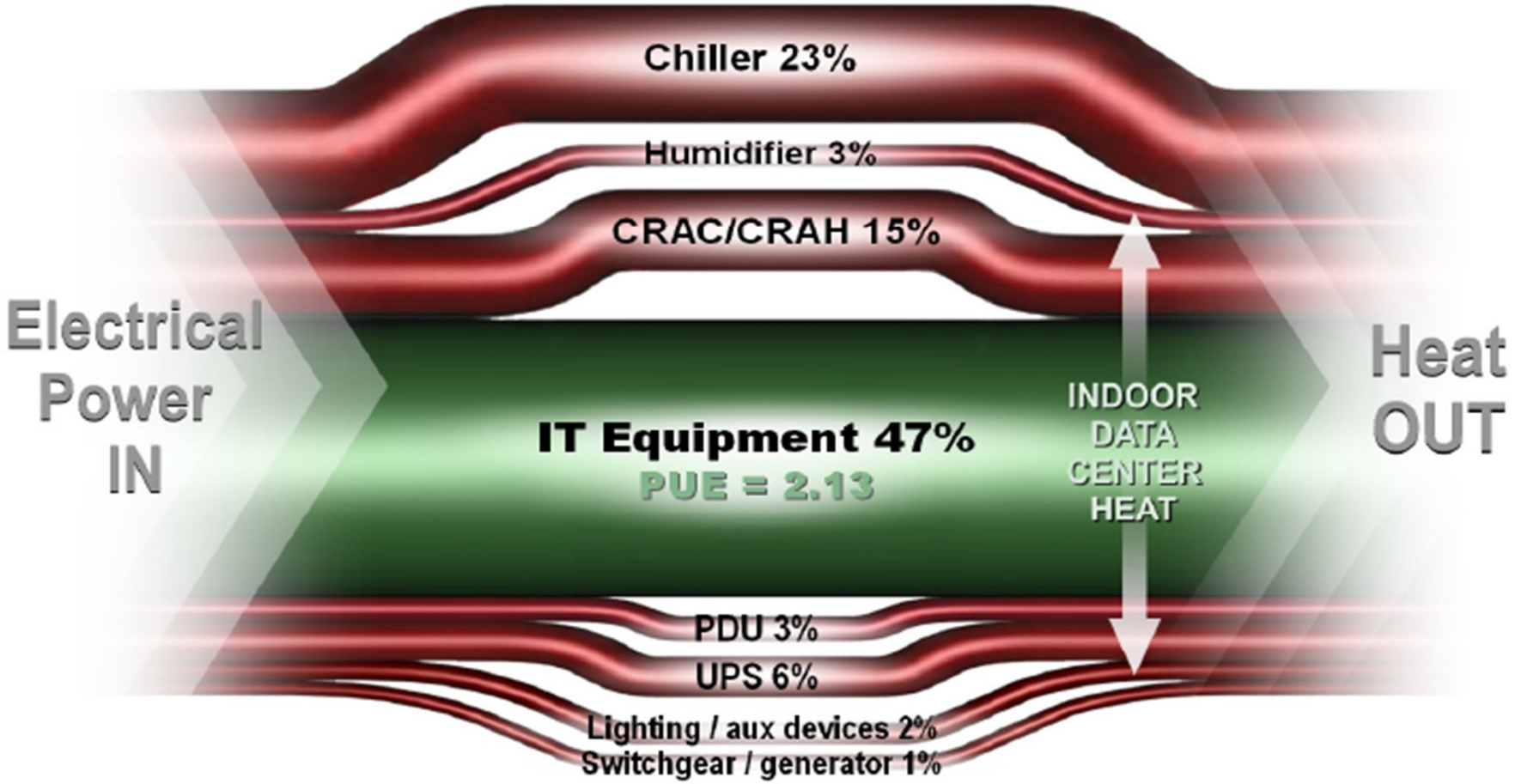
Energy aware management software

Raised Floor Airflow

Rack Temperatures Increase with Distance from CRAC



Cooling matters (courtesy of APC)



Power Challenge at LRZ

System	Era	Peak Performance	Power Consumption	Investment Costs	Total Operating Costs (incl. Power)	Power Bill
HLRB I: Hitachi SR8000	*2000 - 2006*	*1.3 TFLOP/s*	*0.5 MW*	*29 M€*	*13 M€*	*3 M€*
HLRB II: sgi ALTIX 4700	*2006 - 2011*	*62 TFLOP/s*	*1 MW*	*35 M€*	*16 M€*	*7 M€*
SuperMUC: IBM iDataPlex	*2012 - 2016*	*3000 TFLOP/s*	*3 MW*	*48 M€*	*35 M€*	*22 M€*

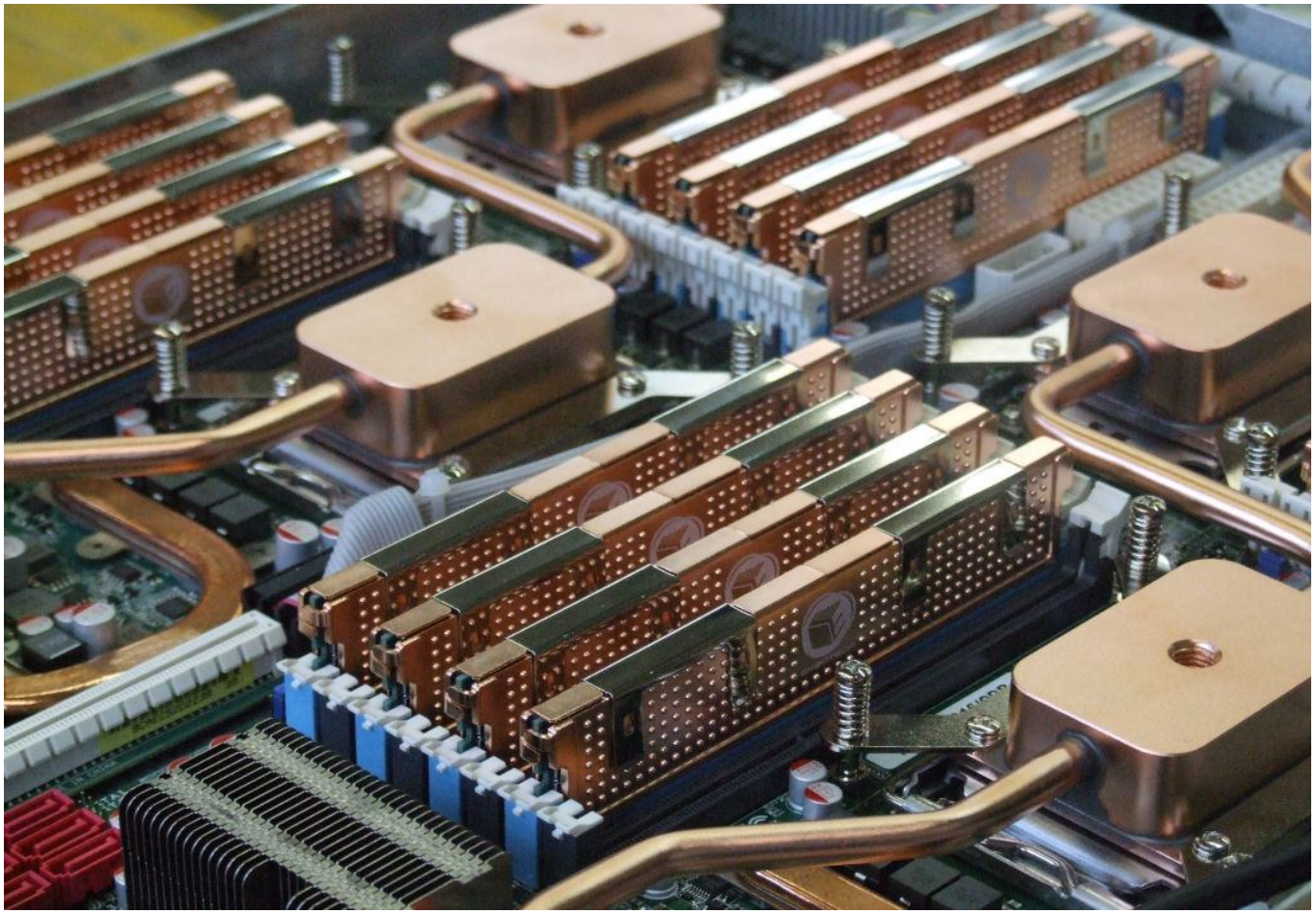
CooLMUC Hardware

- The world's first AMD based direct warm/hot water-cooled cluster with
  - 178 nodes (2x8-core AMD CPU and 16GB RAM per node)
  - Infiniband QDR network
- Power monitoring for nodes, network equipment and cooling hardware
- Closed racks (no dependence on room air conditioning)
- Waste-heat reuse through a SorTech adsorption chiller

MEGWARE COMPUTER CLUSTER

KLIMA
KÄLTE
UMWELT

Partnership
for Advanced Computing
in Europe

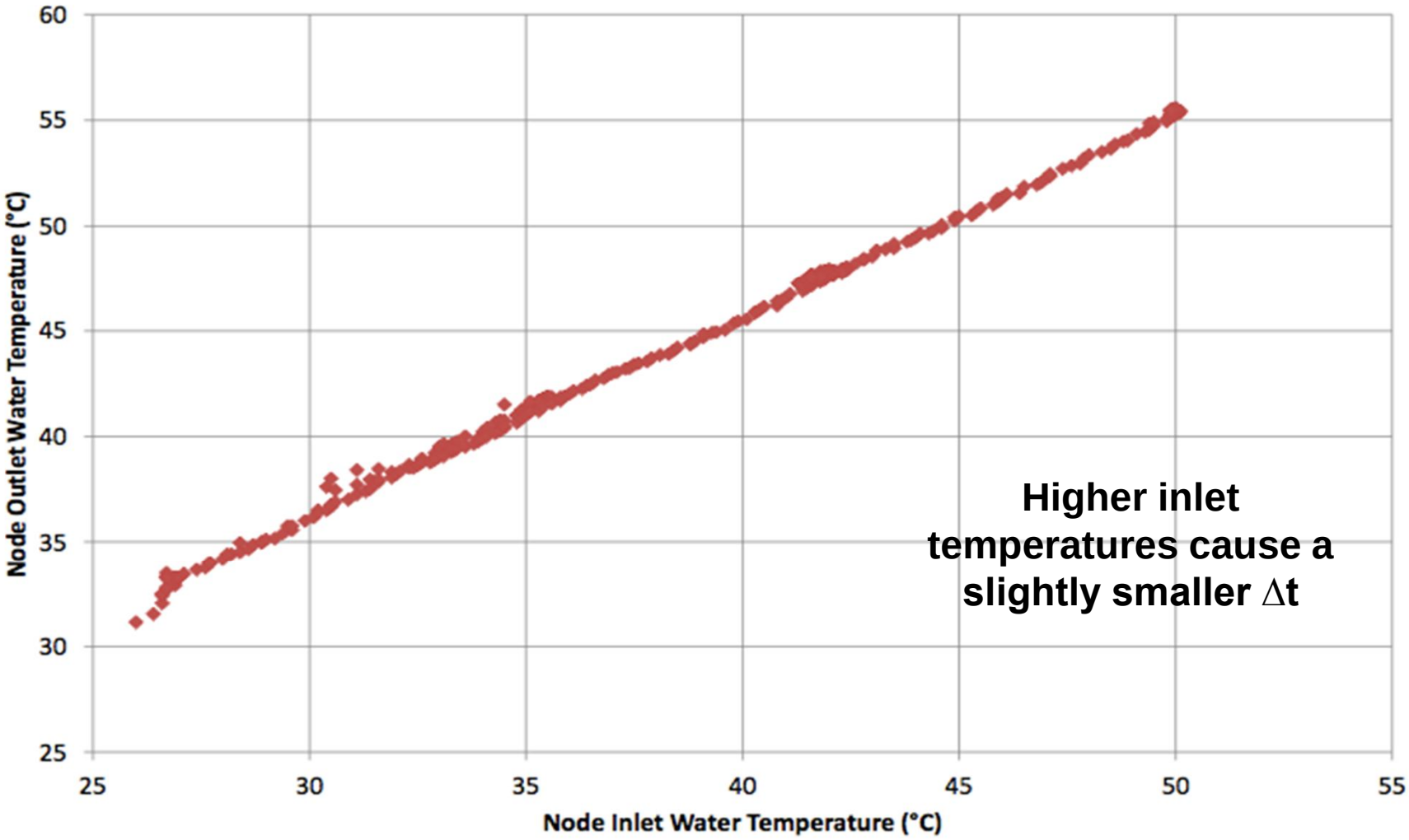


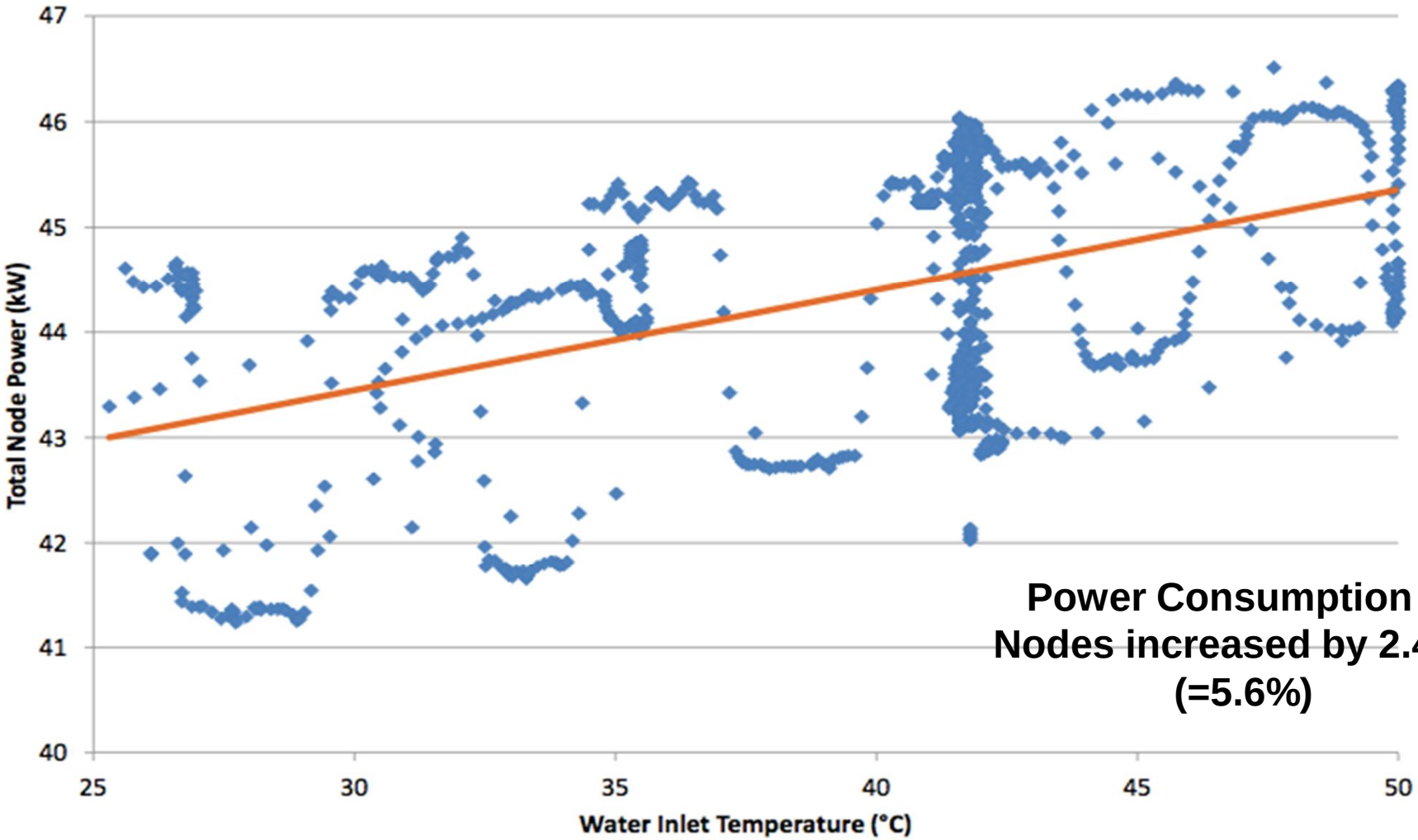
Direct water cooled CPUs, Infiniband HCAs, Chipsets, ...

Warm water cooling first results

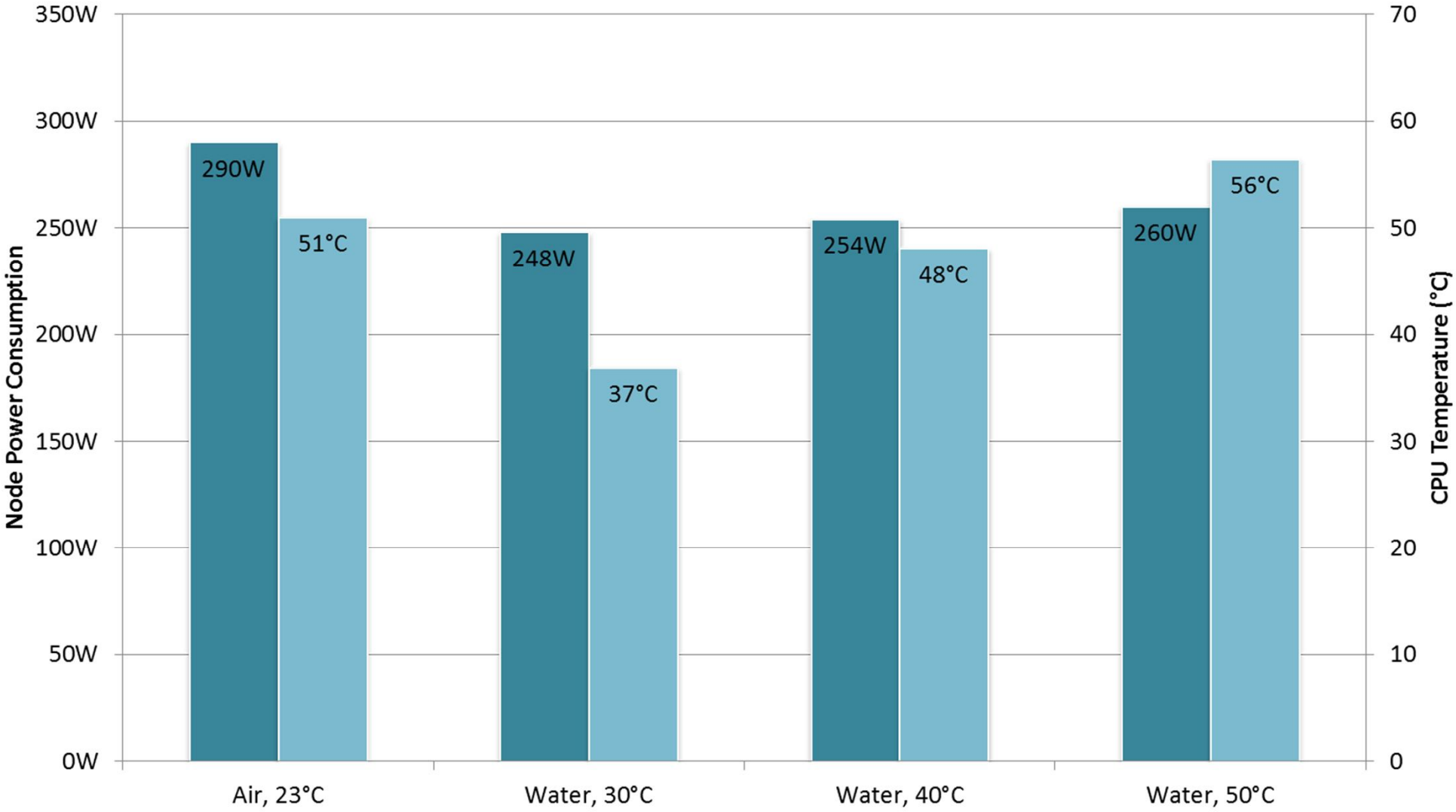
- This is **really early evaluation** data, freshly collected!
- Measurement equipment accuracy is not checked!
- Results are specific to current CooLMUC prototype setup!

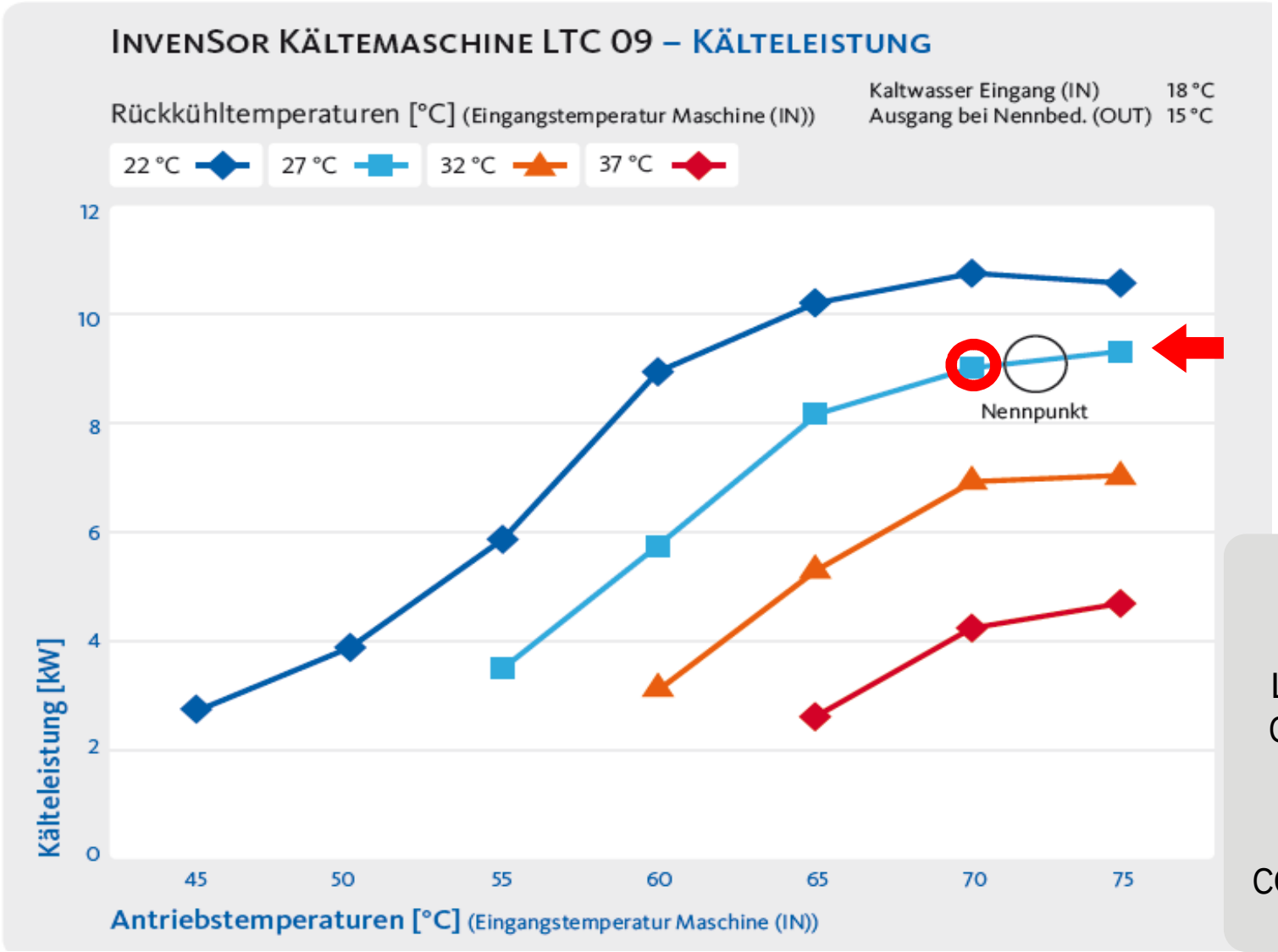
Use all possible caution when drawing conclusions!


Node Outlet Water Temperature (°C)
60
55
50
45
40
35
30
25
25
30
35
40
45
50
55
Node Inlet Water Temperature (°C)
Higher inlet temperatures cause a slightly smaller Δt



Power Consumption of Nodes increased by 2.4kW (=5.6%)


350W
300W
250W
200W
150W
100W
50W
0W
Node Power Consumption
70
60
50
40
30
20
10
0
CPU Temperature (°C)
290W
51°C
248W
37°C
254W
48°C
260W
56°C
Air, 23°C
Water, 30°C
Water, 40°C
Water, 50°C



Example Setting

Outside Temperature:
26,1 ºC

Cond. Outlet Setpoint:
70,0 ºC

Water Inlet:
59,8 ºC

Power Consumption

(Baseline: 27 ºC Water Inlet Temperature)

Leakage Overhead: 3,13 kW
Cooling Power: 9,00 kW
COP: 3
(coefficient of performance)

COP (Compression based): 4

Adsorption Summary

- Efficiency and usefulness depends on:
  - Outside temperature
  - Water temperature
  - Components
  - Adsorption machine COP
- Will work well if:
  - Leakage overhead from components is small
  - COP can compete with compressor based cooling
  - Hot water will be created anyways for other reuse purposes

Other reuse of waste heat

Application	Temperature	Loss due to Leakage*	Comment
Hot Water Heating	50 - 90 ºC	5,9 - 15,4 %	If nothing else works …
Swimming Pool	40 - 60 ºC	3,6 - 8,3 %	Ok, if you would need the energy anyways
Underfloor Heating	30 - 35 ºC	1,2 - 2,4 %	* Good
Concrete Core Activation	25 ºC	0,04 %	* Very Good

*Reference: 18 ºC Inlet Temperature

Outlook - What REALY happens to the waste heat?

Bonus Slides

PRACE Funded Prototypes – 1IP Phase 1

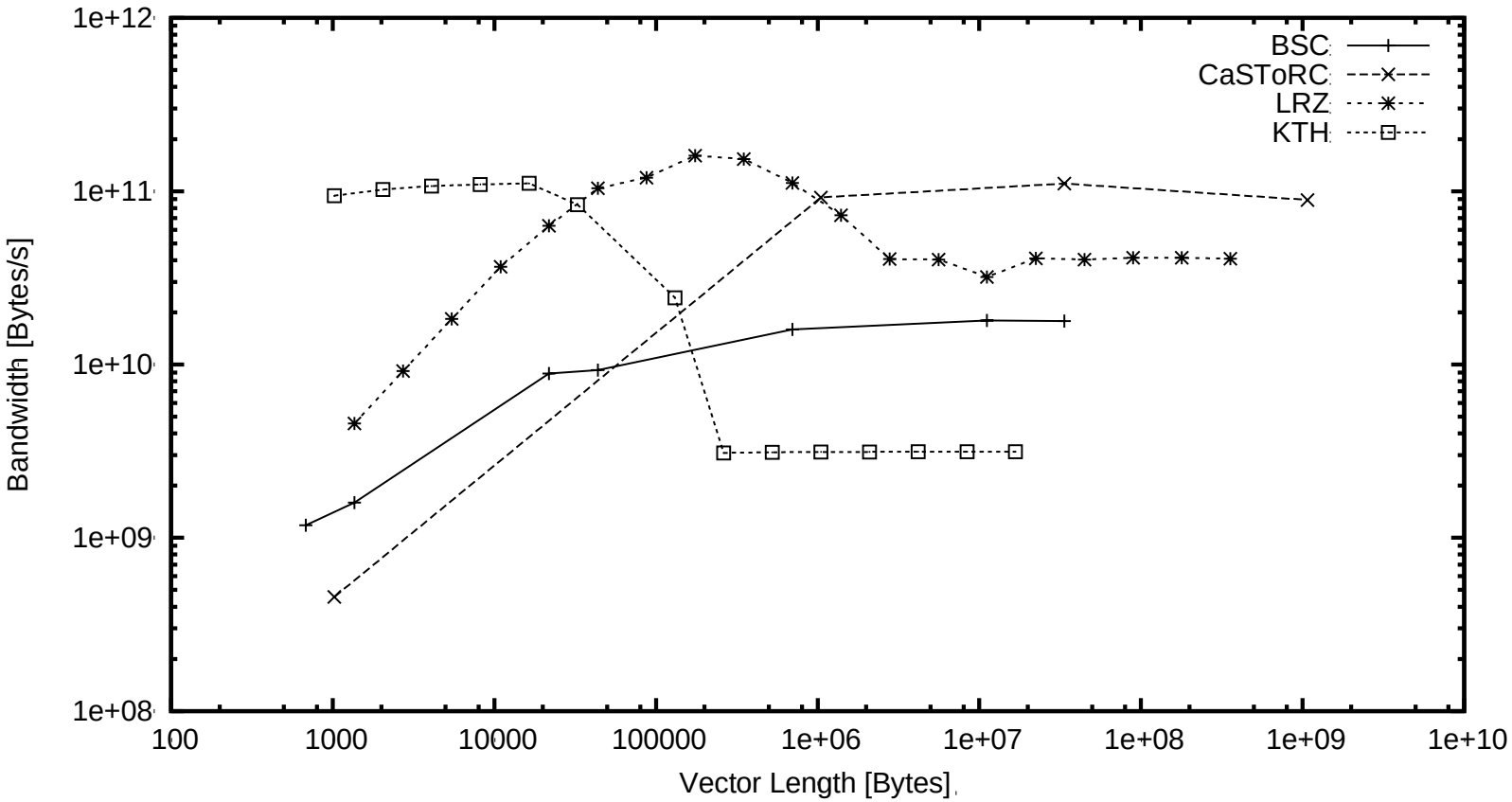
- I/O
  - Exascale IO (CEA & CINES)
  - Novel Exascale I/O Concepts (JSC)
- Accelerators
  - Interconnect Virtualization @ CaSToRC (CaSToRC)
  - DSP HPC Node (KTH)
- Memory
  - NUMA-CIC (UiO)
- Energy-to-Solution
  - Energy-to-Solution @ LRZ (LRZ)
  - Energy-to-Solution @ JKU (JKU)
  - Energy-to-Solution @ BSC (BSC)
  - Energy-to-Solution @ PSNC (PSNC)

Warning

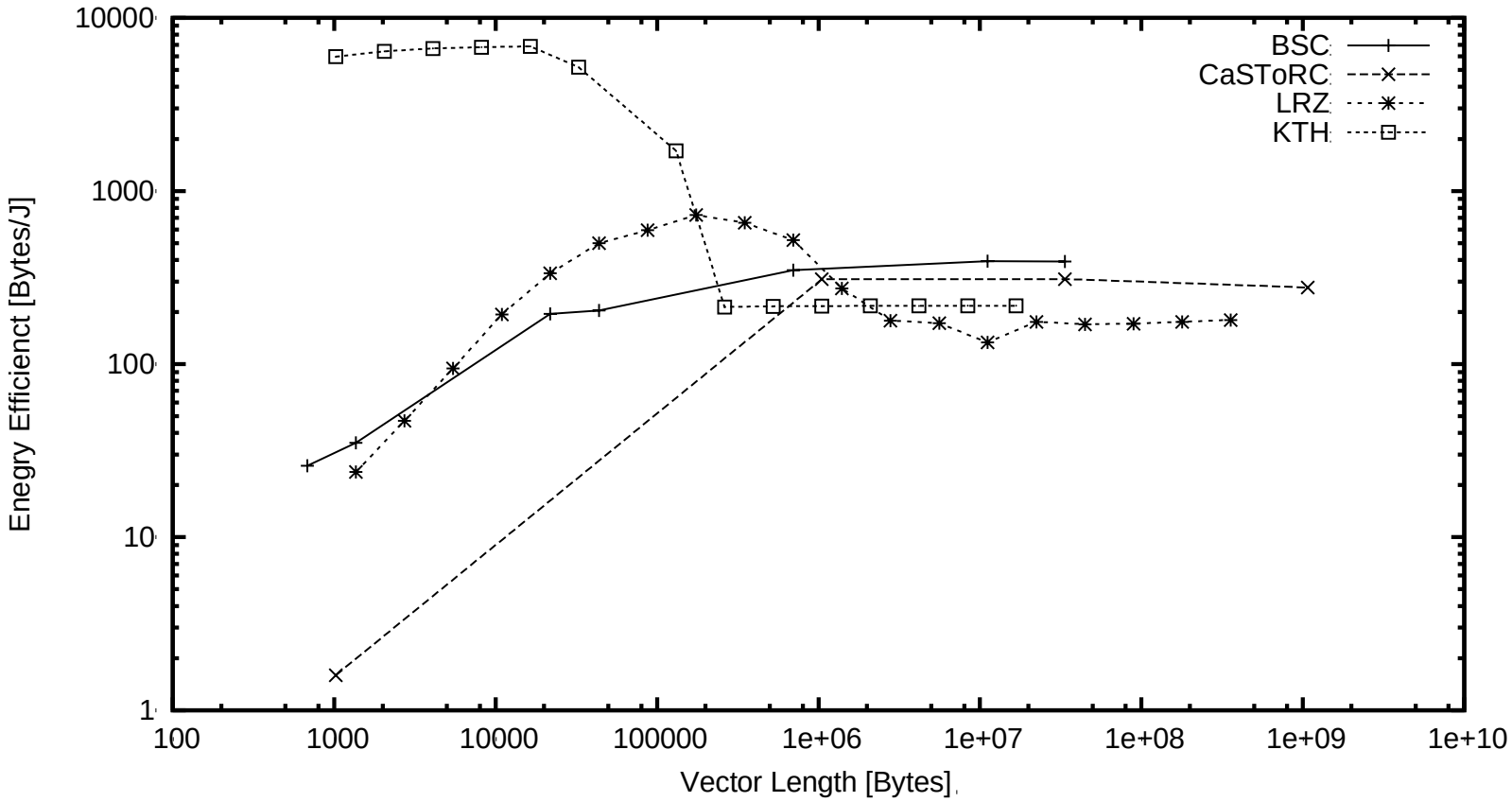
- This is **really early evaluation** data, freshly collated!
- The prototypes span from single chip to clusters!
- Measurement equipment accuracy is not checked!
- Energy measurements vary in scope!
- Most of the codes may be optimized in the future!

Use all the caution possible when drawing conclusions!

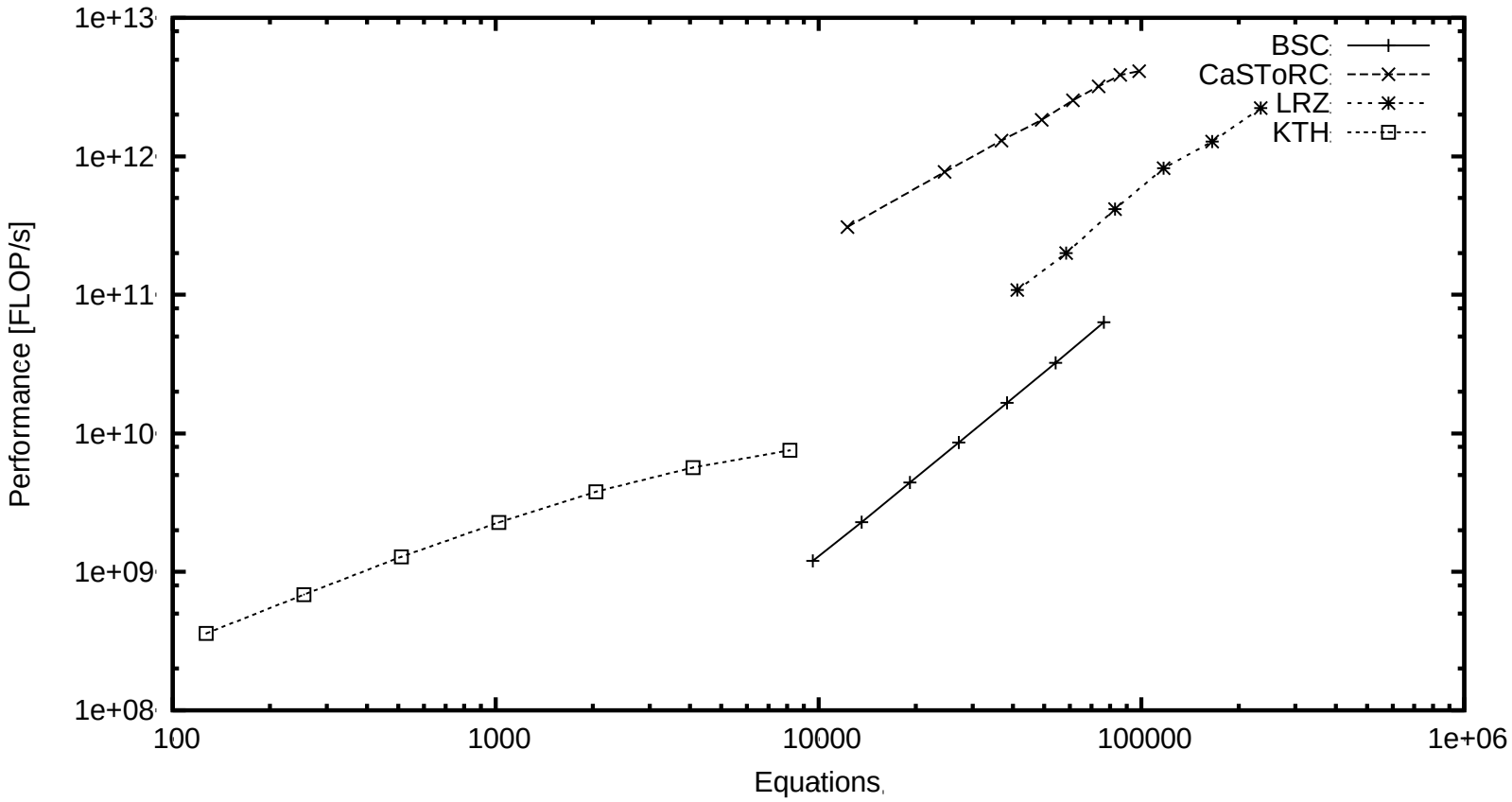
STREAM Performance Space



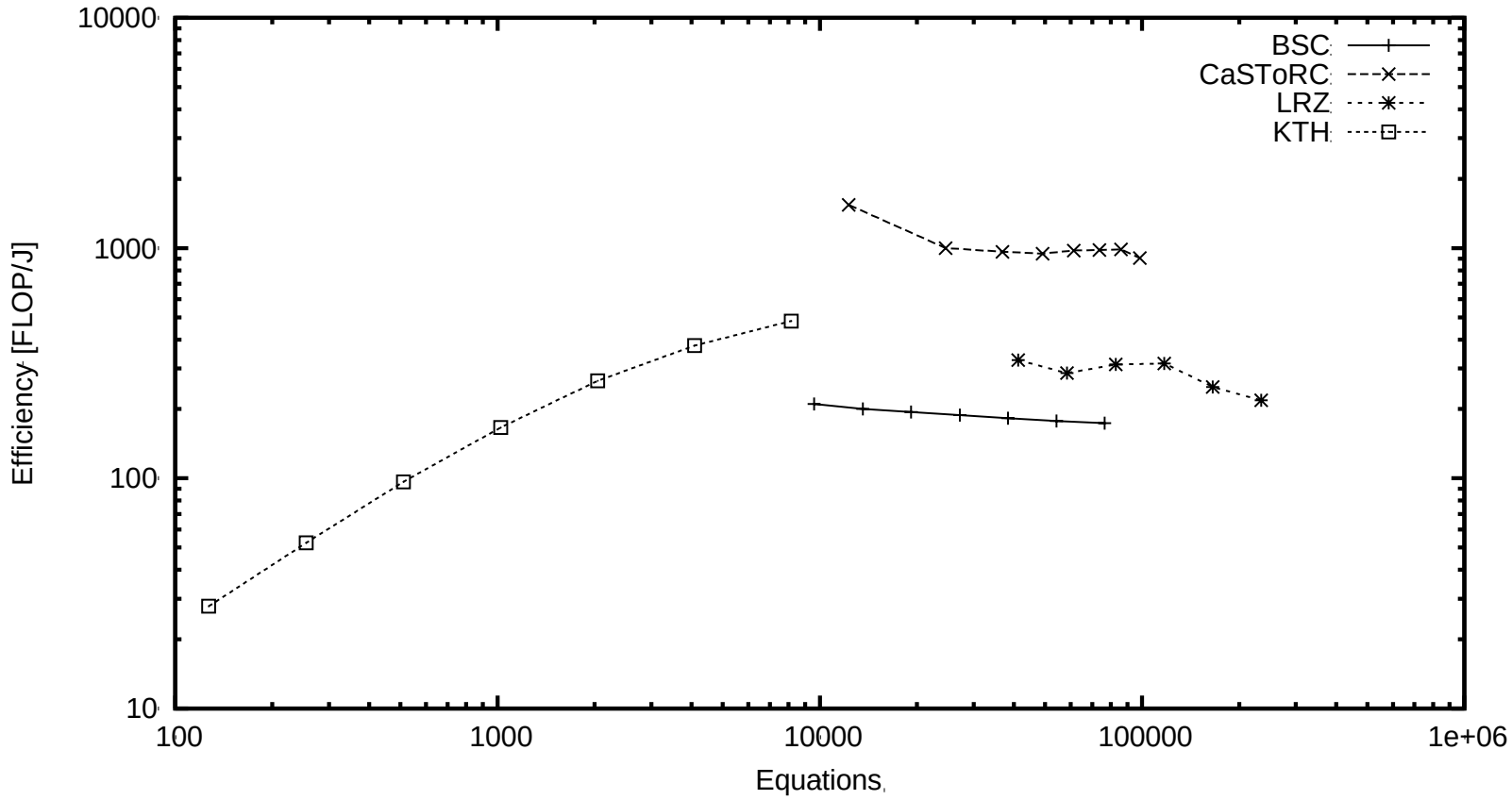
STREAM Energy Efficiency

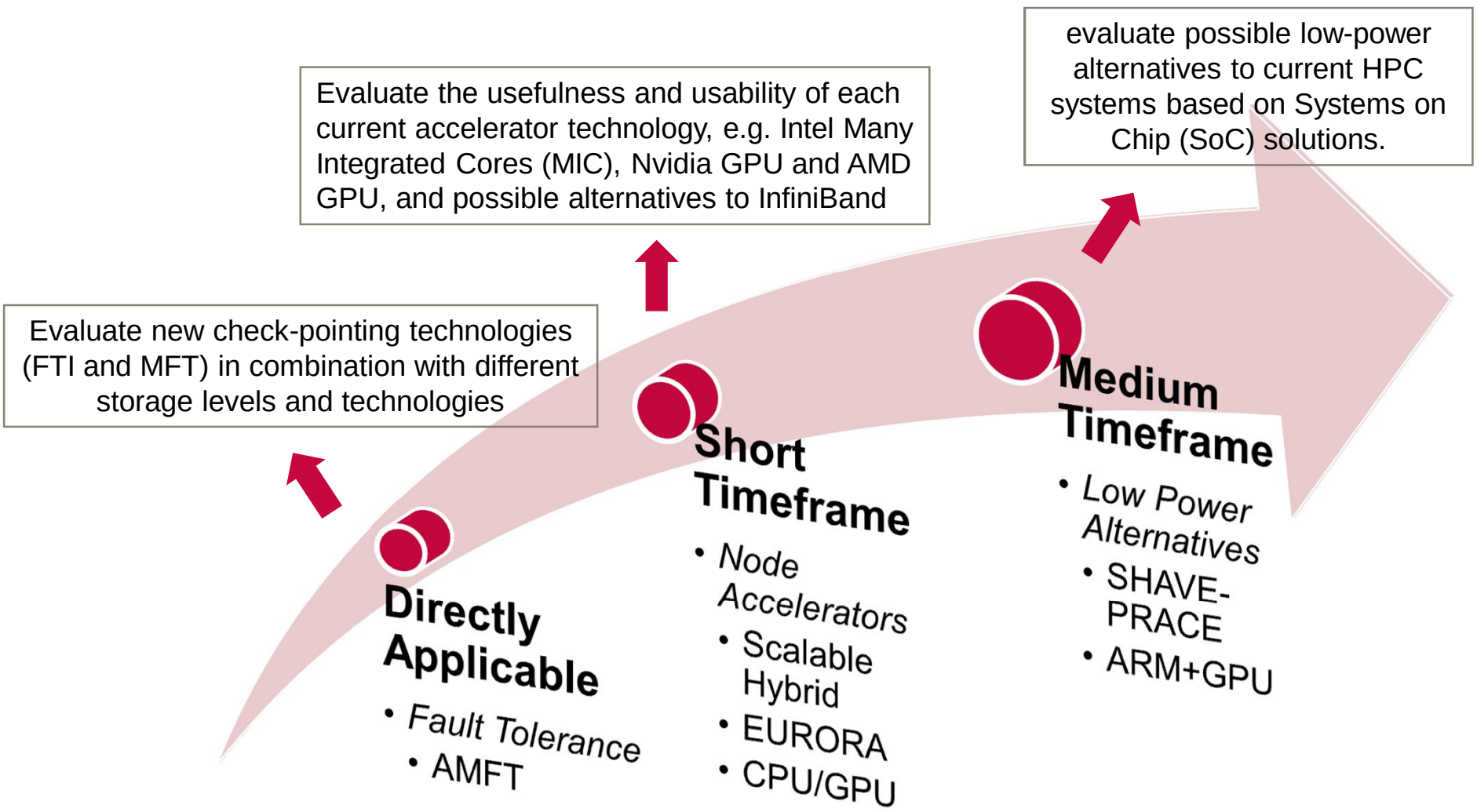


Linpack Performance Space



Linpack Energy Efficiency

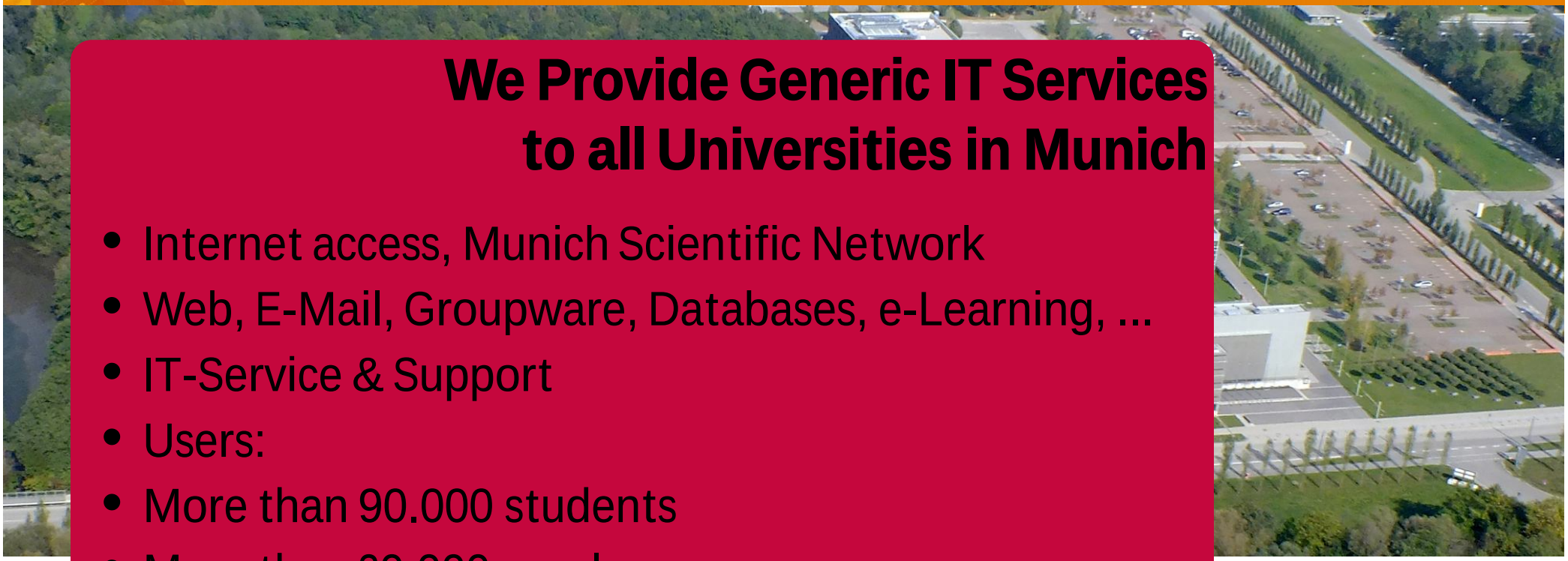



evaluate possible low-power alternatives to current HPC systems based on Systems on Chip (SoC) solutions.
Evaluate the usefulness and usability of each current accelerator technology, e.g. Intel Many Integrated Cores (MIC), Nvidia GPU and AMD GPU, and possible alternatives to InfiniBand
Evaluate new check-pointing technologies (FTI and MFT) in combination with different storage levels and technologies
Directly Applicable
• Fault Tolerance
• AMFT
Short Timeframe
• Node Accelerators
• Scalable Hybrid
• EURORA
• CPU/GPU
Medium Timeframe
• Low Power Alternatives
• SHAVE-PRACE
• ARM+GPU

Backup Material

PARTNERSHIP
FOR ADVANCED COMPUTING
IN EUROPE

PRACE

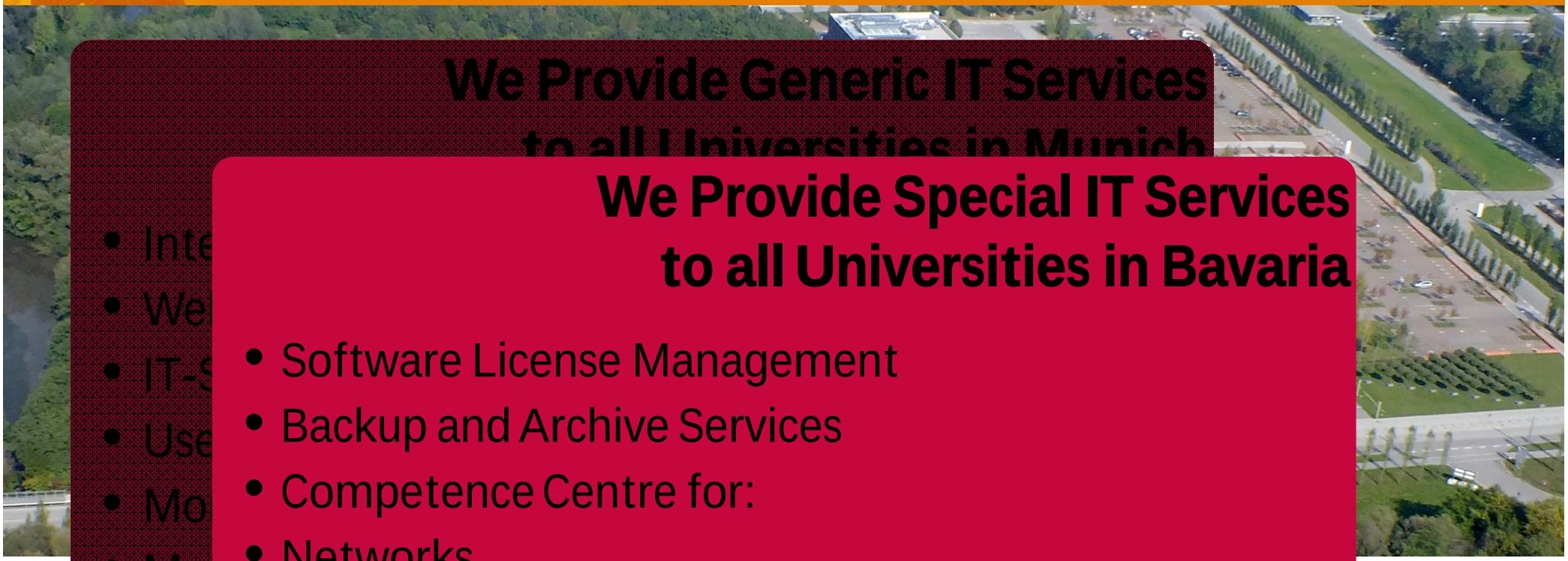


We Provide Generic IT Services to all Universities in Munich

- Internet access, Munich Scientific Network
- Web, E-Mail, Groupware, Databases, e-Learning, ...
- IT-Service & Support
- Users:
- More than 90.000 students
- More than 30.000 employees

PARTNERSHIP
FOR ADVANCED COMPUTING
IN EUROPE

PRACE



We Provide Generic IT Services to all Universities in Munich

- Inte
- We
- IT-S
- Use
- Mo
- Mo

We Provide Special IT Services to all Universities in Bavaria

- Software License Management
- Backup and Archive Services
- Competence Centre for:
- Networks
- High Performance and Grid Computing
- IT-Management


What we do...
Partnership
for Advanced Computing
in Europe
PRACE
We Provide Generic IT Services
We Provide Special IT Services
We Provide Supercomputing Services
to German and European Users
• Tier-0 Supercomputing Centre (SuperMUC)
• Member of the European HPC Community PRACE
• Research on Future HPC Systems:
• Hardware Architectures
• Programming Models & System Software
• HPC Centre Infrastructures

Processor heat density

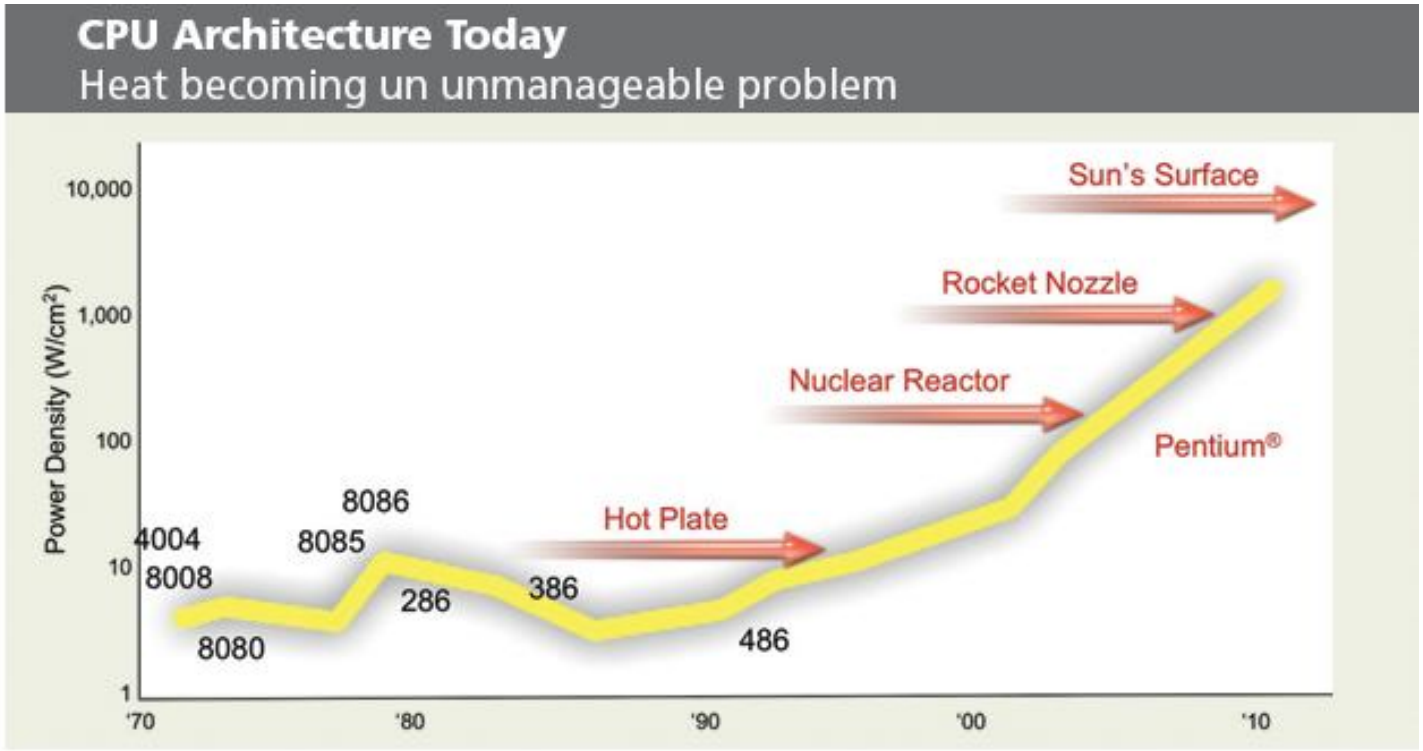
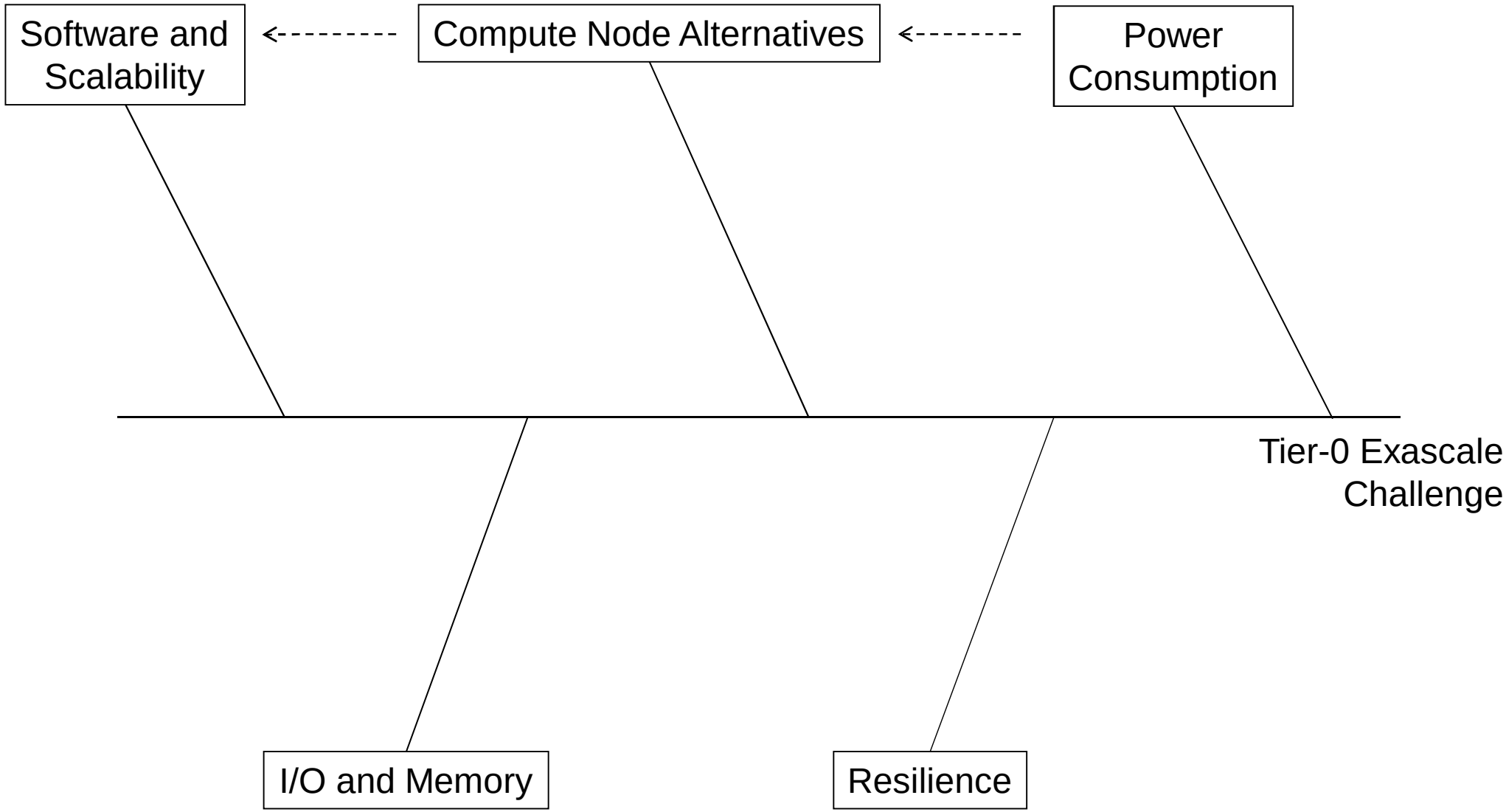
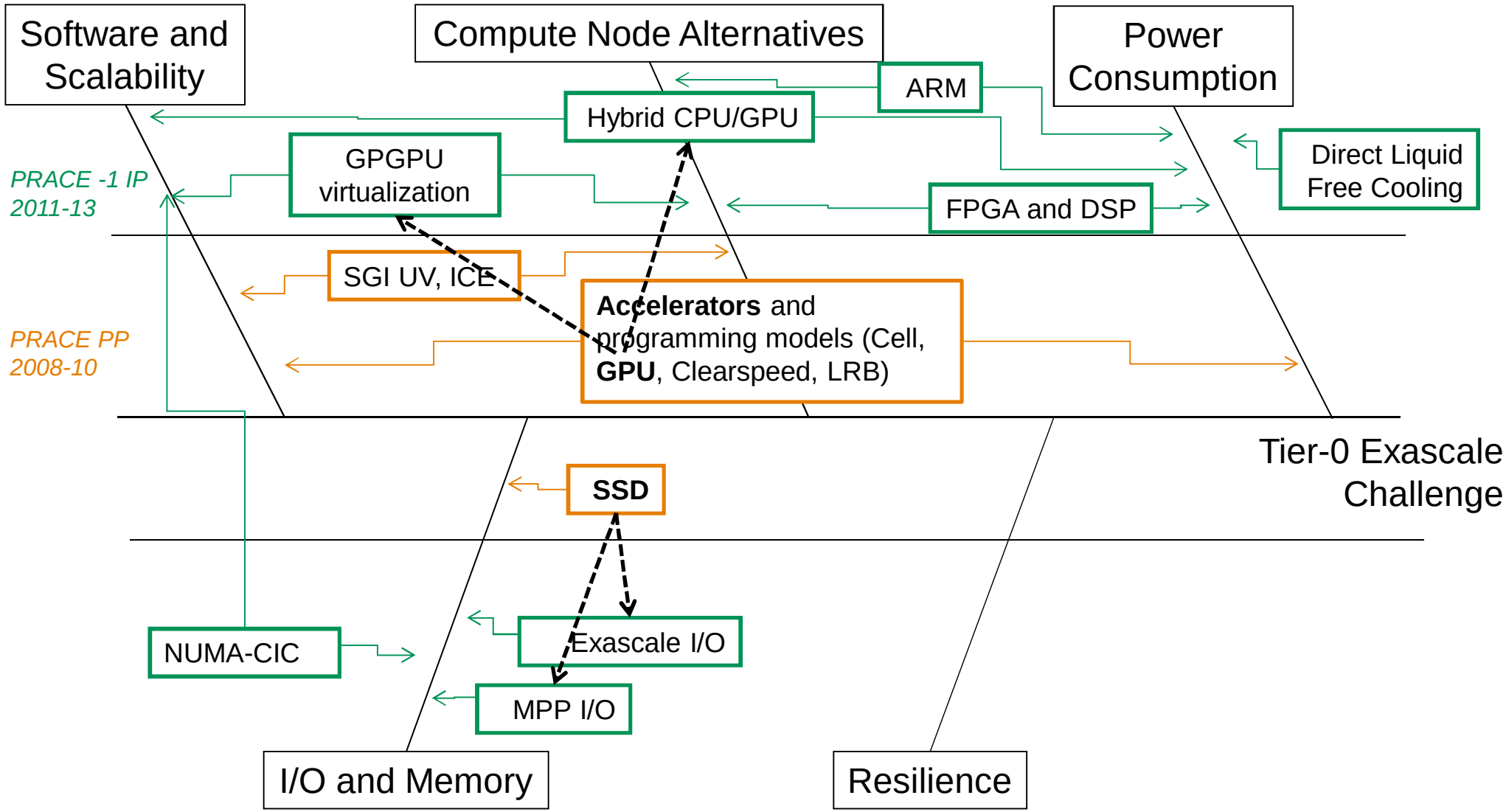
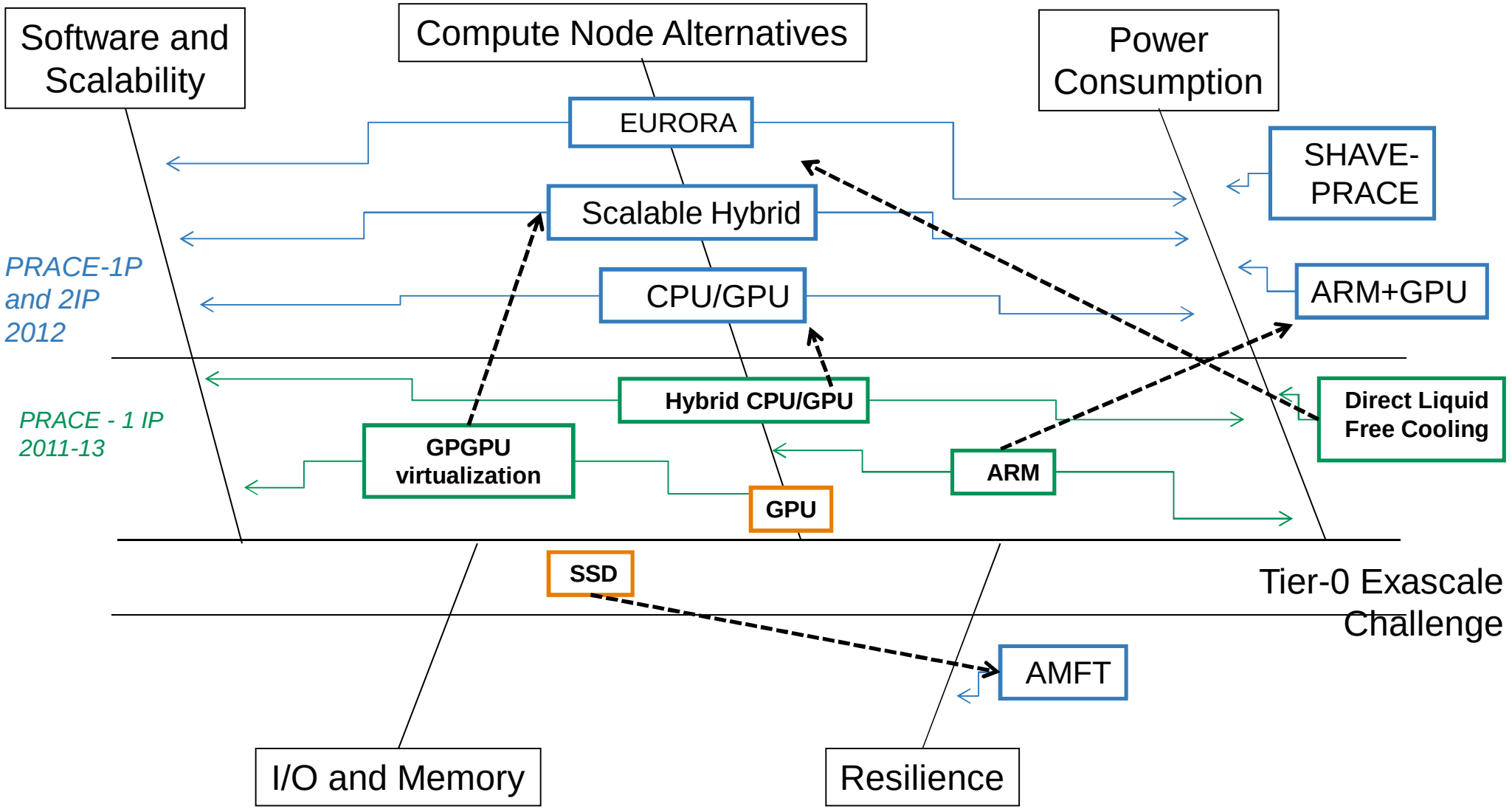


Figure 1. In CPU architecture today, heat is becoming an unmanageable problem. (Courtesy of Pat Gelsinger, Intel Developer Forum, Spring 2004)


Software and Scalability
Compute Node Alternatives
Power Consumption
Tier-0 Exascale Challenge
I/O and Memory
Resilience


Software and Scalability
Compute Node Alternatives
Power Consumption
ARM
Hybrid CPU/GPU
GPGPU virtualization
Direct Liquid Free Cooling
FPGA and DSP
PRACE -1 IP 2011-13
SGI UV, ICE
Accelerators and programming models (Cell, GPU, Clearspeed, LRB)
PRACE PP 2008-10
Tier-0 Exascale Challenge
SSD
NUMA-CIC
Exascale I/O
MPP I/O
I/O and Memory
Resilience


Software and Scalability
Compute Node Alternatives
Power Consumption
EURORA
Scalable Hybrid
CPU/GPU
SHAVE-PRACE
ARM+GPU
PRACE-1P and 2IP 2012
PRACE - 1 IP 2011-13
Hybrid CPU/GPU
GPGPU virtualization
GPU
ARM
Direct Liquid Free Cooling
SSD
AMFT
Tier-0 Exascale Challenge
I/O and Memory
Resilience

Summary

- Investigate novel technologies that might provide unique solutions for the Tier-0 exascale challenge
- Get first hand access to latest technology – before general availability
- Assess technologies, components, and systems for their suitability and maturity
- Guide development of future generation HPC hardware

Vision

- Strengthen Europe as leading innovator for high performance computing by working with European industry and technologies

- PRACE prototyping is a key element for the technological evolution of the PRACE Research Infrastructure

Thanks for Your attendance!

Questions?