E-Infrastructures
H2020-EINFRA-2014-2015

EINFRA-4-2014: Pan-European High Performance Computing Infrastructure and Services

PRACE-4IP

PRACE Fourth Implementation Phase Project

Grant Agreement Number: EINFRA-653838

D5.4
HPC Infrastructures Workshop #7

Final

Version: 1.0
Author(s): Huub Stoffers, SURFsara
Date: 20.10.2016

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: EINFRA-653838**	
	Project Title: PRACE Fourth Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: D5.4	
	Deliverable Nature: Report	
	Dissemination Level: PU *	**Contractual Date of Delivery:** 30 / 04 / 2017
		Actual Date of Delivery: 31 / 10 / 2016
	EC Project Officer: Leonardo Flores Añover	

* - The dissemination level are indicated as follows: **PU** – Public, **CO** – Confidential, only for members of the consortium (including the Commission Services) **CL** – Classified, as referred to in Commission Decision 2991/844/EC.

Document Control Sheet

Document	**Title: HPC Infrastructures Workshop #7**	
	ID: D5.4	
	Version: 1.0	**Status:** ***Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2010	
	File(s): D5.4	
Authorship	**Written by:**	Huub Stoffers, SURFsara
	Contributors:	François Robin (CEA), Jean-Philippe Nominé (CEA), Eric Boyer (GENCI), Torsten Wilde (LRZ), Michael Ott (LRZ), Norbert Meyer (PSNC)
	Reviewed by:	Enver Ozdemir, UHEM; Veronica Teodor, FZJ.
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	20/09/2016	Template, document front matter	Huub Stoffers
0.2	20/09/2016	Introduction, sessions 1-5, 7, PRACE sites, findings	Huub Stoffers

0.3	21/09/2016	All sessions and sections, a few things missing in particular sections. List of external documents and list of abbreviations incomplete	Huub Stoffers
0.4	30/09/2016	Corrections and additions after proof readings of draft v. 0.3 by Philippe Segers and Gert Svensson. Merging of late presenter feedback on some summaries	Huub Stoffers
0.5	04/10/2016	Full version, draft for internal review	Huub Stoffers
1.0	20/10/2016	Full version, after PRACE internal review	Huub Stoffers

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, HPC Facility, Datacentre, Energy efficiency, Cooling, Datacentre Monitoring, Exascale

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] http://www.prace-project.eu

[2] http://geopm.github.io/geopm, Github source code repository for the Intel GEO project

[3] Integrated Energy-Aware Management of Supercomputer Hybrid Cooling Systems, Christian Conficoni, Andrea Bartolini, Andrea Tilli, Carlo Cavazzoni, and Luca Benini to appear in IEEE Transactions on Industrial Informatics

List of Acronyms and Abbreviations

AISBL	Association International Sans But Lucratif (legal form of the PRACE-RI)
ARM	Formerly known as Acorn RISC Machine, family of processors developed by a British company ARM Holdings
BMC	Baseboard Management Controller
BMS	Building Management System
CORAL	Collaboration of Oak Ridge, Argonne and Lawrence Livermore Laboratory

CPU	Central Processing Unit
CRAC	Computer Room Air Conditioner
CUDA	Compute Unified Device Architecture (NVIDIA)
DC	Data Centre (or Direct Current)
DCEM	Datacentre Energy Management or Data processing and Communications Energy Management (ETSI KPI)
DCP	Data Centre Performance
DLC	Direct Liquid Cooling
DOE	Department of Energy (US)
DP	Double Precision, usually 64-bit floating point numbers
DRAM	Dynamic Random Access memory
EC	European Commission
EEHPCWG	Energy Efficient HPC Working Group
EESI	European Exascale Software Initiative
ESFRI	European Strategy Forum on Research Infrastructures
ETSA	European Telecommunications Standards Institute
FDR	Fourteen Data Rate, 14 Gb/s data rate per InfiniBand lane
FP	Floating-Point
FPGA	Field Programmable Gate Array
GB	Giga (= 2^{30} ~ 10^9) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network. The current EC-funded project as of 2015 is GN4.
GFlop/s	Giga (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GPFS	General Parallel File System, a high performance clustered parallel file System, developed by IBM
GPU	Graphic Processing Unit
HDEEM	High Definition Energy Efficiency Measurement
HET	High Performance Computing in Europe Taskforce. Taskforce by representatives from European HPC community to shape the European HPC Research Infrastructure. Produced the scientific case and valuable groundwork for the PRACE project.
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPE	Hewlett-Packard Enterprise
HPL	High Performance LINPACK
IB	InfiniBand
IBM	Formerly known as International Business Machines
IPMI	Intelligent Platform Management Interface
I/O	Input/Output
ISC	International Supercomputing Conference; European equivalent to the US based SCxx conference. Held annually in Germany.
KB	Kilo (= 2^{10} ~10^3) Bytes (= 8 bits), also Kbyte
KNL	Knight Landing (Intel Xeon Phi product)
KPI	Key Performance Indicator
LANL	Los Alamos National Laboratory (USA)
LINPACK	Software library for Linear Algebra

LLNL	Lawrence Livermore National Laboratory (California, USA)
MB	Management Board (highest decision making body of the project)
MB	Mega (= 2^{20} ~ 10^6) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MPI	Message Passing Interface
MVA	Mega Volt-Ampere
NERSC	National Energy Research Scientific Computing Center (California, USA)
NREL	DOE's National Renewable Energy Laboratory (Colorado, USA)
NVRAM	Non-Volatile Random Access Memory
OLCF	Oak Ridge Leadership Computing Facility at ORNL (USA)
ORNL	Oak Ridge National Laboratory (Tennessee, USA)
PA	Preparatory Access (to PRACE resources)
PB	Peta (= 2^{50} ~ 10^{15}) Bytes (= 8 bits), also PByte
PCP	Pre-Commercial Procurement
PCIe	Peripheral Component Interconnect express, also PCI-Express
PCI-X	Peripheral Component Interconnect eXtended
PFlop/s	Peta (= 10^{15}) Floating point operations (usually in 64-bit, i.e. DP) per second, also PF/s
PMBus	Power Management Bus
PMI	Platform Management Interface
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PRACE 2	The upcoming next phase of the PRACE Research Infrastructure following the initial five year period.
PRIDE	Project Information and Dissemination Event
PSU	Power Supply Unit
PUE	Power Usage Effectiveness
RAPL	Running Average Power Limit
RCU	Rack Control Unit
RI	Research Infrastructure
SLURM	formally known as Simple Linux Utility for Resource Management, job scheduler for Linux kernel
SoC	System on a chip
TB	Technical Board (group of Work Package leaders)
TB	Tera (= 2^{40} ~ 10^{12}) Bytes (= 8 bits), also TByte
TCO	Total Cost of Ownership. Includes recurring costs (e.g. personnel, power, cooling, maintenance) in addition to the purchase cost
TDP	Thermal Design Power
TFlop/s	Tera (= 10^{12}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
Tier-1	Denotes the national or topical level of a conceptual pyramid of HPC systems

List of Project Partner Acronyms

BADW-LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Germany (3rd Party to GCS)
BILKENT	Bilkent University, Turkey (3rd Party to UYBHM)
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion, Spain
CaSToRC	Computation-based Science and Technology Research Center, Cyprus
CCSAS	Computing Centre of the Slovak Academy of Sciences, Slovakia
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3rd Party to GENCI)
CESGA	Fundacion Publica Gallega Centro Tecnológico de Supercomputación de Galicia, Spain, (3rd Party to BSC)
CINECA	CINECA Consorzio Interuniversitario, Italy
CINES	Centre Informatique National de l'Enseignement Supérieur, France (3rd Party to GENCI)
CNRS	Centre National de la Recherche Scientifique, France (3rd Party to GENCI)
CSC	CSC Scientific Computing Ltd., Finland
CSIC	Spanish Council for Scientific Research (3rd Party to BSC)
CYFRONET	Academic Computing Centre CYFRONET AGH, Poland (3rd party to PNSC)
EPCC	EPCC at The University of Edinburgh, UK
ETHZurich (CSCS)	Eidgenössische Technische Hochschule Zürich – CSCS, Switzerland
FIS	FACULTY OF INFORMATION STUDIES, Slovenia (3rd Party to ULFME)
GCS	Gauss Centre for Supercomputing e.V.
GENCI	Grand Equipement National de Calcul Intensiv, France
GRNET	Greek Research and Technology Network, Greece
INRIA	Institut National de Recherche en Informatique et Automatique, France (3rd Party to GENCI)
IST	Instituto Superior Técnico, Portugal (3rd Party to UC-LCA)
IUCC	INTER UNIVERSITY COMPUTATION CENTRE, Israel
IT4I	IT4Innovations National supercomputing centre at VŠB-Technical University of Ostrava, Czech Republic
JKU	Institut fuer Graphische und Parallele Datenverarbeitung der Johannes Kepler Universitaet Linz, Austria
JUELICH	Forschungszentrum Juelich GmbH, Germany
KTH	Royal Institute of Technology, Sweden (3rd Party to SNIC)
LiU	Linkoping University, Sweden (3rd Party to SNIC)
NCSA	NATIONAL CENTRE FOR SUPERCOMPUTING APPLICATIONS, Bulgaria
NIIF	National Information Infrastructure Development Institute, Hungary
NTNU	The Norwegian University of Science and Technology, Norway (3rd Party to SIGMA)
NUI-Galway	National University of Ireland Galway, Ireland
PRACE	Partnership for Advanced Computing in Europe aisbl, Belgium
PSNC	Poznan Supercomputing and Networking Center, Poland
RISCSW	RISC Software GmbH

RZG	Max Planck Gesellschaft zur Förderung der Wissenschaften e.V., Germany (3rd Party to GCS)
SIGMA2	UNINETT Sigma2 AS, Norway
SNIC	Swedish National Infrastructure for Computing (within the Swedish Science Council), Sweden
STFC	Science and Technology Facilities Council, UK (3rd Party to EPSRC)
SURFsara	Dutch national high-performance computing and e-Science support center, part of the SURF cooperative, Netherlands
UC-LCA	Universidade de Coimbra, Labotatório de Computação Avançada, Portugal
UCPH	Københavns Universitet, Denmark
UHEM	Istanbul Technical University, Ayazaga Campus, Turkey
UiO	University of Oslo, Norway (3rd Party to SIGMA)
ULFME	UNIVERZA V LJUBLJANI, Slovenia
UmU	Umea University, Sweden (3rd Party to SNIC)
UnivEvora	Universidade de Évora, Portugal (3rd Party to UC-LCA)
UPC	Universitat Politècnica de Catalunya, Spain (3rd Party to BSC)
UPM/CeSViMa	Madrid Supercomputing and Visualization Center, Spain (3rd Party to BSC)
USTUTT-HLRS	Universitaet Stuttgart – HLRS, Germany (3rd Party to GCS)
WCNS	Politechnika Wroclawska, Poland (3rd Party to PNSC)

Executive Summary

The 7th European Workshop on HPC Centre Infrastructures was held at the LRZ building, in Garching near München, Germany, between April 19, 2016 – April 22, 2016. BSC, CEA, CSCS, LRZ, PDC-KTH, and PSNC have collaborated in the committee organising the workshop.

This workshop, upon invitation only, was very successful, with 78 participants coming from Europe, America, Australia, and Asia.

In addition to the standard topics (presentations from vendors and sites), the workshop focused on two important interrelated subjects:

- Monitoring of datacentre infrastructure;
- Analysis of power usage efficiency, i.e., the analysis of power usage in relation to other aspects of datacentre functionality.

The PRACE closed session, held at the end of the workshop, gathered attendees from Tier-0 and Tier-1 sites. Five site representatives gave an update on specific datacentre infrastructure developments and the session gave the opportunity for exchanges between experts from the PRACE sites.

The workshop made the identification of important trends and assessments on the situation in Europe in terms of energy efficient large HPC centres possible.

1 Introduction

The 7th European Workshop on HPC Centre Infrastructures was held at LRZ, in Garching near München, Germany, between April 19, 2016 – April 22, 2016. BSC, CEA, CSCS, LRZ, PDC-KTH, and PSNC, have collaborated in the committee organising the workshop, using some PRACE-4IP/WP5 manpower for this purpose, as well as PRACE sponsorship. The program committee consisted of the following members:

- Javier Bartlome, BSC, Spain
- Ladina Gilly, CSCS, Switzerland
- Herbert Huber, LRZ, Germany
- Norbert Meyer, PSNC, Poland
- Jean-Philippe Nominé, CEA, France
- François Robin, CEA, France
- Gert Svensson, KTH, Sweden

This workshop, upon invitation only, was very successful, with 78 participants coming from 14 countries in Europe, USA, Australia, Japan, and Singapore (See Figure 1).

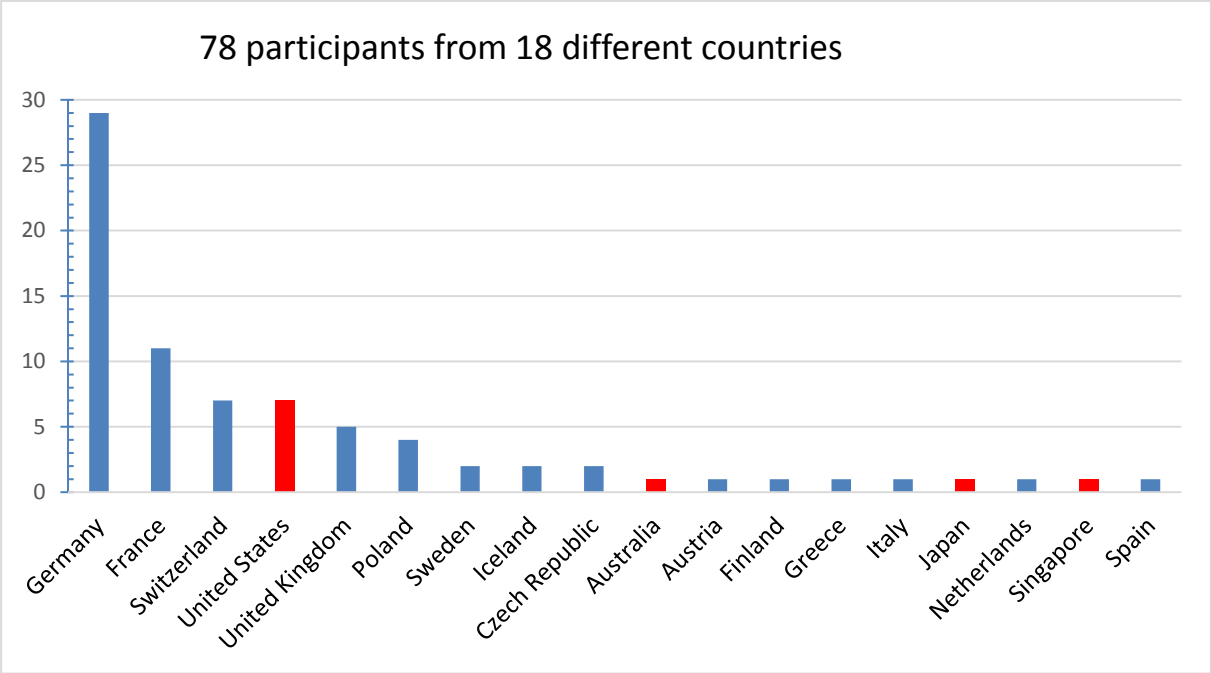

Figure 1: Nationality of workshop participants

28 attendees of the Workshop were associated with PRACE sites (Tier-0 as well as Tier-1) or the PRACE organisation. The two days of plenary sessions mixed these attendees from PRACE with experts from other invited sites and experts from HPC system integrators and other producers of technologies relevant for HPC and/or HPC datacentre infrastructure. 13 non-PRACE European sites and 7 non-European sites[1] participated in the sessions, as well as 13 vendor organisations. The distribution of participants over the diverse backgrounds or associations is shown in Figure 2.

This document reports on the proceedings of the workshop and follows the chronological order of activities during the workshop. It includes summaries of the presentations given at

[1] This includes Microsoft, which was present in its capacity of worldwide datacentre entrepreneur, rather than as a software vendor, as a "non-European datacentre site"

the workshop, including updates from several PRACE sites, on the current state of their HPC datacentre infrastructure and the way in which it is monitored and managed. Vendors of HPC processor architectures and system integrators report on their views on the road map for HPC, i.e.: on what it may have to be accommodated in the datacentre in the near future.

Besides the usual site updates, this year's workshop aimed to focus on data centre monitoring. This focus is reflected in some of the presentations given by vendors, and in the following presentations:

- The Energy Efficient HPC Working Group (EEHPCWG), supported by the US DOE (http://eehpcwg.lbl.gov/), an active and influential initiative with a growing interest for collaboration with this workshop and present now for the fourth time, reporting on datacentre dashboards;
- A keynote session with Microsoft reporting, in the role of global data centre entrepreneur, on the evolution in the design of its data centres and global data centre provisioning;
- The presentation of (preliminary) results of a survey, conducted by PSNC, on data centre monitoring at PRACE sites;
- An in-depth presentation on the development of an integral system for monitoring and consolidation of monitoring data by LRZ;
- A keynote session with Intel reporting on its global energy optimisation program, which aims to provide a platform for monitoring the energy efficiency of applications.

At the end of the workshop participants from PRACE sites gathered in the PRACE closed session in which site updates were presented and planning for further development of PRACE-4IP WP5 took place.

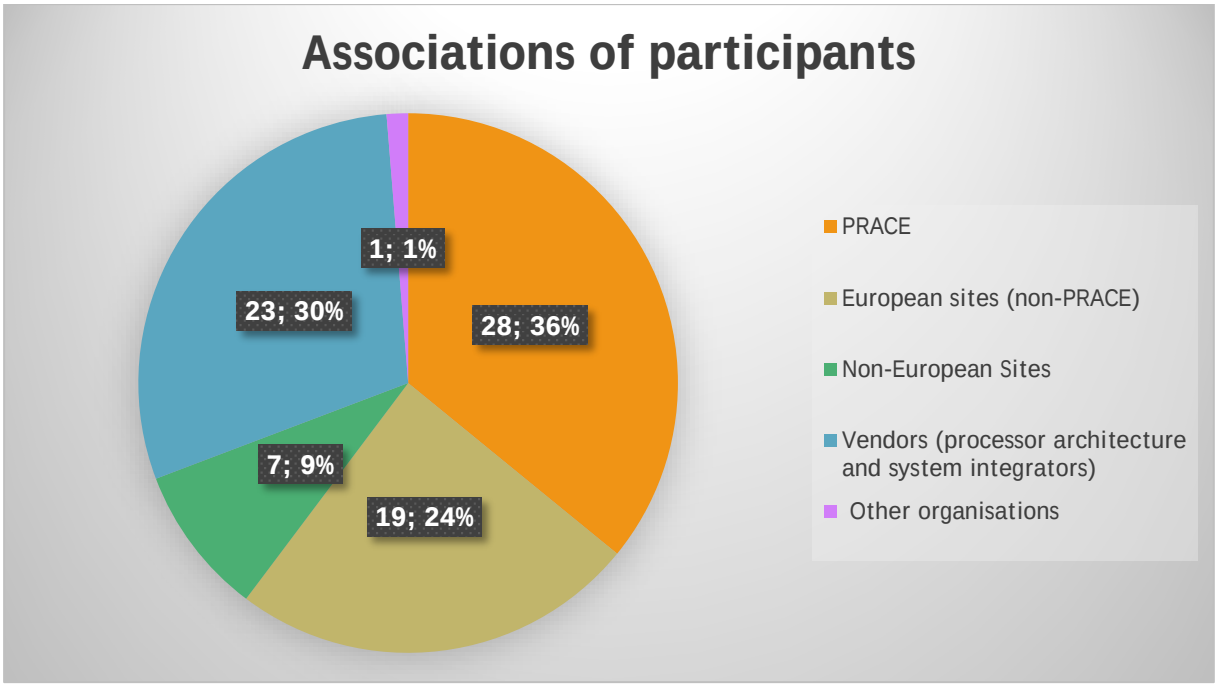

Figure 2: Background, association, or expertise of workshop participants

The document intends to give an audience interested in developments in HPC data centres and energy efficiency an overview based on reports from major HPC centres and vendors of HPC technology.

Figure 3: 7th Workshop participants, LRZ, Germany

2 Programme, Content and Speakers

Session I – Keynote and Site updates US, Chair: François Robin

- Keynote 1: Data centre monitoring and analysis at LRZ, Torsten Wilde and Detlef Labrentz (LRZ)
- ORNL, Jim Rogers (ORNL)
- NERSC, Brent Draney (NERSC)

Session II - Site updates Asia, Chair: Torsten Wilde

- A*STAR, Lim Ching Kwang (A*STAR CRC)
- RIKEN, Fumiyoshi Shoji (RIKEN AICS)

Session III – Site updates Europe (Part 1), Chair: Norbert Meyer

- Keynote 2: Microsoft Azure datacentres, Ralf Wigand, Senior Program Manager Global Ecosystems (Microsoft)
- ECMWF, Andy Gundry (ECMWF)
- CEA, Jean-Marc Ducos and Frédéric Souques (CEA)

Social Event - Guided City-tour Freising, dinner at the Bräustüberl, Weihenstephan

Session IV – Energy Efficiency and Monitoring

- Keynote 3: GEO Power management Framework, Jonathan Eastep (Intel)
- Energy efficient HPC Workgroup, Natalie Bates (EEHPCWG / LLNL)

Session V – Vendors of future HPC processor architectures, Chair: Jean-Philippe Nominé

- ARM, Geraint North (ARM)
- IBM, Klaus Gottschalk (IBM)
- Intel, Andrey Semin (Intel)
- NVIDIA, Axel Koehler (NVIDIA)

Session VI – Site updates Europe (Part 2), Chair: Norbert Meyer

- Data Centre Monitoring at PRACE sites, Norbert Meyer (PSNC)
- CSCS, Ladina Gilly (CSCS)
- LRZ, Herbert Huber (LRZ)

Session VII – System integrators, Chair: Ladina Gilly

- Atos/Bull, Jean-Philippe Sibers (Atos)
- Cray, Wilfried Oed (Cray)
- HPE, Sammy Zimmerman (HPE)
- Lenovo, Luigi Brochard (Lenovo)

LRZ Datacentre Tour

Social Event – Dinner at Gasthof Mühle, Ismaning

PRACE 4IP/WP5 Session (reserved for PRACE partners), Chair: François Robin

- CINECA (Italy)

- FZJ (Germany)
- GRNET (Greece)
- IT4Innovation (Czech Republic)
- SURFsara (The Netherlands)

3 Session I – Keynote 1 and Site updates US

3.1 Keynote 1: Datacentre monitoring and Analysis at LRZ – Torsten Wilde, Detlef Labrentz

LRZ currently has 3160 m^2 of IT equipment floor space and 6393 m^2 of data centre infrastructure floor space. The data centre has a power supply of 2 x 10 MW. LRZ is powered entirely by renewable energy and has committed itself to a long term global optimisation strategy to improve the energy efficiency of the data centre. Apart from an environmental motivation, there is the need to reduce the total cost of ownership. Much is to be gained by improving the energy efficiency: in the year 2000 the cost of 1 kWh for LRZ was 0.07€. In the year 2015 it was 0.161€. The annual power bill currently is over 5 Million €. As the German "Energie Wende" is in part financed by raising taxes on energy usage, the cost of electricity is bound to continue rising this year and beyond.

LRZ discerns four pillars for its global optimisation strategy. For each pillar a measurable optimisation goal, a key performance indicator, has been defined:

1. Building infrastructure; the goal is to improve the PUE;
2. IT systems hardware, including network and storage hardware; the goal is to reduce hardware power consumption;
3. System software, including batch systems and compilers; the goal is to optimise resource usage, tune the system;
4. Application usage of CPUs, GPUs, FPGAs; the goal is to optimize application performance.

The comprehensive approach implies that many data points, covering many different aspects of the datacentre, have to be collected and consolidated into an environment that allows data processing, analyses and reporting. LRZ has developed the PowerDAM framework for this purpose, which at present interfaces with three of the four pillars: building infrastructure, systems hardware, and systems software - mainly the batch systems and IPMI sensors. Many relevant sensors are present in all kinds of equipment, including building automation systems, in addition, LRZ has built many additional sensors into the datacentre itself.

PowerDAM does not change anything in the data collected at the data source level. It merely extracts those data, with a time resolution of 1 minute, and imposes a logical structure on the individual data sources by grouping them into hierarchies of systems with racks and nodes, or systems with circuits and devices. An eclectic approach towards importing had to be taken as many of the circuits and devices containing relevant data do keep logs and/or allow some form of monitoring, but the design never envisaged the incorporation of the monitoring into a larger framework. During the development of PowerDAM – which is ongoing – many problems of interoperability, sometimes down to the level of character encoding – had to be solved.

Despite the fact that LRZ has taken care to check and properly calibrate many sensors, not all data gathered by PowerDAM are of equal quality and accuracy. Building automation systems primarily collect data to use them in control decisions. Changes in values that are deemed irrelevant for such decisions are often not recorded in the logs. In addition, sensor readout may be limited by connectivity. As this constitutes problems for some types of data analysis, options for data verification and validation are now being explored.

The LRZ consolidation of its monitoring data has already proven useful for a variety of analyses, ranging from analyses of energy to solution and power usage of specific HPC

applications, to analyses of the mode of operation and energy efficiency of pumps. Though PowerDAM does not interface with applications directly, analyses of application behaviour and power usage by applications is possible via the batch system, as the batch system produces a track records of resources (nodes, racks) that were used in application runs.

3.2 ORNL – Jim Rodgers

ORNL is preparing for OLCF-4 (Oak Ridge Leadership Computing Facility #4) system to be called "Summit" which is expected to arrive in 2017. ORNL's current largest system is a "Titan", a 200 cabinet Cray XK7 system. Summit will be an IBM + NVIDIA system, with multiple Power9 CPUs and multiple NVIDIA Volta GPUs in a single node communicating over NVIDIA NVLink. NVLink enables "Unified Memory", which is expected to considerably simplify taking full advantage of the GPU power by programmers, as they no longer need to worry whether data resides within the CPU or the GPU. A non-blocking fat-tree dual rail EDR InfiniBand interconnect by Mellanox is the medium for internode communications and node file system access. Summit will deliver more than 150 PFlop/s of compute capacity. The aggregate bandwidth of 120 PB of IBM GPFS file system storage will be around 1TB/s.

In 2016 two separate early access systems – "stepping stones to Summit" – of the IBM Garrison server family will be delivered. These have 2 IBM Power8 processors and 4 GPUs per node and NVLink. The early access systems will support system administrative preparative activities and application readiness activities. The water cooled version of Garrison allows prototyping of the water delivery system anticipated for Summit.

A new facility is built to accommodate Summit – and systems after Summit that in many ways are different from the current. There will be no raised floor, just a cement slab capable of carrying the weight. Consequently, everything – water, power, network – has to be delivered to the cabinets from above. Rear door heat exchange in addition to direct liquid cooling is feasible, but there will be no additional air conditioning in the computer room.

The current average power use at ORNL is about 11 MW. The average power use of Titan alone is 5.2 MW. The expected power usage of the Summit system alone is 9.6 MW. Eventually, up to 20 MW extra will be needed. In reality, ORNL is aiming at a scalable design from 10 to 20 MW to accommodate Summit and future systems. Jim Rogers labels the new electrical installation as "still pretty standard" but the mechanical design as "not so standard". Salient innovative features in this area are:

- Pipes made of polypropylene-random (PP-R) thermoplastic in the secondary loop;
- Reverse-return (Tichelmann) mechanical piping systems providing balanced piping losses without the need for flow control or balance valves;
- There will be supplementary chillers for very warm days, but warm water inlet of 20ºC provides substantial "free cooling" days.

The facility's PUE is expected to improve considerably: most energy in cooling will be spent on pumps.
Jim Rodgers envisages that the lifetime of future HPC systems will go to 6 or 7 years, rather than the 3 or 4 years that were usual. Consequently, there are more worries about corrosive parts and electronics being damaged over time by exposure to acidic gases. Considerable thought has been given to filtering air and monitoring air quality. In addition, computer rooms have overpressure to reduce infiltration whenever doors are opened.

3.3 NERSC – Brent Draney

NERSC current HPC systems are “Edison”, and “Cori”, both Cray XC systems. Brent Draney points out that the delivery of the Intel Xeon PHI Knights HILL (KNH) based “Aurora” system is delayed and that a system “Theta” will be accommodated at NERSC to bridge the gap. Theta is based on Intel Xeon PHI, Knights Landing (KNL) and has a compute capacity of over 3 TFlop/s. Its peak power usage is around 1.7 MW.

NERSC has a new home, a new datacentre overseeing the San Francisco Bay Area. This region is notorious for its sudden bursts of considerable seismic activity. The design of a datacentre for this area needs to take this into account. The new NERSC building has a data centre floor of 20,000 ft^2 that is seismically isolated from the building. The raised floor is hanging from a structure and can move in swing, move in the horizontal plane. About 45 cm of motion is allowed. In an event of seismic activity, the equipment on the floor will be “shaken, but not stirred”. Figure 4 shows some implementation details of the seismic isolation of the raised floor at NERSC from beneath.

Figure 4: The raised floor at NERSC, seismically isolated from the building

The new facility is designed for a PUE of 1.1. It currently has 5 power substations of 2.5 MW, providing a 12.5 MW redundant power feed. The design allows expansion to 11 substations that provide a 27.5 MW redundant feed (42 MW non-redundant). 10 MW of liquid cooling capacity is currently provided by cooling towers. The design allows an expansion up to 20 MW. In addition, 2 MW of 100% outside air capable air cooling capacity is available. Both water and air quality are monitored for various sorts of contamination. Computer room waste heat is used to heat the office floors of the building.

There is a network between the old and the new datacentre facility with a bandwidth of 400 Gbps. Live migration of GPFS file systems by means of induced failover is possible and has indeed been used to migrate about 20 PB of data flawlessly in about 11 days.

4 Session II – Site updates Asia

4.1 A*STAR CRC – Lim Ching Kwang

A*CRC (A*STAR Computational Resource Centre) is the main HPC data centre for the A*STAR research institution of Singapore. It is not a research institution itself, but it studies state-of-the-art HPC technologies, and engages in discussions with vendors on the merit of new trends in HPC.

It runs a state-of-the-art data centre at the 17th floor of a high rise building. A*STAR also hosts the HPC system of Singapore's National Super Computing Center (NSCC). The combined computational power of the systems hosted in 2016 is slightly above 500 TFlop/s. With the new NSCC HPC system, A*STAR moved from chiller-supported cooling to ASHRAE W4 (hot water, chiller-less) direct liquid cooling. The new 1PFlops system (built by Fujitsu) uses direct warm water cooling (based on Asetek) with an inlet temperature of 40º C and an outlet temperature of 45ºC. It is GPU accelerated and uses the PBS Pro Scheduling system. The cooling is done via dry coolers. The goal for the data centre upgrade is a PUE of 1.4 from a current PUE of 2.5.

The power distribution is modular and from the possible maximum of 2MW only 1MW are installed. The current data centre power draw is 900kW.

To save more energy the data centre and offices are using intelligent lighting system. This system uses collected data (occupancy, daylight, task tuning) to adjust the artificial lighting according to work tasks. An integrated heat map software can show occupancy and can help to identify areas that are mostly used which can help with office space planning.

An interesting technology project driven by A*CRC is the InfiniCortex network which is a worldwide InfiniBand network based on Obsidian Strategic's Longbow technology.

4.2 RIKEN AICS – Fumiyoshi Shoji

RIKEN AICS (Advanced Institute for Computational Science) is the main HPC Data Centre in Japan. It hosts the K-computer which was #1 on the Top500 list in 2011 (and still was #5 as of June 2016 with Rmax at 10.5 PFlop/s), and which is developed in a collaboration between RIKEN and Fujitsu. The facility has an average power consumption of 14.7MW of which 10.2-12.2 MW is used by the K-computer and 2.3-3.6 MW is used by the data centre cooling infrastructure.

A unique property is the use of two 5 MW Gas Turbines used for tri-generation (electricity, heating, cooling). The Gas turbine efficiency increases with generated power and the gas price affects its usefulness. For a long time the gas piece was higher than the electricity price and the power generation level was suppressed accordingly. The facility uses 70% water cooling and 30% air cooling. Using the gas co-generation to generate cold water via adsorption chillers helps a lot to achieve higher energy efficiency, by re-using waste heat, and to keep cooling costs down.

Early 2016 the cooling tower building was modified to improve efficiency. Another energy saving activity was to reduce the number of active fans in the installed air handlers from two to one.

Future plans include the use of free cooling, air storage battery, a thermal storage pool, and the implementation of a smart energy management system.

5 Session III – Site updates Europe (Part 1)

5.1 Microsoft Azure Datacentres – Ralph Wigand

While Microsoft is still primarily seen by many as a software company, its global data centre footprint is, in fact, huge. As of April 2016, Microsoft operates more than 140 data centres in more than 40 countries, and the number is still growing rapidly. Its global network connecting its data centres is also one of the two largest in the world. Its datacentres are organised in 30 cloud regions. All regions have a replicating peer region – e.g. Dublin and Amsterdam are peer regions. Globally, Microsoft's current investment in data centre infrastructures is above 15 billion US Dollars and more than 550 million users are registered in Azure's Active Directory user administration.

Microsoft can offer various levels of "software defined availability" of clouds and virtual machines that can include, among other things, synchronous replication in-region, asynchronous cross-region replication, traffic management for cross-region HA, allocation across fault domains, and coordinated updates within availability sets.

Looking back in time, an evolution spanning five generations of Microsoft datacentres is discernible. The company's policy is to not disclose any specific figures on actual power consumption of its datacentres, but they can specify how the PUE of its datacentres has improved with each generation. Every generation has a name that matches the typical style of deployment of that generation. Table 1 provides an overview of the generations, the period in which they were deployed, and the typical PUE.

Gen.	Period	PUE	Name, typical characteristic
I	1989 - 2005	2.0+	"Colocation", server by server deployment
II	2005 - 2009	1.4 – 1.6	"Density", rack by rack deployment, higher density
III	2009 - 2012	1.2 – 1.5	"Containment", deployment in containers
IV	2012 - 2015	1.12 – 1.2	"Modular"
V	2015 -	1.07 – 1.09	"Software Defined"

Table 1: Generations of Microsoft Datacentres

1st generation datacentres have not existed for some time and 2nd generation datacentres are disappearing. The trend over the generations is for more infrastructure convergence and to more resilient software. Separate datacentres used to be built for specialized server farms, for HotMail, for Bing, etc. Some of them are still running production but new datacentres are no longer deployed in that way. Microsoft now deploys one platform for all services and this has dramatically improved failover capability and at the same time reduced cost.

HPC sites typically operate one or perhaps two datacentres. This workshop bears witness to the fact that many HPC sites are very capable of performing detailed causal analyses of phenomena observed in their specialised centres. While Microsoft's datacentres are not for HPC, operating and monitoring that many datacentres make a different, more statistical, approach feasible. Microsoft has performed "big data" analytics on massive datasets comprising event logs of numerous components and pieces of equipment in its own datacentres. Making use of a machine learning approach to classify "patterns in all sorts of datacentre events", they are now in a position in which they use current monitoring data to predict which components have a very high probability of failing in the near future. This makes quick proactive replacement of components and/or quick proactive data migration possible.

Microsoft is also involved in several highly experimental datacentre-related research projects. In an energy innovation project, the possibilities of in-rack fuel are explored in which natural gas is converted directly into electricity. In project Natick, small "datacentres" or equipment containers are put in seawater. The underwater datacentres may be cooled easier and cheaper than those on land. In addition, in some coastal areas the tidal motion of the surrounding water might be usable very directly to generate power.

5.2 ECMWF – Andy Gundry

The European Centre for Medium-Range Weather Forecasts (ECMWF) specialises in medium range, i.e. for up to about 2 weeks ahead, global numerical weather prediction. It generates two high resolution forecasts and 51 Model ensembles from 149 million observations per day, collected by a high number of very heterogeneous sources, such as ships, buoys, and several types of satellites. For this purpose it runs a 24/7 operational centre at its Reading site in the UK. Figure 5 presents the various types of sources and their approximate numbers that play a part in the daily data collection for ECMWF.

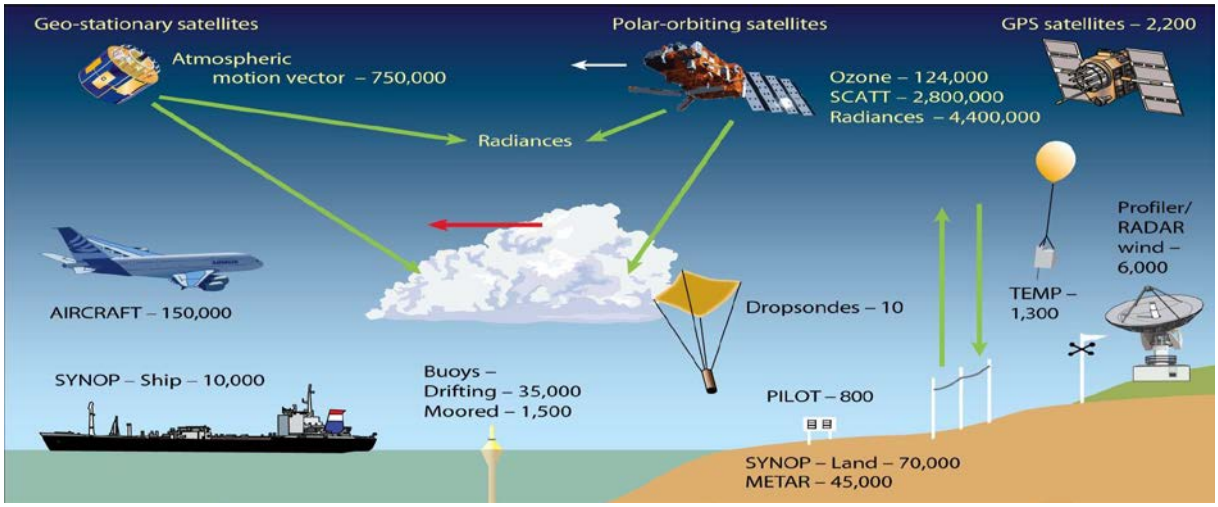

Figure 5: ECMWF daily collection of 149 million observations from heterogeneous sources

For operational resiliency, the centre operates two identical, yet separate, 3.6 PFlop/s Cray XC30 supercomputers. Each system comprises 3,504 nodes, each with two 12-core Intel Ivy Bridge CPUs and 64GB of main memory. Both systems were upgraded in the first half of 2016 replacing the Ivy Bridge processors with Broadwell CPUs (18-core), double the memory in each node to 128GB, and added a few additional XC40 nodes. The new systems yielded a peak performance of around 8 PFlop/s each.

On the infrastructure side, ECMWF is putting quite some effort in monitoring everything from power and cooling to water quality and data hall environment. Power monitoring covers power consumption at the system level (not individual nodes) as well as the quality of utility-provided power. Monitoring of the cooling infrastructure proved to be particularly tricky as most sensors available are analogue and additional accuracy loss may be introduced by analogue/digital converters (ADCs). All sensors (including power) are therefore checked regularly for accuracy and calibrated if needed. The water quality of the cooling loops is being monitored by multiple sensors: pH, conductivity, TDS (total dissolved solids), and molybdate concentration. The cooling loops also include a filter to remove bacteria, and chemical and bacterial analyses of water samples are regularly performed in-house.

5.3 CEA – Jean-Marc Ducos and Frédéric Souques

CEA operates two computing centres at its Bruyères-le-Châtel site: TERA for the Military Applications Division and TGCC that hosts publicly available HPC systems.

In summer 2015 TGCC installed the "INRA Cube" for the Institut National de la Recherche Agronomique, a cluster system in an APC hot-aisle containment. From an infrastructure point of view, the cluster is self-contained and only requires connection to electricity and a chilled water supply.

In April 2016, CEA deployed the CCRT-D system for the Centre de Calcul Recherche et Technologie, a 1.36 PFlop/s bullx machine. The system is available for industrial research partners and will replace the older CCRT-C system that will be decommissioned by the end of 2016.

In March 2017, CEA will deploy for GENCI phase 1 of the CURIE2 system, with a peak performance of up to 15 PFlop/s that will replace the current CURIE machine that is scheduled for decommissioning in late 2017. Phase 2 of CURIE2 will be rolled out early in 2018 and will add another 15 PFlop/s.

With the retirement of the older CCRT-C and CURIE systems and the commissioning of the successor systems, TGCC will gradually switch from air-cooling to direct-liquid cooling of most systems.

TERA is in the process of installing the TERA1000 system in multiple phases. The final TERA1000 system will consist of two different architectures: TERA1000-1 will be a standard x86 system with 2,7 PFlop/s and equipped with Intel Broadwell processors. TERA1000-2 will be based on Intel's Knight's Landing and provide a peak performance of 25 PFlop/s. The first phase of TERA1000-1 was installed in April 2015 with a peak performance of 148 TFlop/s. The second phase is currently being deployed and will add the remaining 2.55 PFlop/s. In parallel, a small part of TERA1000-2 is currently being deployed. The whole system will be in operation by the summer of 2017. The older TERA100 system will be gradually phased out, with the first half being shut down in the summer of 2016, and the second half in the summer of 2017.

As with TGCC, the TERA system will gradually phase out air-cooled system and switch to direct-liquid cooling for the bulk of the cooling. However, as the new TERA1000 system will be warm-water cooled, it will yield a lower PUE of approximately 1.1. TGCC, on the other hand, uses mostly cold-water cooling and yields a PUE of 1.2 to 1.3.

TGCC experimented with Ultra Capacitor Modules (UCM) to accommodate voltage drops in the power supply of their HPC systems. Voltage drops below 30% are typically not an issue and drops above that threshold can typically be tolerated by the power supplies if they are shorter than 125ms. However, longer lasting voltage drops may be critical and could be dealt with by UCMs. Yet, CEA found that these events are quite rare and hence may not justify the deployment of UCMs. UCMs may, however, be also beneficial in case of shorter voltage drops as they also help with reduced power supply performance in case of such events.

With the installation of the new Bullx systems, CEA observed new constraints in terms of load capacitance of false floors and concrete slabs which are typically around 1000kg/m² and which are exceeded by the new Atos/Bull systems and hence require reinforcement.

6 Session IV – Energy Efficiency and Monitoring

6.1 Keynote 3: GEO power management framework - Jonathan Eastep (Intel)

To reach the goal of building an exascale system within a 30MW power envelope by 2022, major improvements in terms of energy efficiency will be required. Process and architectural improvements as well as increased integration will help to improve energy efficiency, but an additional factor of approximately 3 will be required to reach the goal. Exploiting application-specific optimisation opportunities and hence using the available energy more wisely will be the key for this. As most of these optimisation opportunities are detectable only during runtime, Intel started the GEO project to develop an application aware runtime environment that is able to coordinate global energy management decisions dynamically and online. Application awareness is achieved through machine learning techniques which typically yield an 1.1 to 1.3 application speedup within the same power envelope. GEO – Global Energy Optimization - is an open-source platform; core development is however driven by Intel which also facilitates SW/HW co-design for performance and power management features in future Intel CPUs.

The GEO runtime support is implemented as a tree hierarchy of agents, each of which owns a sub-problem in their particular sub-tree. Each compute node runs its leaf agent on a dedicated core to allow for fast reaction times and deep analysis. Telemetry data is aggregated towards the root to minimize network traffic. All leaf agents use the same “Leaf Decider” to tune the energy policy within the compute node, whereas all non-leaf agents run the same “Tree Decider” to coordinate the energy policy across nodes. GEO currently only implements one energy optimization strategy which minimises then runtime of an application within a power cap. However, additional strategies can be implemented by means of plugins. Currently, Intel is seeking collaborators to develop a strategy that finds a good trade-off between low power consumption and low runtime to execute an application at its most efficient point.

The source code is available on Github [2].

6.2 EEHPCWG – Natalie Bates

The Energy Efficient HPC Working Group aims to promote and to accelerate energy efficient HPC by exploring innovative approaches, peer to peer exchange, sharing best practices and taking collective action. It currently has approximately 600 members from more than 20 countries.

One of the key tools for energy efficient HPC operations is the availability of a dashboard – an easy to read display that conveys actionable information. The information should be stakeholder specific, i.e., with individual views tailored for directors, facilities managers, system managers, and other stakeholders of HPC sites. Additionally, the Metrics team within the EEHPC WG promotes the use of ITUE (IT-Power Usage Effectiveness) and TUE (Total-Power Usage Effectiveness) as better alternatives to the widely used PUE as they also account for power usage efficiency within the IT equipment.

In the field of resource management and job scheduling, the approaches to limit the power consumption of HPC equipment seem to differ quite a bit at different sites: for US and Japanese sites, the scope is on power capping techniques to limit the total power draw of the IT equipment, i.e., to do as much compute as possible within a given power envelope. For European sites, the scope is more on optimizing the energy to solution, i.e. using a little energy as possible for any given compute task.

7 Session V – Vendors of future HPC processor architectures

7.1 ARM – Geraint North

The ARM business model is the licensing of instruction set architectures and micro architectural implementations, which partners use to create their own system-on-chip (SoC) products[2].

The ARM Development Solutions Group has 25 engineers in Manchester, UK; its mission is to enable the ARM software ecosystem for HPC and the enterprise: optimised math libraries for ARM-compatible microarchitectures, support for ARM-ported open source HPC software, performance tools and new compiler technology.

ARM® Cortex®-A series of processors is divided in 3 processor tiers: High Performance, High Efficiency, and Ultra High Efficiency. ARMv8-A Cortex-A57 and Cortex-A72, recent representatives of the High Performance 64-bit tier of the portfolio, can deliver HPC capabilities comparable to x86 architectures within lower power budgets.

Examples of server-class SoCs now available are Applied Micro X-Gene (8 cores), AMD Opteron A1100 (8 A57 cores), NVIDIA Tegra X1 (4 A57 cores + 4 A53 cores, and a Maxwell 256-core GPU). Next processors released (as of April 2016, date of the workshop) would be Cavium ThunderX (48 cores/chip); Applied Micro X-Gene 3 (32 cores); Qualcomm (24+ cores/chip); Broadcom Vulcan. NVIDIA Tegra X1 can benefit from power management and monitoring capabilities via Jetson TX1 Developer Kit, that can collect, expose and visualize events collected from different sources including ARM-based CPUs.

ARM and other players are putting specific efforts on the HPC software ecosystem development for 64-bit ARM targets, a crucial dimension and key success factor for the dissemination of such architectures, at different levels:

- Performance libraries: commercially supported math libraries such as BLAS, LAPACK, FFT are available for 64-bit ARMv8 platforms, based on NAG technology
- Open Source packages: ARM provides and supports optimisations for silicon vendors and set-up for many Linux distributions, for a variety of compilers, MPI implementations, maths libraries, data management libraries, profiling tools, prebuilt Python
- Third-party commercial software for 64-bit ARM include Allinea's DDT, MAP and Performance Reports, as well as PathScale EKOPath Fortran/C/C++ compiler, optimized BLAS and math libraries, OpenACC support.

7.2 IBM – Klaus Gottschalk

The HPC commitment of IBM is clearly reasserted, in the broader and evolving context of IT industry, where business and consumptions models are changing, and Moore's law slowdown and processor technology can no longer provide sufficient price/performance gains. Access and delivery modes are increasingly including scale-out datacentres and cloud infrastructures, with a redefined and evolving relationship between data and compute needs (a more data-centric vision), exacerbating data movement and storage as well as capacity elasticity issues.

[2] At the time of the workshop, the Japanese IT conglomerate Softbank had not yet announced that they were buying ARM Holdings, the UK-based microprocessor designer for about $32 billion. The announcement was made in July 2016.

This calls for innovative approaches combined at different levels: processor architecture enhancement, mixing CPU design together with acceleration and hybridization through proper interfaces and partnerships; filesystem scalability; workflow and cluster management.

OpenPOWER is the processor architecture for HPC and Big Data; IP can be licensed to enable semiconductor partners to build chips. Open interfaces allow tight integration with accelerators using CAPI and NVLink to deliver more Flops/W (NVIDIA, Xilinx, Altera), as well as storage (CAPI Flash) and networking (Mellanox). Thus, system and software open developments allow innovative POWER-based servers from partners.

The Open Power Foundation has more than 230 members (research organisations including application development, software and system development and integration companies, I/O-storage companies, board-chip-SOC designers and vendors, etc.). Important collaborations have been established with US DOE, MPCDF and FZJ in Germany, STFC in UK.

Current POWER 8 has 12 cores (8 threads per core so 96 threads in total), up to 1 TB per socket, up to 230 GB/s sustained memory bandwidth, and direct accelerator interconnect.

The OpenPOWER-based HPC roadmap is summarised in Table 2 (IBM vision with Mellanox and NVIDIA):

	2015	**2016**	**2017**
Interconnect	FDR IB PCIe Gen3	EDR IB CAPI over PCIe Gen3	Next-Gen IB Enhanced CAPI over PCIe Gen4
GPU	Kepler PCIe Gen3	Pascal NVLink	Volta Enhanced NVLink
CPU	POWER 8 CAPI interface	POWER 8+ NVLink	POWER 9 Enhanced CAPI and NVLink

Table 2: OpenPower-based HPC roadmap

7.3 Intel – Andrey Semin

Intel Scalable System Framework approach is meant to address holistically all HPC growing challenges and needs. Challenges encompass all the commonly understood system bottlenecks and "walls" (memory, I/O, energy…), the diversity and even diverging infrastructure trends (for numerical computation, machine learning, big data processing and visualization), and the needs for flexible and elastic access and delivery at every scale.

The corresponding technological answers encompass 4 axes:

- Compute with Intel Xeon processors (multi-core), and Xeon Phi as processor or co-processor (many-core);
- Storage/memory with resp. Lustre solutions (support for various HPC features), and SSD/Optane based on 3D XPoint technology (in an effort to bring memory closer to compute);
- Fabric with OmniPath architecture, True Scale fabric and Silicon Photonics;
- Software with a portfolio of tools, libraries and other stack elements.

Latest Xeon processor E5-2600 v4 aka Broadwell-EP (14 nm) accommodates up to 22 cores and 44 threads. Latest Xeon Phi x200 family (codename Knights Landing) is x86 binary-compatible, up to 3+ TFlop/s, with a 2D mesh architecture and up to 72 out-of-order cores,

with up to 16 GB of on-package memory – able to act either as memory (flat mode), cache, or in hybrid mode mixing cache and flat.

The3D XPoint memory standard can work as both DRAM and Flash. 3D XPoint based Optane SSDs and DIMMs are promising important improvement in density, latency and endurance, and opening perspectives of more pervasive and efficient SSD usage (instead of disk storage) together with larger and faster local memories closer to compute.

7.4 NVIDIA – Axel Koehler

Tesla P100 is NVIDIA's latest GPU architecture of the Tesla series. It delivers 5.3 TFlop/s in double precision performance, NVLink interconnect, 16 GB of HMB2 memory with 720 GB/s of bandwidth and the page migration engine to handle data transfers automatically. It is designed to address a broad range of compute-intensive usage from HPC to deep learning. This latter area is a strong emerging application driver. The NVIDIA DGX-1 solution is designed for deep learning, composed of 8 Tesla P100 and an optimized software stack including cloud management.

The Tesla GPUs come with a suite of tools for easier monitoring and management of power: Data Centre GPU Manager (DCGM) can implement system policies, monitor GPU health, diagnose system events, etc. It can operate either in standalone (daemon) or embedded (client process) mode, and is able to manage groups of GPUs.

Enhanced power and clock management features comprise:

- Dynamic power capping applied to a single GPU or to a group to better drive power density
- Synchronous clock boost to dynamically modulate multi-GPU clocks in unison across multiple GPU boards, based on the target workload, power budget and other criteria, so as to deliver predictable performance
- Fixed clocks for fixed performance.

A number of job statistics can be collected (whether GPUs are used, and how many of them, error and warnings including ECC errors, GPU health data, etc.). DCGM is meant to cache 1 to 4 hours of data. There are plans to release plugins to push data to various time series database tools or contribute to open source plug-ins for such tools or metric publishing tools.

8 Session VI – Site updates Europe (Part 2)

8.1 Datacentre monitoring, a survey of PRACE sites – Norbert Meyer

In the first quarter of 2016 a survey, that inquired into the solutions currently employed by these sites for monitoring their datacentre infrastructure, was distributed among PRACE Tier-0 and Tier-1 datacentres. PSNC conducted the survey and is compiling a report for PRACE. All information reported at the workshop is based on the responses received from large data centres (PRACE hosting partners) and large infrastructures of the PRACE general partners.

The Survey revealed that HPC Data Centres have plenty of different monitoring platforms to choose from. For example:

- Tridium JACE system – Niagara
- Siemens Desigo
- Graphite/Graphana
- Pronix POC (Power Operation Centre)

- Check_MK
- Messdas
- InTouch
- UGM 2040
- Desigo Inside
- Johnson Controls Metasys
- Siemens WinCC
- SCADA PANORAMA
- Victor Web LT
- StruxuWare Power Monitoring Expert Data Center Edition.

This large selection of options does however not mean that each of these monitoring platforms by themselves provide all required services. In fact, each platform is responsible for only a limited part of the monitoring of datacentre infrastructures and functionalities. The most common solution encountered is the use of three or more independent software platforms for monitoring various types of equipment. For example, one platform is responsible for heat exchangers (HxB), Air Conditioning (Chillers and CRACs) and generators; another for web cam, TV, fire protection, door access codes and other security features installed in the building; and a third one for live status, event logging, triggering of alarms, and generating problem tickets.

It is practically impossible to take care of data centre monitoring utilising just one of the above mentioned monitoring platforms. However, one centre developed a system, completely from scratch, that allows integrating all monitoring into a single global system. The development of this solution implied searching for complementary solutions and creating the individual networks consisting of various pieces of software adapted to the specific needs of the datacentre and the specific equipment used there. But even this system had to be fairly eclectic and did not result in a common interface.

The lack of a global, structured, “full-stack”, monitoring view is reported as the biggest problem by most survey respondents. Without a well-structured dashboard view it is difficult to have full control over the datacentre system and optimize it using information that is related and acquired from different platforms.

Ideally, one well prepared platform would meet such relevant features like monitoring water temperatures of all cooling loops, energy optimization functionality, tight monitoring and early detection of problems, view of historical data, historical trend monitoring and easy programming and log system. Such a platform should support the already implemented features, among others, asset and capacity management, workflow, Application Programming Interface (API), mobile & web-based interface, real-time monitoring, trends & reporting, data centre floor visualization and simulations.

Standardization on a common API, an open interface for monitoring, provided by all suppliers is currently clearly lacking, as are options for easier configuration of third party access to collected data, to graphs, and to the triggering of alarms. Such access should be customisable per user. Many respondents see a need to relate data centre monitoring data to data of the HPC resource management system. A connection between the two could lead to integrating common IT monitoring software and datacentre monitoring tools.

The current multitude of software necessary for relatively precise control over the datacentre not only affects the amount of collected information, but also the number of people (FTE, Full Time Equivalent) dedicated to maintaining and managing the datacentre hardware. There is a correlation between complexity of the complementary software used in a datacentre, its size

and the number of employees. The more complex the system is, the more people are needed, although according to studies an average of 2 FTE staff is enough to support the DC infrastructure and monitoring. It should be emphasized that FTE excluded functions such as managing computing, data infrastructures and services correlated with these infrastructures.

The complexity of monitoring systems is shown by a wide range of equipment connected to them. This group includes: electrical distributions, cooling systems (including data from chillers, coolers, pumps, valves, heat pump and other technology), UPS, high voltage system, low voltage system, dryers and air conditioning systems for IT technology, flood detectors, smoke and heat detectors, fire prevention system, air quality, temperature and moisture, oxygen and carbon dioxide levels, fuel storage and management, basic security (video, access control, door status indication) and more.

An **inventory software platform** plays a very important role in data centres. The HPC centres use their own implementations, where an inventory is handled via mysql databases or an inventory data catalogue is based on an Oracle database. Also, applications such as the web based Python/Django, CommScope iTRACS are being deployed. Open source platforms like Ralph (based on the Django framework) or Argos, are used as well.

Monitoring data are usually collected via a **separate internal LAN network** dedicated for that purpose. The monitoring network is protected against external threads and unauthorised access. Usually the network is not physically connected to any other network and there is no routing of the network to the outside world. At the moment, there are two common types of access to the system monitoring data:

- From a machine on the dedicated (internal) LAN;
- Through a VPN from outside, connecting to a "stepping stone" that has an interface on the monitoring LAN.

As the monitoring systems become more complicated, they require additional servers and storage to capture and analyse data. To support the monitoring platform most DCs use additional IT systems: ranging from one to four servers. The reason for deploying several servers is generally not performance related. Most centres insist on high reliability and high availability. For that, at least a HA-cluster consisting of two machines is needed.

8.2 CSCS – Ladina Gilly

The new datacentre (DC) at CSCS (Lugano, Switzerland) began operation at the end of March 2012. Now, after four years of use, three achievements in the operation of the centre stand out as particularly successful:

- Continuous fine tuning of the operation has led to the PUE being well below the design specification of 1.25, despite the low load. The usual PUE value for a traditional DC is between 1.6 and 1.5, for the same climate regions as Lugano.
- The addition of micro turbines in the lake-water cooling circuit allow CSCS to generate power from the free-falling water returning to the lake. This allows the recovery of one third of the power used to pump the lake water to CSCS.
- By tendering for framework contracts for a planning team and key contractors, CSCS is able to execute new projects, for instance, for the installation of new supercomputers very efficiently. The framework contracts also ensure price stability and continuous development of knowledge and experience over a four-year period.

Successes are great and must be mentioned, but often more can be learned from unexpected problems, from the way in which they were solved, and also from the approaches – both successful and not - that were taken to tackle them. The presentation by CSCS focused on

some challenging issues that have arisen, mainly in mechanical parts of the infrastructure. The examples of reduced flow rates over heat exchangers (HE), stray currents in ball bearings of pumps, and the water quality in closed cooling circuits were discussed.

Flow rates over HE

During the first 18 months of operation reduced flow rates over HE were detected. For some time the presence of organic deposits in the filters led involved parties to suspect that the problem was caused by iron bacteria that had suddenly appeared in Lake Lugano, form where the data centre draws its water for cooling. Based on this hypothesis numerous checks and tests were conducted, such as water quality, HE performance, pump performance. Ultimately, the source of the problem was found to be unexpectedly high pressure drops in the piping. These were being caused by an unfortunate conjunction of multiple welded joints, diameter changes, and tight bends in polyethylene (PE) pipework on the primary side of the HE. As the pipework on the secondary side of the HE was made of stainless steel the pressure drop problem only appears on the primary side.

PE piping requires joints with either sleeves or welding. Due to space limitations it was not always possible to use sleeves. Welded joints have the downside of containing “welding lips” that protrude inside the pipe and can cause resistance. In this specific case the issue is compounded by the numerous changes in diameter and tight bends. The problem can be solved by replacing the PE piping with stainless steel piping and optimising the geometry of the pipework to avoid tight bends. This change was planned to be implemented by mid-September 2016.

Stray currents in ball bearings of pumps

The ball bearings of the pumps on the cooling loops were wearing out significantly faster than they should and showed signs of damage from stray currents. An initial attempt to solve the issue by using ceramic coated ball bearings proved to be insufficient. A series of measurements and audits revealed an imperfect execution of electromagnetic compatibility (EMC) wiring. The issue can be resolved by:

- Completing EMC installation
- Adding sinus filters between the variable frequency drive (VFD) and pump motors
- Tuning the VFD frequency to match that of the sinus filter

The goal was to fix this issue by September 2016.

Water quality in closed loops

During operation, tests showed that the water quality in the secondary cooling circuits was gradually degrading because of oxidation processes. The planning teams had assumed that water treatment was unnecessary at the given operating temperatures. Continuous testing of the water quality proved this assumption to be incorrect. The solution, implemented in 2015, was the addition of water treatment cartridges to all cooling loops to correct, stabilise, and then maintain the water quality at the desirable level.

Other projects

On the electrical side of the infrastructure, CSCS has decided that it will add a generator-powered electrical distribution network to facilitate the maintenance of the main power breakers without causing an interruption to the UPS-supplied HW.

The UPS batteries are monitored continuously to gain further insight into battery deterioration and life expectancy.

CSCS is determined to resolve any remaining construction issues by March 2017 at the latest, which is when their five-year warranty period ends.

8.3 LRZ – Herbert Huber

HPC datacentres like LRZ generate a huge amount of heat, i.e. waste energy, which in principle can be re-used. However in practice, it may be hard to find a partner that can effectively use the total amount of waste heat all year long. New houses and buildings in Germany have very good thermal isolation and heating requires about 50W/m^2 or less. SuperMUC's waste heat would suffice to heat at least 40,000 m^2 of office space! So what to do, especially in summer? There is no nearby SPA centre or market garden that could use the waste heat. It would be a very practical solution if the waste heat can be profitably re-used at the datacentre itself.

LRZ has started doing just that, by exploring adsorption cooling which re-uses the waste heat for chilled water production. As the datacentre hosts systems other systems besides SuperMUC, that are not direct liquid-cooled, the chilled water can productively be used by CRACs and/or water-cooled rear doors.

There is a continuous process of two alternating phases - adsorption and desorption (or regeneration) - going on in the adsorption chiller:

- Adsorption – the accretion of water vapour on the adsorbent's surface.
  Warm water, located in the condenser, is fed into the evaporator. Heat is extracted from the water through the evaporation process and cold water for the cold distribution circuit is produced. Water vapour, formed in the evaporation process, is taken up by the adsorbent, dry silica gel. Once the gel is saturated, the regeneration starts.
- Desorption – Drying the adsorbent (regeneration). The water, heated by the HPC system is channelled over a heat exchanger through the adsorption chamber in the chiller aggregate. The added heat dries the silica gel, resulting in a discharge of water vapour which flows into the condenser where it liquefies again. Once the silica gel is sufficiently dry the addition of heat in the adsorber stops.

The concept of an adsorption chiller is presented in Figure 6 below.

A new system – CooLMUC-2 - is operational at LRZ. The R&D work was done by LRZ in co-operation with Sortech (http://www.sortech.de), a German SME company known for developing and manufacturing innovative, energy efficient solutions for the generation and storage of cold and heat – anywhere where refrigeration is needed and excess heat is available.

CooLMUC-2 is prototypical and re-uses the waste heat generated by 8 racks of Lenovo NeXtScale systems: 6 compute racks with 384 nodes in total, 1 InfiniBand rack, and 1 management rack. The theoretical peak performance of this compute system, which is operated with water inlet temperatures between 30ºC and 50ºC, is 466 TFlop/s. It showed an actual compute performance of 366 TFlop/s during a run of the HPL benchmark (78.54% efficiency). Its maximum power consumption is 149.5 kW. The maximum output into the warm water cooling loop is 104 kW of waste heat. The average output is about 89 kW and the adsorption chillers are able to generate a chilled water volume from this with an average cooling capacity of 48 kW.

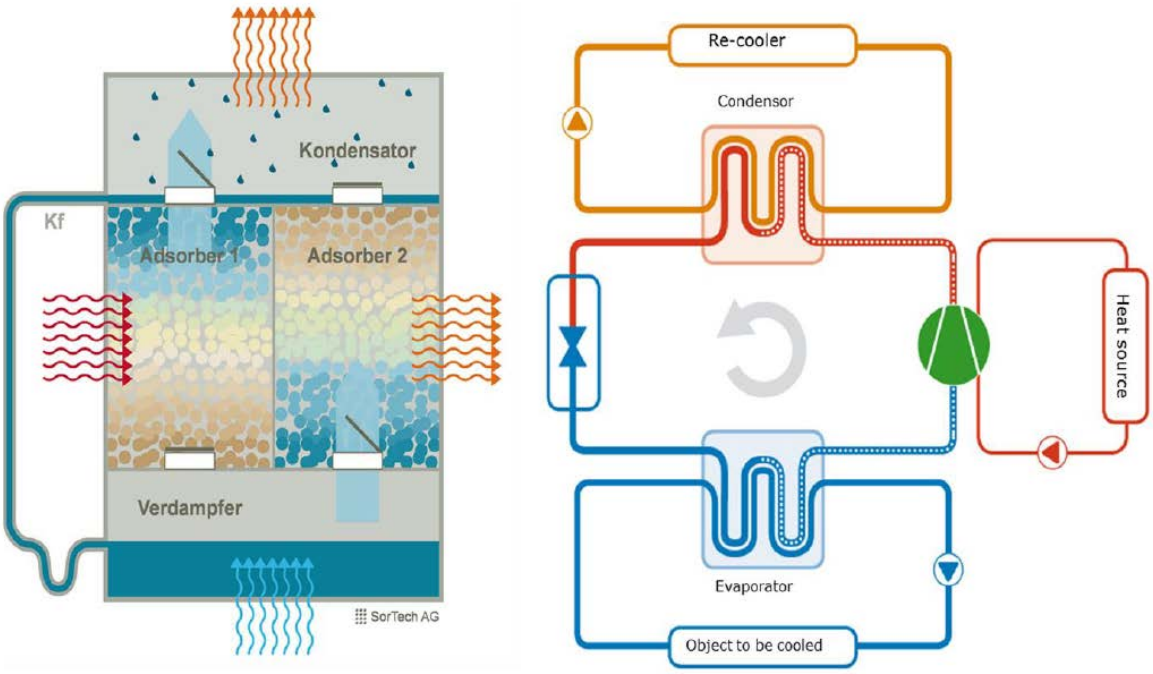

Figure 6: The concept of an adsorption chiller

The adsorption coolers at LRZ contain silica get units. Sortech is developing a new technology that uses zeolite-coated metal fiber heat exchangers. A new prototype, based on a type of a zeolite with low driving temperatures, and therefore more suitable for datacentre cooling, is under development now. It is expected to be more than twice as efficient as the silica gel based coolers and will be installed and tested at LRZ at the end of 2016.

9 Session VI – System integrators

9.1 Atos/Bull - Jean-Philippe Sibers

Sequana is a new system integration design by Bull-Atos, that is targeting exascale computing. It is aimed primarily at the high end market of HPC and other demanding applications and offers “ultra-dense and scalable” integration of a wide range of technologies. The interconnect can be InfiniBand EDR from Mellanox, and, from 2017 onwards, BXI from Atos. Processors can be from the Intel Xeon and Xeon PHI families, from ARM. NVidia Pascal GPU accelerators can also be accommodated. A Sequana “Cell” comprises 2 compute cabinets (up to 288 compute nodes) and one management and interconnect cabinet. With the options available in 2016 a single cell can provide up to 2 PFlop/s.

The design particularly focusses on cooling improvements, power efficiency, and easier integration in the datacentre infrastructure. Sequana is 100% direct liquid-cooled. The cell has an embedded heat exchanger which accepts warm water inlet up to 40ºC. The local flow rate of warm water is controlled to maximize outlet temperature independently of the system’s compute load.

Energy management relies on high frequency sampling of on-board components in each compute node. HDEEM (High Definition Energy Efficiency Measurement) is developed by Atos in a collaboration with TU Dresden.

Future developments target to support extreme CPUs, up to 400W with warm water cooling. Atos also wants to improve the insulation of cabinets, to reduce uncontrolled ambient air

cooling and to increase inlet and outlet water temperature further. If the outlet temperature is higher and highly controlled, opportunities for efficient heat reuse are increased. Atos foresees that wasted heat will be reused, via adsorption cooling, to feed air-cooled cabinets and water-cooled rear doors of cabinets with 15ºC water.

9.2 Cray – Wilfried Oed

Cray Engineer Wilfried Oed pointed out that one important function of the many sensors in the new Cray HPC systems is to make each cabinet very aware of the load that it is running and how that load may be rapidly fluctuating. The cooling sub-system of each cabinet in a Cray XC series system is highly "load aware" and is regulated by adjusted flow rate. The system can be described as an indirect water-cooled system. Within the cabinet components are air-cooled, by means of fans. The air inside the cabinet is water cooled such that nearly all heat is dissipated to water. The cooling capacity of the water is adjusted to the load by valves that regulate the flow rate.

The CRAY XC product line implements HSS – Hardware Supervisor System – which relies on dedicated hardware and software. Data is collected at the node, blade, and cabinet level, for each component (CPU, node, memory, fans, blowers), a specific daughter card is developed for KNL.

Cray points out that more than just system regulation can be achieved with the data points collected by HSS. All data collected can be retrieved and processed by RUR – Resource Utilization Reporting. RUR is a scalable accounting tool for gathering diverse usage data and reporting to system administrators and users. The system is designed to relate the collected data to job execution, thus, to reveal the details of power consumption, cooling requirements, and component behaviour at the job level. RUR collects data in two phases: before and after the application run. The application specific staged data is the delta of these two. Post processing can provide usage statistics and can be used for each application for time to solution and energy to solution optimization by clock frequency adjustment.

9.3 HPE - Sammy Zimmerman

HPE reports that during the last year they have regularly confronted their customers and potential customers with a "Chiller Chat" – a small informal and slightly provocative questionnaire in which they ask datacentre personnel questions like: "Why are you running chillers in your datacentre at all?"

The answers they get in response to such questions are often not very valid nor very convincing. The two most frequently heard responses are:

- Most often - "Because I have them anyhow";
- Sometimes - "Because I need to – the water inlet for my system is between 7ºC and 10ºC".

Indeed, depending on climate, chillers may be needed. But on closer inspection, it turns out that chillers are usually still run 24/7 while they are actually only required for a very small fraction of the time – in many cases only a few hot days per year. In recent years the cooling requirements of smart HPC systems have evolved dramatically and daily practices and policies at many datacentres yet have to take those changes into account.

Centres could gain a lot from liquid cooling planning. The basis for such a plan is a psychrometric chart, i.e. a chart with local climate specific data relevant to the possibility of

free cooling, such as representative average “wet bulb” and “dry bulb” temperature measurements for each month of the year.

Although not everything in the ASHRAE standards and guidelines is equally well applicable to HPC, they constitute a very good place to start for reviewing enterprise datacentre and cooling solutions and planning. The outcome of such planning will be very location specific; e.g., Hyderabad, London, and Stockholm do not have the same climate, therefore, the optimal use of dry cooling, evaporators and chiller differ significantly. The energy savings resulting from the right decision can be significant. The recommended trend is to go to the high ASHRAE water temperature class.

9.4 Lenovo - Luigi Brochard

To deliver the most cost effective HPC solution for data centres, Lenovo considers all cooling technologies: air cooling, air-cooled with rear door heat exchangers, direct water cooling. Direct liquid cooling should be used for highest energy efficiency. Hot water cooling and high heat to water ratio, make free cooling and energy re-use more feasible. Large scale usage of such systems may result in a PUE for the data centre of 1.1 or less and an ERE much less than 1. In terms of TCO, it is cost-beneficial in less than a year if the data centre is new, regardless of the electricity costs. However, not every data centre is new and for existing data centres return on investment will depend on the electricity costs. At the other end of the spectrum, air-cooled solutions cannot cope with a dense footprint and will keep the PUE of a datacentre around 1.5. However, provided the space is available, they can be accommodated in any datacentre and offer maximum flexibility and the broadest choice of configurable options, in particular when associated with rear door heat exchangers.

Lenovo has performed return on investment studies that take the details and context of datacentres into account: their geographic locations, the weather conditions and the local energy cost. In existing air-cooled data centres the payback period of direct liquid cooling strongly depends on the electricity price. If the kWh price is less than $0.12, systems with a lifetime of four years or less will never reach the break-even point.

With NeXtScale systems, Lenovo can deliver a wide range of solutions. For Lenovo NeXtScale solutions two levels of total cost of ownership (TCO) analysis are offered:

1. High level analysis, based on node configuration, assumed PUE for an existing data centre, and the customer’s energy pricing. The output is an assessment of which cooling method is optimal for the customer’s datacentre context.
2. Detail analysis, taking more details of the node and data centre configuration into account, as well as the target workload to determine power. The PUE is determined based on location and weather, and the customer’s energy pricing is taken into account. The outcome is a more detailed TCO assessment of chosen cooling methods and predicted PUE.

Lenovo NeXtScale systems have global power management and monitoring. xCAT enables fine-grained monitoring of component temperatures and power usage. Platform LSF, an energy aware batch scheduler and Integrated Management Modules (node-level system management) are coupled to a centralized database for application workload energy optimization.

A collaboration focusing on EAS (Energy Aware Scheduling) was initiated between IBM and LRZ in 2012. This work has been integrated in LoadLeveler features and is now available in LSF 9.1.2 release integrated in Lenovo NeXtScale product line. This includes the reduction

of power consumption of idle nodes, power optimisation of workloads, energy accounting, and options for energy policy based scheduling. Two policies are supported:

1. Minimise Time to Solution: The goal is to run an application to a higher frequency if performance scales with frequency according to a defined threshold.
2. Minimise Energy to Solution: The goal is to save energy within a maximum performance degradation threshold.

Lenovo is developing new features to deliver more energy efficient solutions starting in 2017.

10 LRZ datacentre tour

The programme included a short tour in the LRZ datacentre, which, among other systems, hosts the PRACE Tier-0 system SuperMUC. The datacentre now consists of two cube-like buildings. The first cube was built in 2006, the second one in 2011 (see Figure 7). LRZ has 3160.5 m^2 of IT equipment floor space and 6393.5 m^2 of data centre infrastructure floor space. The data centre has a power supply of 2 x 10 MW (20 kV). The data centre is powered entirely by renewable energy. The average PUE of the centre is around 1.22.

Figure 7: The LRZ double cube building

The current petascale flagship machine at LRZ is the SuperMUC system, which was constructed in two phases. Both phases use direct warm-water cooled system technology. They have a common programming environment and share a 10 PB and a 5PB GPFS file systems.
Phase 1 (see Figure 8) consists of IBM System x iDataPlex racks:

- 3.2 PFlop/s peak performance
- 9216 IBM iDataPlex nodes in 18 compute node islands
- 2 Intel Xeon E5-2680 processors and 32 GB of memory per compute node
- 147,456 compute cores
- Network IB FDR10 (fat tree)

Figure 8: SuperMUC phase 1

Phase 2 (see Figure 9) is based on Lenovo NeXtScale WCT:

- 3.6 PFlop/s peak performance
- 3072 Lenovo NeXtScale nx360M5 WCT nodes in 6 compute node islands
- 2 Intel Xeon E5-2697v3 processors and 64 GB of memory per compute node
- 86,016 compute cores
- Network IB FDR14 (fat tree)

Figure 9: SuperMUC phase 2

On the floor we also encountered the CooLMUC-2 system reported on in the LRZ presentation (this document, section 8.3, p. 21). The white cabinets, shown below, in the pictures of Figure 10, are the adsorption chillers using the waste heat from a 466 TFlop/s Lenovo system to produce chilled water.

Figure 10: Adsorption chillers of CoolMUC-2

11 PRACE 4IP/WP5 Session

11.1 CINECA (Italy) – Carlo Cavazzoni

CINECA has a 2014-2020 plan for acquiring supercomputers and data storage equipment with a target of 50 PFlops/50 PB in 2019-2020.

The current phase consists in installing in 2016/2017 a new supercomputer (Marconi) provided by LENOVO. This supercomputer includes an Intel Xeon partition (2.1 PFlop/s in 2016, 4.5 PFlop/s in 2017) and a KNL partition (11 PFlop/s in 2016), with an Intel Omni Path interconnect (Fat Tree). The power usage is expected to be around 2.5 MW (running LINPACK), 2.2 MW (regular workload). A part of this machine is dedicated to the EUROfusion workloads.

The cooling, considering a computer room at 30°C, will be done by free-cooling during 4 months a year, and during 4 months a year with chillers. For the remaining 4 months, the type of cooling will depend on the temperature. In order to reach a target PUE of 1.2-1.3, coordinated cooling strategy (See [3]) will be implemented in order to optimize air and water cooling taking into account the thermal constraints.

11.2 FZJ (Germany) – Willi Homberg

FZJ (Forschungszentrum Juelich) is a research centre located in Juelich, Germany. JSC (Juelich Supercomputing Centre), as part of FZJ, is in charge of supercomputer operation, application support and R&D work. The total staff of FZJ is 5800 employees whereof about 200 belong to JSC. JSC operates two main computing rooms, one in building 16.3 (old building– total available power 2 MW), one in building 16.4 (new building – total available power 8 MW).

In terms of supercomputer systems (in building 16.4), FZJ is following a dual track approach: general purpose clusters on the one hand (currently Jureca, 1.8 PFlop/s, to be enlarged by a booster system within next year), highly scalable systems on the other hand (currently Juqueen, 5.9 PFlop/s soon to be replaced by a 50 PFlop/s system).

In terms of R&D, JSC is involved in several projects, and hosts early systems related to these projects (in building 16.3), including DEEP (EU funded exascale research project), Human Brain Project (HBP) and Quantum Chromodynamics Parallel Computing on the Cell (QPACE).

Regarding infrastructure, the cooling is provided in both rooms by a mix of water cooled rear doors, direct water cooling and air cooling (CRAC).

Datacentre monitoring includes the following functions:

- Energy management, IT equipment and infrastructure equipment, based on a Messdas software;
- Building management: leakage detection, data centre access supervision, safety monitor, fire alarm;
- IT monitoring and management of tickets;
- System monitoring with the LLVIEW software.

Several updates are scheduled for 2016 in the two computer rooms including several pilot systems (HBP, Huawei, Seagate), extensions (Deep-ER and Just) and exchange (QPACE3).

11.3 GRNET (Greece) – Antonios Zissimos

GRNET (The Greek Research and Technology Network) manages four new data centres in Greece. The total capacity is 84 racks and 1 MW in terms of IT power capacity.

Regarding HPC resource (hosted in the MINEDU data centre located in the main building of the Ministry of Education), the new Tier-1 supercomputer (ARIS) is now in full production. This system, provided by IBM, consists of 426 dual socket IBM NextScale nodes, each configured with Intel E5-2680 v2 10-core processors and 64GB main memory. The interconnect topology is a full non-blocking fat tree based on IB-FDR technology. Storage capability is provided by an IBM ElasticStorage (GPFS) appliance offering 1PB of raw storage space.

This supercomputer is air cooled with hot aisle containment provided by APC. Four chillers (4x117 kW), with free cooling are providing the cooling. The system is entirely protected by UPS (1.5 MW total capacity) because of the power fluctuation of electricity in the area.

This system was ranked #468 in the Top500 list of June 2015 with a LINPACK performance of 170 TFlop/s. Since then, an optimized run was able to reach 180 TFlop/s. The results in term of application performance and scalability are excellent. Access policy is close to the PRACE access policy in terms of principles and organization (2 calls per year + preparatory access). The system usage is already very high.

An extension to the HPC resource composed of fat nodes, GPU nodes, Phi nodes (provided by Dell) and additional storage, is currently in the commissioning phase. It will bring an additional performance of 254 TFlop/s peak.

The Louros data centre is located in the north of Greece. With this data centre, GRNET experiments the concept of green data centre with a project installed close to a dam (100% renewable energy) and a river. This data centre, mostly used for disaster recovery for cloud services, is targeting a PUE of 1.2 at full load by using cold water from the nearby river for cooling (maximum river temperature is 15.5°C) with chiller for back-up in case of a problem. This data centre is built with containers, for IT equipment (hot aisle containment) and also for infrastructure equipment.

11.4 IT4Innovations (Czech Republic) – Branislav Jansik

Located in Ostrava, the mission of IT4Innovations, is

- To deliver scientifically excellent and industry relevant research in the fields of high performance computing and embedded systems;
- To provide state-of-the-art technology and expertise in high performance computing and embedded systems and make it available to Czech and international research teams from academia and industry.

The current IT4Innovations supercomputer is Salomon (#47 in Top500/November 2015) with a peak performance of 2 PFlop/s (1.5 PFlop/s RMAX). This machine includes 1008 compute nodes with 24 Haswell cores per node and 864 Intel Xeon Phi 7120P. The cooling is provided by a mix of direct liquid cooling (for Haswell nodes – input 32°C/output 36°C) and of rear door heat exchangers (for Xeon Phi nodes).

Salomon is installed in the new datacentre (the previous system was installed in a container). The main specifications are the following.

- The power supply is fully redundant. It is based on two 22 kV independent lines. A diesel generator of 2.5 MW, coupled with a rotary UPS, can provide the power (at 22 kV) needed for the supercomputer in case of a power failure. One motivation for having generators is the fact that flood can occur in the area and disrupt power distribution, however, power quality in general is very good;
- The cooling combines warm water for the compute racks and cold water cooling. The warm water is provided by dry coolers on the roof, the cold water by compressors installed also on the roof. Heat pumps make the heat reuse for heating the building possible;
- The distribution of power to the compute racks is made under the raised floor (80 cm) by power bus on which power distribution boxes are attached where needed;
- Fire suppression is made via oxygen reduction (around 15%) using redundant N2 generators.

After one year of operation the experience with the new datacentre is very positive. The coupling of large rotating wheels (20 seconds backup) with diesel generator works well. Several minor problems had to be solved including:

- Flying wheels rotating at 2800 rpm are very noisy. The noise disturbed surrounding area at the beginning. This was solved by adding dampers;
- Flying wheels bearings can overheat, this is possibly connected to 2 months long stop-start cycles. Bearings were damaged on rotating elements,which can be due to either overheating or electromagnetic compatibility issues;
- Oxyreduct systems were starting to work more often than expected. This was due to the fact that the building was underpressured. The problem was fixed by inspection and replacement of air-conditioning particle filters on the building side;
- Heating the building by reusing heat from the computer works well, however, when switching to the district heating system, the water in the pipes coming from the district heating system is cold and cools the building at the start. The problem was fixed by allowing a very small inflow from the district heating system.

11.5 SURFsara (The Netherlands) – Huub Stoffers

SURFsara (previously SARA, before the merger with SURF) has long been its own provider of HPC datacentre services - until the split of SARA into SARA and VANCIS, where VANCIS became the company commercially providing, among other things, datacentre services to SARA and other parties. The current datacentre is fairly old and has a poor PUE (>1.5) by modern standards. It is also used to the maximum, making growth and transitions to new systems very difficult. In this situation, SURFsara after investigating different solutions, decided to rent space in a brand new commercial datacentre located close to the SURFsara offices.

In this datacentre, designed for a PUE of 1.22, SURFsara is renting 800 m^2 of private datacentre space on two floors plus a separate tape robot room on the third floor. The power committed is 1.5 MW (extensible to 1.8 MW) with UPS and generators. The cooling capacity is 1 MW of water-cooling.

Since the cost of operation in the new datacentre is lower and the entire floor surface is rented from the start, there is a compelling business case for moving all systems located in the old

datacentre, including the supercomputer, to the new datacentre as soon as possible. This move is planned during the June-November 2016 timeframe and is organized in several waves. The move of most systems is a "lift and shift" scenario with a short period of no service during the move.

The extension to Cartesius (196 Sequana cells for an additional power of 260 TFlop/s) will be installed in Q4-2016 in the new datacentre. Some of the new management and network equipment to integrate the Sequana island into Cartesius is first used to facilitate the move and to shorten the period in which there is no supercomputer service: a temporary "bridgehead" can be built in the new datacentre, to test components that have been moved, while the core of Cartesius is still running in production in the old datacentre, albeit at a reduced capacity.

Such a large move involves taking care of plenty of "details", including:

- Data loss mitigation measures;
- Vendor and specialised moving companies;
- Insurance;
- Run books;
- Freeze period.

The move from a VANCIS datacentre to a private commercial datacentre motivated the set-up of a datacentre office (DCO) at SURFsara. Previously SURFsara had no DCO which saves some money but causes serious problems due to the lack of focus on datacentre details in terms of planning, calculation and mistakes.

This newly created DCO took care of organizing the move in order to reduce the downtime for the users and will control the usage of the capacity rented by SURFsara. In addition, the DCO will be in charge of capacity planning, financial overview and inspections and quality insurance checks.

The new information system used by the DCO for managing datacentre information (layout of IT equipment, cables, power usage, heat dissipation, etc.) is based on "patch manager" (from patchmanager.com). This software provides a graphical interface for easy access to data.

12 Main findings

The presentations given during the 7th European Workshop on HPC Centre Infrastructures:

- Reveal important trends in terms of energy efficient HPC centre infrastructures and, more generally, energy efficient HPC datacentres.
- Give hints to assess the situation in Europe in this domain.

12.1 Trends

The evolution of HPC technology is the driving force that needs to be, as much as possible, understood and anticipated since facilities are multi-decade investments compared to systems that have a typical lifetime of 5 years.

The HPC roadmaps now target the delivery of production capable exascale compute facilities in 2023 – 2025. Notwithstanding this delay, there is a sustained trend towards more powerful systems. Vendors of HPC processor architectures, system integrators, and datacentre site representatives all alike foresee direct liquid cooling (DLC) with higher water outlet temperatures to be omnipresent in new densely packed HPC systems. Some of the newly built datacentres to accommodate HPC facilities are designed without a raised floor. However, in

older centres and in datacentres catering to other systems besides HPC, the return on investment of direct liquid cooling for less dense systems – e.g. storage cabinets or non HPC servers – will vary greatly across datacentres, depending on the age of the centre and the local electricity price.

Higher outlet temperatures of DLC cabinets increase the options for productive heat reuse. Some datacentres have been able to make profitable arrangements with external parties that have a demand for their waste heat, thus reducing the total cost of ownership for the datacentre. Others have found a suitable application, such as heating the datacentre office during part of the year, but have more difficulty finding enough local partners that can productively re-use the heat all year round. At least one system integrator (Atos) and one site (LRZ) have pointed to adsorption cooling technology as a productive way to re-use waste heat of HPC systems within the datacentre itself. Other cabinets are cooled by CRACs or by water-cooled rear doors that require water at about 15ºC.

The condition that exascale systems stay within a 30 MW power envelope, however, is not very likely attainable by means of energy efficiency improvements in computing hardware and datacentre infrastructure alone. A substantial part of the improvement has to come from users and the way in which they run their applications. Applications and/or users must become more energy aware and time to solution approaches to application optimization should be integrated with - or perhaps even replaced by - energy to solution approaches.

Probably the clearest exponent of this trend at the workshop was Intel's GEO project, which provides a runtime environment that analyses application behaviour through machine learning techniques and aims specifically at application speedup within the same power envelope. The trend focuses more on energy efficient application runs, which is also reflected in the effort of system integrators to provide tooling with their systems that allows energy consumption of components, as recorded by many sensors in the machine, to be related to job execution. Sites still account resource usage in terms of core hours rather than kWh, but energy efficiency incentives at the application level can be present in other forms. LRZ, for instance, allows applications to run on a higher clock frequency only if they have been subject to analysis in which high energy efficiency was demonstrated.

Global energy optimisation approaches drive monitoring to higher levels of complexity as more and more parameters from hardware and infrastructural equipment have to be related to batch system and application behaviour data. This poses challenges that go beyond mere improvement of control over the infrastructure. Even though control systems get more complex in their own right and combine more and more parameters, they chiefly deal with current values. Global energy optimization of applications requires consolidation of many data points for statistical analyses.

Both for dashboards giving a comprehensive overview of the datacentre, and data consolidation on behalf of analyses, more interoperability between devices at the monitoring level is needed, but is currently lacking. Many state-of-the-art HPC sites presently cope with this situation in eclectic ways, combining several complementary partial solutions in their own way. Microsoft deals with many datacentres and has reported that massive data collection over these centres and feeding the data into machine learning processes has resulted in higher reliability through prediction of failure and subsequent proactive replacements. Without sufficient standardisation of their datacentres this would have been more difficult. However, HPC sites usually deal with a single highly specialised and customised data centre, or a few at most. While a top down implementation of standardization is unfeasible and undesirable, more interoperability, an open interface, and an API for monitoring should be high on the agenda.

12.2 Current situation in Europe

Regarding Europe, the workshop is an important place for sharing best practices among HPC centres and especially the HPC centres of PRACE partners. It contributes to the high level of expertise and stimulates the implementation of advanced technologies in HPC centres. It confirms that HPC centres in Europe are able to host and operate large supercomputers in an energy efficient way.

In European HPC datacentres, direct warm water cooling of at least the dense compute racks, is becoming a consolidated best practice. Centres are now exploring possibilities to extract even higher waste heat temperatures from these systems and novel ways of heat re-use within the data centre itself. Given the variation in climate conditions in Europe, chillers will be on the way out of most data centres when technologies that can produce chilled water for less dense systems by exploiting the waste heat of HPC systems mature.

Despite the emphasis on energy efficiency, the power needed to drive new HPC systems with higher compute capacities still tends to increase. In other words: the increases in energy efficiency cannot yet compensate completely for the growth of the demand for compute capacity. Because the price of energy also tends to increase – although at different rates in different countries – this also has an increasing effect on the total cost of ownership. HPC centres in Europe display a very high awareness of cost and power consumption issues. 30 MW is probably a barrier that will not be crossed, not even by an exascale system – In Europe or elsewhere – in 2022 or 2023.

In their main production environments centres have not yet radically shifted from time to solution policies to energy to solution policies. However, as monitoring tools enabling more direct attribution of changes in resource consumption by IT and other data centre infrastructure equipment to the behaviour of application runs are becoming available, centres can adopt the more integral approach to run applications energy efficiently rather than to provide an energy efficient HPC infrastructure for just any generic HPC application. This approach will inevitably blur the lines between application and infrastructure optimization and will drive monitoring to greater complexity and sophistication. In their presentations at the workshop, various system integrators and architects of future HPC processors, have shown that they are well aware of the growing demand for means to collect and insightfully organize data on system and application behaviour.

Not all monitoring data have to be subject to causal analyses before they can contribute to effective control and pro-active maintenance of IT and datacentre infrastructures. Promising initiatives that report the use of machine learning techniques and statistical approaches to reduce complexity of the vast amounts of monitoring data were noted at the workshop.

In the context of these trends the special focus of this workshop on monitoring and the presentation of monitoring practices being developed at PRACE sites proved to be a very timely choice.

13 Conclusions

Like its predecessors, the 7th HPC Infrastructures Workshop has been very successful in bringing together experts working on HPC site infrastructures. Figure 11 shows that the annual workshop has become an institute that is capable of consistently attracting a stable, if not a growing, number of experts in the field.

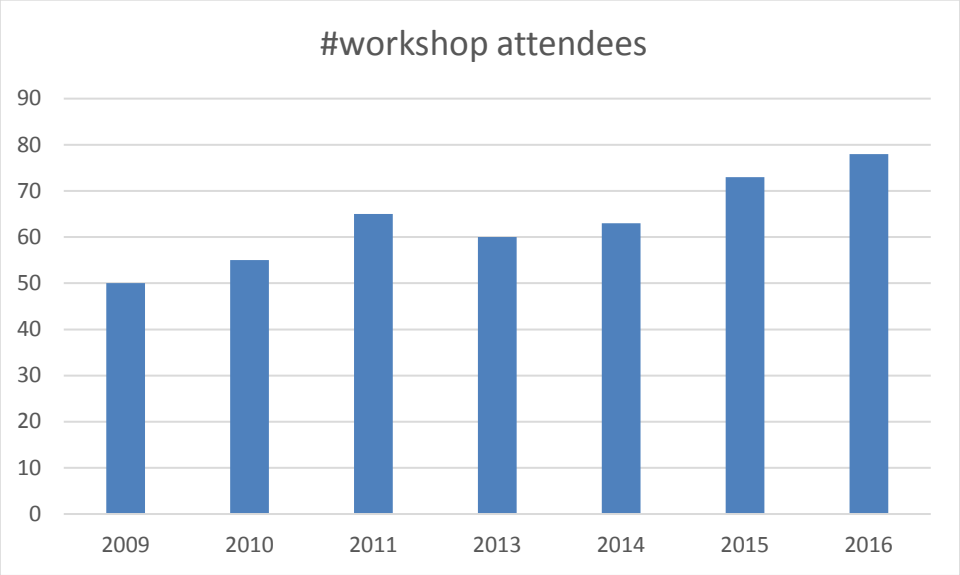

Figure 11: Number of workshop attendees in a historical perspective

The workshop made possible the identification of important trends and an assessment of the current situation within Europe. This year's special focus on monitoring was very well in line with observable trends and with the importance European centres attribute to this topic and the efforts they are currently spending in this area of interest.

Regarding trends, it is worth noting that:

- Denser racks with power consumption of 100 kW or more and weighing more than 1000 kg will become common in the near future;
- Direct liquid cooling (DLC) is the state-of-the-art cooling technology for these compute racks;
- The aim for DLC in the near future is to produce even higher outlet temperatures, as heat re-use of hot water has more applications and is generally easier and more efficient than heat re-use of lukewarm water;
- Opportunities for effective heat re-use outside the data centre are site specific and vary across sites and promising techniques for heat re-use in the datacentre, building the HPC system into the cooling plant for less dense systems, are currently being used in experimental systems;
- The emphasis on energy efficient IT equipment and datacentre infrastructure cannot yet completely compensate for the demand for more compute capacity that drives the deployment of new HPC systems. The power consumption of future top HPC systems is expected to still rise, but not to exceed the 30 MW barrier;
- There is a trend towards more encompassing, sophisticated monitoring that allows the attribution of resource consumption at various levels directly to application runs. Development of this type of monitoring is a prerequisite for integral energy to solution optimization of HPC applications.

The next workshop will take place in April 2017 and will be hosted by CSCS, in Lugano (Switzerland).