E-Infrastructures
H2020-EINFRA-2016-2017

EINFRA-11-2016: Support to the next implementation phase of Pan-European High Performance Computing Infrastructure and Services (PRACE)

PRACE-5IP

PRACE Fifth Implementation Phase Project

Grant Agreement Number: EINFRA-730913

D5.3

HPC Infrastructures Workshop #8

Final

Version:	1.0
Author(s):	Huub Stoffers, SURFsara
Date:	20.11.2017

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: EINFRA-730913**	
	Project Title: HPC Infrastructures Workshop #8	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: D5.3>	
	Deliverable Nature: Report	
	Dissemination Level: PU *	**Contractual Date of Delivery:** 30/04/2018
		Actual Date of Delivery: 30/11/2017
	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 2005/444/EC.

Document Control Sheet

Document	**Title: HPC Infrastructures Workshop #8**	
	ID: D5.3	
	Version: 1.0	**Status:** *Final*
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2013	
	File(s): PRACE-5IP-D5.3.docx	
Authorship	**Written by:**	Huub Stoffers, SURFsara
	Contributors:	Javier Bartlome (BSC), Ladina Gilly (CSCS), Jean-Philippe Nominé (CEA), François Robin (CEA), Susanna Salminen (CSC), Gert Svensson (KTH), Torsten Wilde (LRZ)
	Reviewed by:	Miroslaw Kupczyk, PSNC
	Approved by:	Veronica Teodor, JUELICH

Document Status Sheet

Version	Date	Status	Comments
0.1	04/08/2017	Draft	Some session reports still missing, some need improvement
0.2	22/08/2017	Draft	Complete proceedings text, no conclusions, no illustrations, no lay-out
0.3	25/08/2017	Draft	Complete draft, sent for comment/review to WP5 members participating in the workshop
0.4	30/10/2017	Draft	Processed comments on 0.3 version of WP5 members participating in the workshop
[0.5	06/11/2017	Draft	With minor corrections and suggestions from WP5, version for PRACE internal review
1.0	20/11/2017	Final version	Minor corrections after PRACE internal review

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, HPC facility. datacentre, Energy efficiency, Cooling

Disclaimer
This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement n° EINFRA-730913. It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements. Please note that even though all participants to the Project are members of PRACE AISBL, this deliverable has not been approved by the Council of PRACE AISBL and therefore does not emanate from it nor should it be considered to reflect PRACE AISBL's individual opinion.

Copyright notices

Table of Contents

List of Figures

References and Applicable Documents

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

[2] https://www.ashrae.org/resources--publications/handbook

[3] https://webstore.iec.ch/publication/33927 (IEC 60755:2017)

[4] https://www.en-standard.eu/csn-en-50600-1-information-technology-data-centre-facilities-and-infrastructures-part-1-general-concepts

[5] https://www.top500.org/project/linpack

List of Acronyms and Abbreviations

aisbl	Association International Sans But Lucratif (legal form of the PRACE-RI)
ASHRAE	American Society of Heating, Refrigeration, and Air-Conditioning Engineers
BCO	Benchmark Code Owner
BEO	Bull Energy Optimizer
BUNA-N	acrylonitrile-butadiene rubber, a synthetic rubber with good resistance to oils and solvents
CAPEX	Capital Expenditure
CGG	Company name, originally an acronym for Compagnie Générale de Géophysique
CMOS	Complementary Metal Oxide Semiconductor
CoE	Center of Excellence
COP	Coefficient Of Performance (for heatpumps or cooling systems)
CPU	Central Processing Unit
CRAC	Computer Room Air Conditioner
CUDA	Compute Unified Device Architecture (NVIDIA)
DARPA	Defense Advanced Research Projects Agency
DBIA	Design Build Institute of America
DEISA	Distributed European Infrastructure for Supercomputing Applications EU project by leading national HPC centres
DoA	Description of Action (formerly known as DoW)
EC	European Commission
ECMWF	European Centre for Medium-Range Weather Forecasts (Reading, UK)
EEHPCWG	Energy Efficient High Performance Computing Working Group
EER	Electrical Efficiency Ratio
EESI	European Exascale Software Initiative
EoI	Expression of Interest

ESFRI	European Strategy Forum on Research Infrastructures
ESIF	Energy System Integration Facility
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
GPU	Graphic Processing Unit
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.
HMM	Hidden Markov Model
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPL	High Performance LINPACK (a benchmark)
HT-DLC	High Temperature Direct Liquid Cooling
IEC	International Electrotechnical Commission
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
LINPACK	Software library for Linear Algebra
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
MCCB	Moulded Case Circuit Breaker
MFlop/s	Mega (= 10^{6}) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MOOC	Massively open online Course
MoU	Memorandum of Understanding.
MPI	Message Passing Interface
MSBs	Main switch boards
MW	Megawatt
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
OIC	Oil Immersion Cooling
PA	Preparatory Access (to PRACE resources)
PATC	PRACE Advanced Training Centres
PDUs	Power Distribution Units
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

PUE	Power Usage Effectiveness
RI	Research Infrastructure
RFQ	Request for Qualifications
RoI	Return on Investment
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
UEABS	Unified European Application Benchmark Suite
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources.
UPS	Uninterruptible Power Supply
VRLA	Valve Regulated LeadAcid

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 (3 rd 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 (3 rd Party to GENCI)
CNRS	Centre National de la Recherche Scientifique, France (3 rd 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)
IT4Innovations	IT4Innovations National supercomputing centre at VŠB-Technical University of Ostrava, Czech Republic
IUCC	INTER UNIVERSITY COMPUTATION CENTRE, Israel
JKU	Institut fuer Graphische und Parallele Datenverarbeitung der Johannes Kepler Universitaet Linz, Austria
JUELICH	Forschungszentrum Juelich GmbH, Germany
KIFÜ (NIIFI)	Governmental Information Technology Development Agency, Hungary
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
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 8th European Workshop on HPC Centre Infrastructures was organised by CSCS and held in Mendrisio, a town near Lugano, Switzerland, between April 5, 2017 – April 7, 2017. BSC, CEA, LRZ, PDC-KTH, and PSNC have collaborated in the committee organising the workshop.

This workshop, upon invitation only, was very successful, with 75 participants coming from Europe, America, Australia, and Asia.

The workshop covered a broad range of topics relevant for HPC Centre Infrastructure management: standards and regulations, energy effciency strategies, total cost of ownership reduction strategies, energy storage technologies, and trends in the design as well as procurement procedures for datacentres. The workshop brought together experts from the vendor side and experts from the HPC date centre facility management side. Several presentations from datacentre sites gave an insightful look into the “kitchen” of their facility management, on topics such as oil immersion cooling, and quality control of UPS batteries.

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

The workshop made possible the identification of important trends and assessments on the situation in Europe in terms best practices in facility management, infrastructure design, and facility procurement of HPC centres.

1 Introduction

The 8th European Workshop on HPC Centre Infrastructures was organised by CSCS and held in Mendrisio, a town near Lugano, Switzerland, between April 5, 2017 – April 7, 2017. BSC, CEA, 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
- Torsten Wilde, 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 75 participants coming from 17 countries, including USA, Australia and Japan. 23 European sites were represented, 17 associated with PRACE, and 6 non-PRACE sites. 4 non-European sites were represented, 1 from Australia, 1 from Japan, and 2 from the USA. For this workshop, no commercial datacentres participated, as none were invited this time.

The workshop brought together experts from datacentres with a rich palette of expertise on the vendor side: besides HPC system integrators and companies engaged in processor technology, vendors of energy storage solutions participated, as did vendors of specialised in testing and monitoring the safety of electrical installations, and vendors of specialised cleaning technologies for water cooling installations.

Figure 1: A workshop session - Mike Patterson of Intel, answering questions of other participants, after his talk on cooling strategies for datacentres

This workshop covered a broad range of topics relevant for HPC centre infrastructure management: standards and regulations, energy efficiency strategies and policies, energy storage technologies, and trends in facility design as well as procurement procedures for HPC datacentres.

At the end of the workshop a PRACE closed session was held. Several PRACE site representatives gave an update on specific datacentre infrastructure developments, there was an update on the PRACE Pre-Commercial Procurement (PCP) (PRACE-3IP) and planning for further development of PRACE-5IP WP5 took place.

2 Programme, content and speakers

Wednesday, April 5th 2017

- Session I - Site updates
  - Site update CSCS – Ladina Gilly, CSCS
  - Experiences with oil immersion Cooling in a processing datacentre 2017 – Michael Garcia, CGG
  - Servers cooling themselves: perspective on adsorption cooling for datacentres – Matthias Hoene, Fahrenheit AG

- Session II - Standards and Regulations
  - EE HPC WG: Connecting infrastructure and HPC systems – Natalie Bates, Energy Efficient Working Group, LLNL
  - Liquid cooling, HPC can't live on W3 alone! – Mike Patterson, Intel
  - Residual Current Monitoring (RCM) and electrical regulations in European countries – Frank Riedo, Riedo Networks

- Session III – TCO
  - Ways to decrease TCO and infrastructure related costs – Willy Homberg, FZJ
  - Using TCO considerations in HPC procurements – Gert Svensson, KTH
  - TCO of adsorption cooling – Torsten Wilde, LRZ

- Session IV - Site Visit to CSCS
- Social Event, Workshop dinner at Ristorante Capo San Martino

Thursday, April 6th 2017

- Session V - Design and Procurement
  - An approach to control the sharp fluctuations of power consumption and heat load of the HPC system – Shoji Fumyjoshi, RIKEN
  - An HPC experience with Design-Build – Steve Hammond, NREL
  - Challenges and opportunities of "slab on grade" datacentre design – Jim Rogers, ORNL

- Session VI - Energy Storage
  - Lead battery life cycle management – Tiziano Belotti, CSCS
  - Lithium batteries – Morten Stoevering, Schneider Electric
  - Application of fuel cells for UPS – Felix Büchi, PSI

- Session VII – Technologies
  - Power Management Framework: A consistent vision to manage your cluster – Abdelhafid Mazouz, Atos

  - ABB Synchronous reluctance (SynRM) motor & drive package - Super premium efficiency for HVAC application – Huy-Hour Lor, ABB
  - ABB Cognisense Motors – Condition monitoring of HVAC LV motors through the Internet of Things – Tom Bertheau, ABB
  - Preventing biofilm in cooling installations with the use ultrasound – Michael Widegren, Creando

- Session VIII – Vendor Panel
  - IoT in the datacentre - a vision of ge2e HPC operations – Martin Hiegl, Lenovo
  - The IBM-ASTRON DOME micro datacentre technology – Ronald P. Luijten, IBM
  - Path to exascale - Options and obstacles – Frank Beatke, HPE
  - Vendor Panel discussion: Martin Hiegl (Lenovo), Ronald P. Luijten (IBM), Frank Beatke (HPE), Mike Patterson (Intel)

- Social Event, Workshop dinner at Ristorante Grotto Loverciano

Friday, April 7th 2017

- PRACE Session
  - BSC
  - Cineca
  - PCP
  - SURFsara
  - CEA
  - GRNET
  - WCNS

- PRACE 4IP/5IP WP5 meeting

3 Session I - Site updates

3.1 CSCS site update – Ladina Gilly

In their site update CSCS reported the definitive solution to a problem in the mechanical parts of the datacentre infrastructure that had bothered them since autumn 2013. The presentation contained insightful details of the path they had travelled towards the definitive solution, including some hypotheses about the root cause that proved wrong, and some intermediate solutions adopted along the way.

The heat exchangers (HE) of the CSCS datacentre use cold water that is pumped from Lake Lugano. These HE have a capacity of 504m3/h each, needed to sustain the full cooling capacity of the centre. However, in October 2013 unexpectedly high pressure drops accompanied by up to 50% reduced flowrates were detected. Subsequent investigation revealed that parts of the infrastructure had clogged up after a rapid ramp-up of the lake water flowrate. The self-cleaning filters in pumping stations were overwhelmed with “algae-like” growth that was pulled in from the suction baskets in the lake. Biological analysis identified the growth as the iron eating bacteria *Leptothrix ochracea*.

At that point the bacteria, and the rubbish they produce became the main suspects for the pressure drop and flowrate problems. The first intermediate resolutions adopted to solve the problem were:

- The opening of the HE to clean them mechanically: the bacterial residue is easily removed with a garden hose.
- The regular cleaning of the suction baskets in the lake, initially by hand every six month. Now they are cleaned every three months by a diving company with an underwater vacuum cleaner. Thanks to the new method this cleaning can take place without interruption to service.
- A reduction of the mesh of the self-cleaning filters from 0.5mm to 0.1mm.
- Cleaning of the filters in front of the HE every two weeks.
- Construction of an additional HE, with 75 kW pumps rather than the 17kW pumps envisaged in the original design.

The reduced mesh size of 0.1mm did not prevent the bacteria from getting in, but most of these measures brought some improvements. However, none of these intermediate steps solved the problem satisfactorily. The flowrate remained far too low. Extensive measurements on the HE and adjacent piping followed and showed that the flowrate issue was due to a number of problematic aspects in the piping:

- Too many diameter changes in rapid succession in the piping trajectory
- The quality of the welding of the polyethylene (PE) piping in a number of instances was suboptimal. This caused “welding lips” to protrude into the pipes and caused additional resistance,
- Too many right-angle curves and T-junctions in the piping, causing excessive turbulence and high pressure drop across the HE.

Refurbishment of the main cooling distribution was undertaken in four phases. The first one focused mainly on the piping. The PE piping was replaced with stainless steel piping, with full recalculation of pressure drops across the system. The piping was also redesigned to reduce the number of right-angle curves and T-junctions. The original pumps were replaced with a slightly more powerful model (37 kW instead of 17 kW). In addition, filters were moved to more optimal positions.

The second refurbishment modified the flow path of the lake water. Originally the water was led directly to the HE. Now it is first routed to a large holding reservoir, to allow sedimentation. This has been proven to significantly reduce the amount of debris and bacteria that reaches the HE.

With that route in place, a subsequent third improvement was the addition of an “elbow” pipe, to raise the level at which water is taken from the holding reservoir, to avoid sucking in sedimented debris and bacteria.

The fourth project focused on tests with various substances to clean the HE chemically, that is, without having to open it. The outcome is that it can be done, but needs careful preparation and precautions. It entails introduction of additional filtering steps to make sure that the chemicals are taken back out again, before the water is returned to the lake. The ability to clean the HE, without having to open it is of crucial importance for the ability of the datacentre to grow. If power consumption is increased to 20 MW, all of the HE capacity is needed. There no longer is the redundancy to take them out, and clean them, one by one.

With all modifications in place, CSCS now achieves a flowrate of 580m3/h with the correct pressure drop. They will continue the regular cleaning of the suction baskets and filters, and will keep monitoring pressure differences in the HE to observe deterioration over time. The mesh size of the filters in front of the HE will be reduced, from 1mm to 0.5mm.The routing via the sedimentation reservoir, has proven to be an important decision. Nonetheless, water quality will continue to be monitored. Chemical cleaning will be done if degradation of the water flowrate or increase of pressure drop is observed.

3.2 Experiences with Oil Immersion Cooling in a processing datacentre - Michael Garcia, CGG

Michael Garcia is part of the datacentre service group at CGG. CGG is a globally operating geoscience and seismic imaging company that provides geological, geophysical and reservoir capabilities to its customers – primarily the global oil and gas industry. CGG operates its own three main datacentres, one datacentre for a particular area of the world:

- London, for Europe, the Middle East, and Africa
- Singapore, for Asia
- Houston, for America

Per site the datacentres typically house more than 1000 GPU units and more than 10,000 CPU sockets. The conditions in these locations differ considerably and so do the cooling technologies used. In London free cooling is used. The one in Houston has two computer rooms, one with tradition CRAC cooling and one with oil immersion cooling (OIC). The room with OIC is in production since June 2011 and now houses equipment with an average power usage of about 1 MW.

The motivation to implement an OIC solution came from studies conducted by CGG in 2009 that investigated the options to upgrade an old datacentre, built in 1992, that was operated with a PUE of about 2.0. The studies showed that the OIC option, developed by Green Revolution Cooling, provided significantly better return on investment than any other feasible option – for this particular case. OIC is not appropriate for all locations and business scenario's. CGG e.g. definitely does not consider switching to oil in their London datacentre, which thrives with free cooling and is operated with a PUE of about 1.05.

In the Houston case, some capital expenditure (CAPEX) would be avoided by the OIC option, some CAPEX would be deferred and distributed over time, and power savings would lead to a significant reductions in operational expenditure (OPEX). More specifically, the raised floor of the old datacentre could simply remain in place. The CRACs could be removed, but were to be replaced by oil pump units. The old "vertical" racks would be replaced by "horizontal" oil containers. IT equipment that would otherwise be stacked, would thus be more spread out across the floor. Comparatively more floor space is needed for the same amount of IT equipment, but the availability or cost of floor space was not a bottleneck in this case.

To compare their air cooled Houston computer room, which is operated at a PUE of 1.34, with their OIC room, CGG use the concept of an "equivalent PUE". Using the PUE metric, the OIC room would appear less efficient than it actually is. This is because fans in servers are considered part of the IT equipment and so the power usage of the fans is counted as IT power usage. But for servers immersed in oil, the fans are taken out. This leads to a power usage reduction of about 20%. Taking this into account, the "equivalent PUE" of the room that uses OIC is 1.05. The power usage of the oil pumps is negligible.

Since the room is in operation since 2011, CGG can report on some oil immersion challenges as well as oil immersion bonuses. Some of the challenges appear to be very component specific. CGG found that the cumulative failure rate for NVIDIA Kepler GPUs was the same for air cooled and oil cooled systems. But for Maxwell, another generation of NVIDIA GPU, the cumulative failure rate in oil was twice as high as in air!

Capacitor plugs tend to absorb oil and swell. At some point solder joints crack and capacitors can fall off the boards. But solutions for this are readily available, protective conformal coating can be applied, and oil resistant plugs are available too. A similar case is thermal paste degradation. Silicon based thermal pastes degrade in oil. The readily available solution is to use oil resistant thermal interface materials, for example Indium. Optical network components in oil degrade over time. CGG only uses copper connections.

Normal power cords stiffen over time and become unreliable. Oil resistant BUNA-N power chords are a suitable alternative.

Among the bonuses of OIC rooms is the sound level. IN CGG air cooled datacentres sound levels reach up to 96 dB. Hearing protection is required and collective in situ system trouble shooting is hampered. In an oil cooled datacentre the sound levels are around 50 dB. Another bonus is that the thermal inertia of an oil bath is about a 1000 times higher than the thermal inertia of air. In case of a loss of cooling water air systems throttle and must shutdown in five to ten minutes. Oil immersed systems easily can bridge an interval of thirty minutes or more. An oil bath also safely absorbs the heat pulses that are caused by synchronous job starts.

3.3 Servers cooling themselves: perspective on adsorption cooling for datacentres – Matthias Hoene, Fahrenheit AG

Fahrenheit is a new brand of the Sortech company. Sortech has been very much a research and development company. It has been awarded over twenty patents in the last 15 years. The Zeolite technology, used in adsorption cooling, is now ready for broader roll out to the market. This brings about a number of changes that warrant the rebranding.

Adsorption is the method of using solid materials for cooling via evaporation. It is a method of two phases:

- Water is evaporated using a solid adsorption material. Evaporation extracts heat and creates cold. Water vapour formed in this process is taken up by the adsorbent.
- Once saturated, the adsorption material is heated to release the water. The discharged water vapour flows into a condenser where it liquefies again.
- Two identical modules operate phase-shifted to provide continuous cooling

Adsorption is heat driven, so it allows the using of residual heat for cooling – rather than electricity. Adsorption cooling is already in use in HPC datacentres, such as LRZ, today. The heat source hot water coming from direct liquid cooled processors. The chilled water produced by the adsorption cooler is used for cooling other components of the system or in the datacentre.

Current adsorption coolers are almost as efficient as using free cooling, but without having to rely on cold weather. Two measures are particularly relevant for quantifying the energy efficiency of the device:

- The Electrical Efficiency Ratio (EER), which is the ratio of cooling power and electrical power consumed
- The thermal Coefficient Of Performance (COP), which is the ratio of cooling power output over heat power input

The EER of Fahrenheit’s adsorption coolers is in the range 15-20. In other words: for every kW of electrical power you put in, you get up to 20 kW of cooling power. The thermal COP is in the range 0.55 – 0.65. In other words: for every kW of waste heat that you put in, you get up to 650 Watt of cooling power.

The technology is evolving. A different adsorbent, zeolite, a material based on aluminium crystals, generally performs better than the originally used silica gel:

- Zeolite crystallisation has an advanced working range (heat dissipation temperature down to 40° C).
- Lower power to weight / volume ratios are feasible. It is easier to maximize the surface of the heat exchanger for a given volume
- Lower electrical energy consumption for the re-cooler, so a higher EER
- Faster adsorption cycle, reduced start up time

However, the thermal COP of zeolite is somewhat lower than that of the more traditional silica gel.

With zeolite as an absorbent, the cooling power per surface area is improved. This allows for more cooling power in a smaller footprint. At least as significant for applicability is that zeolite can be adapted to the temperature range of the application.

With the possibilities for miniaturisation and current power densities combined, in principle the point is reached where the adsorption unit can be rack-fitted, which could make hydraulic integration easier. Fahrenheit has not built this yet and is not sure whether this is the way to go. They would appreciate some feedback from the community whether they would prefer rack-integrated adsorption over separate dedicated adsorption cooling units.

The CoolMUC-2 machine at LRZ is an early adopter of adsorption cooling technology produced by Sortech – the predecessor of Fahrenheit. Their cooling units are based on silica gel adsorbent. The key results for CoolMUC-2 are:

- 120 kW of waste heat from the HPC racks is removed and is reused to generate
- 50 kW of cold produced for the storage units
- 9 kW of electric power is consumed in this process to drive pumps, fans, etc.

In new datacentres, the investment case for adsorption cooling is quite compelling. Pay back, via electricity savings, are in the order of one year. In existing datacentres, case by case evaluation is necessary. But typically energy savings pay back the investment in a reasonable amount of time.

The maintenance cost of adsorption coolers are expected to be comparatively low. They have pumps on the outside, but other than that, there are no moving parts, no mechanical components that wear out.

Discussion with the audience focus on other elements of the business case for adsorption cooling. Of course a lot depends on the local context of the datacentre: adsorption cooling makes most sense in a situation where there is otherwise unproductive waste heat to use as an input and where there are also components that need the cold water produced for their cooling. Statements about the time interval for return on investment cannot but make assumptions on the price per kWh, since the pay back must mainly come from electricity saved. The calculations of Fahrenheit assume a price of 15 Euro cents per kWh.

Feedback from the audience on Fahrenheit's question about usefulness of rack-builtin adsorption units, was that rack-builtin would allow for easier adoption of adsorption cooling by parties who are in a rented datacentre in which they have less of a say on the infrastructure that is extraneous to the racks.

4 Session II - Standards and Regulations

4.1 EE HPC WG: Connecting infrastructure and HPC systems – Natalie Bates, Energy Efficient Working Group, LLNL

Economics of energy efficiency

Current development in HPC datacentre construction and operations is shifting the cost balance. For example, at NCSA the ratio of IT investment to facility was in 1984 15-to-1 (IT system was 15 million $US, facility was 1 million $US). In 2003 the ratio changed to 5-to-1 (IT system was 226 million $US, facility was 85 million $US). The ratio of OPEX and CAPEX for the Bluewater system over a 5-year lifetime is estimated at OPEX 40% and CAPEX 60%. If OPEX becomes dominate current CAPEX focused procurements will need to change.

JSRM (Job Scheduling & Resource Management)

The EEHPCWG is working on a survey of sites that use or plan to use energy/power aware scheduling. This survey includes 10 HPC datacentres situated in the USA, Europe, and Asia. The goals of the work is to suggest future steps in improving HPC datacentre scheduling. This can be related to a changing landscape in which HPC datacentres live. For example, tighter power grid integration where features such as managing and predicting total load of site and/or IT system might be required.

There are two papers available on website. The latest focused on power/energy contracts of HPC sites. One of the results is that current contracts encourage demand flexibility but this does not translate to demand / response. Also high equipment costs prevent more flexibility. A new paper "Taxonomy of contract elements" is work in progress. It looks into datacentre power contracts and how they affect the approach to power and energy by the datacentres.

The main insights are:

- fixed energy costs encourage energy savings
- power bounds encourage simple power management
- dynamic power costs encourage sophisticated power management

Liquid Cooling Controls

The WG is currently working on defining data inputs for controls system related to different types of liquid cooling.

Dashboard

The dashboard team is currently defining important information that needs to be seen (visualized) by HPC datacentres.

4.2 Liquid cooling, HPC can't live on W3 alone! – Mike Patterson, Intel

In the context of designing new datacentres, Mike Patterson too often encounters a too simplified or too limited view on cooling issues. In a somewhat provocative mode, he poses the following questions:

- "The future is ALL about liquid cooling, right?
- And obviously, WARM water – ASHRAE W3 or even warmer - has the best TCO, right?"

Is an infrastructure for warm water cooling all that is needed in a state of the art and future proof datacentre? No, for most datacentres, it definitely is not.

The Guidebook of the American Society of Heating, Refrigerating, and Air-conditioning Engineers (ASHRAE) [2] is the best resource on liquid-cooling that is available today. Although the name suggests otherwise, members of over 140 nations now participate in ASHRAE. Technical

Committee 9.9, for “mission critical facilities” is the one that deals, among other things, with datacentre cooling and is the largest ASHRAE technical committee, with over 100 members. The committee includes all relevant equipment vendors. Air cooling thermal environments was the primary work, resulting in air-cooling classes A1, A2, etc. Liquid cooling was added later. Facility inlet water temperature is key in defining the liquid cooling classes W1 – W5. But the guide also covers designs, materials of construction, and water quality.

So, what does this authoritative source on the matter have to say? Mike points out that the following points are very important to take home:

- Water quality is very important – if it is ignored, you will fail
- For maximizing datacentre efficiency while minimizing TCO, the best *starting point* for your new datacentre design is to run AS COLD AS POSSIBLE without a chiller.

The latter rule, by the way, applies both for air-cooling and for liquid-cooling.

What it means for a server that it was designed to work at higher temperatures is that it will not fail at these temperatures. But it will still perform better at lower temperatures, be able to “turbo up” more frequently, etc. The gain in performance could be an easy trade for extra cooling capacity. And so could be the increased longevity and reliability in the long run.

W3 will work well for dense computing equipment. 32°C is “cold” for a CPU, but “hot” for some other equipment. Cooling water at W3 does not allow for dehumidification. 32°C water results in air that is too warm – about 37°C. W3 does not work for current archives, for storage and service nodes. Neither is it a good temperature for humans, for technicians that have to do maintenance in the computer room.

There will still be a need in a datacentre for low temperature water. For most datacentres a combination of W3 water cooling and A1 air cooling is optimal. HPC cannot live on W3 alone!

4.3 Residual Current Monitoring (RCM) and electrical regulations in European countries – Frank Riedo, Riedo Networks

Switching power supplies generate leakage currents over the ground connection. In HPC systems this adds up to deadly current. Faulty grounding can kill, therefore, the datacentre installations need to be tested regularly for this. Though regulations differ from counry to country, most are based on IEC standards such as IEC 60755, “General safety requirements for residual current operated protective devices” [3]. The European standards series EN 50600 “Information technology - datacentre facilities and infrastructures” [4] aims to end the lack of a single standard for datacentre facilities and infrastructures within Europe. It is comprehensive, but currently its status is that of a non-mandatory best practice guide. However, most national standards already require periodic testing of residual current devices (RCD) that break the power when a critical level is detected. Traditionally, the datacentre installations under test would need to be off-line during the tests.

There is an opportunity for new technologies that do not require off-line testing. Residual current monitoring (RCM) is one technique that can be used. The main benefits of RCM are:

- Residual current levels are monitored continuously
- RCM can react to changes long before the critical threshold is reached
- RCM sends alarms, it does not cause automatic shutdown
- Resistance measurements are not required

For optimal localization of errors, the RCM unit needs to be placed as close to final power consumers as possible. RCM can be integrated in newer power distribution units (PDUs).

5 Session III – Total cost of Ownership (TCO)

5.1 Ways to decrease TCO and infrastructure related costs – Willy Homberg, FZJ

“Ways to decrease TCO and infrastructure related costs” is the topic of a forthcoming PRACE-5IP project white paper. Jülich Supercomputing Centre (JSC) is taken as an example for initial reflections on this topic – Willi Homberg, from JSC, is leading the white paper effort.

JSC envisions a ‘dual track’ approach where systems can be combined in a mix of conventional clusters and more specific elements like ‘booster’ modules (e.g. cluster of KNL) in the supercomputing centre – together with suited global data and resource management.

PRACE-5IP WP5 white paper purpose and goals are to:

- Define best practices for using TCO analysis in the procurement of HPC systems
- Make aware of the most relevant cost factors which must be considered when purchasing, installing or upgrading, and operating an HPC system
- Draw attention to the increasing importance of infrastructure works in view of the upcoming Exascale systems and their strongly growing resource consumption
- Develop strategies for cost reduction without decreasing the value of investment

TCO analysis and insight can help get the best computing capacity for a given TCO, and/or get a defined computing capacity for the lowest budget.

TCO cost categories can be broken down into classical categories: investment (IT equipment), initial application porting incl. staff training; upgrades along the lifespan of the system; building and technical facilities adaptation; operational costs: IT equipment and facility maintenance, power, cooling, staff, insurance, security and monitoring.

Cost cutting strategies and the related parameters that can be considered are manifold and sometimes lead to different trade-offs:

- Density/space footprint
- Re-use of sustainable infrastructures
- Right-sizing of cooling and power

- Cooling technologies (full air / cold doors / DLC …)
- Power management (load management, redundancy level, tracking power distribution losses)
- More dynamic power management (energy-aware scheduling, virtualization…)
- Fault-monitoring and prediction for enhanced availability

A survey of state-of-the-art practices in these areas has started and includes six European HPC centres so far. The final report, a white paper, is planned to be delivered in October 2017.

5.2 Using TCO considerations in HPC procurements – Gert Svensson, KTH

This is a testimony and report of experience from KTH – the biggest computing centre in Sweden, with Linköping. KTH has used some TCO concepts in their recent procurements and would welcome discussions with other centres that might have similar experience.

TCO places a single value on the complete lifecycle of a purchase. In an attempt to better reflect the true cost, not only the acquisition price is taken into account, but also maintenance cost, and in principle the cost of all elements that are needed to make the system used and useful.

TCO elements considered are:

- Purchase price - installation cost - maintenance: related information is coming from the vendor
- Facility adaptation: cost to prepare the facility for a specific system; here again the vendor describes what is needed with sufficient detail; then either the buyer or an appointed expert/consultant does a cost estimate, or the vendor can be given the entire installation setup – KTH preferred management by local facility people
- Cost of power: this is application and workload dependent; the vendor is asked to provide measurements/limits for different standard (HPL) [5] or specific benchmarks (real applications). Some formula can be used to calculate a ‘normal’ power consumption estimate (e.g. 0.8 x HPL-power)
- Cost of cooling: this is climate-dependent; heat re-use, also climate-dependent, can reduce cost; this is also influenced by output temperature of water, not easily guaranteed by vendors. KTH purchases district cooling.
- Cost of using existing equipment for cooling and power: depreciation and maintenance costs are assigned to equipment using a key based on respective power usage. Estimates are calculated (via a spreadsheet) using vendor measurements on benchmarks.
- Costs of using space in the computer hall: layout of equipment is left to the vendor; KTH rents its facility, but floor space has so far not been an issue
- Cost of staff for operation of the system: system-dependent, especially in terms of application porting. This cost has not been considered directly in KTH procurements so far.

It is discussed, and widely acknowledged, that TCO must be related to productivity, scientific outcome – it is clueless if taken from a mere accounting standpoint.

Procurements can be steered in different ways, given a set of minimum requirements:

- Seeking lowest price: requirements fulfilled with lowest TCO
  Fixed price: highest capacity that fulfil the requirements under maximum TCO

More complex combinations can be used, with value functions for extra features beyond minimal requirements. The selected bid would then fulfil all minimum requirements with lowest evaluation cost.

In many organizations budgets are split in different areas and funds can not be transferred between the areas (this is the case at KTH: one budget each for purchase, maintenance, power

and cooling). To handle this, additional minimum requirements are introduced (e.g. purchase + maintenance + power + cooling < budget). The vendor can choose the largest system that fulfils the requirements. The selected bid fulfils all minimum requirements with the lowest evaluation cost.

KTH used these approaches, using simple Excel sheets for the vendors. It was not an excessive amount of work and the approach was well-received by the vendors.

5.3 TCO of adsorption cooling – Torsten Wilde, LRZ

In the LRZ datacentre chiller supported cooling is about four times as expensive as chiller-less cooling. LRZ’s long term goal is it to remove all mechanical chillers. Adsorption chillers are an option for this, when used together with high temperature Direct Liquid Cooling since it can generate cold water close to the efficiency of chiller-less cooling.

The LRZ CoolMUC-2 system is used to explore the potential of adsoption cooling. CoolMUC-2 is a Lenovo NeXtScale cluster with HT-DLC. It is operated with water inlet temperatures between 30˚C and 50˚C, and runs with all season chiller-less cooling. The system has 384 compute nodes (2 x14 core Intel Haswell Xeon E5-2697 v3 per node). Its peak performance is 466 TFlop/s peak performance. CoolMUC-2 held position #356 on the Top500 list of June 2016.

Of 120kW of IT power, 86kW is captured in the HT-DLC loop (node inlet temp 45°C), and 46kW of cold water (at 21°C) is generated by the adsorption chillers to cool 12 SuperMUC storage racks. The setup has a Coefficient of Performance (COP) of 18.38 if the heat radiation into air and the increased CMOS leakage at 45°C is not taken into consideration.

The complete system energy balance analysis enables a precise characterization of possible scenarios:

Scenario 1 (Adsorption chiller setup):

The complete setup has a COP of 11.3 at 45°C cooling inlet temperature (heat capture rate of HT-DLC cooling loop of 72%). With a yearly energy consumption of 1177 MWh and a kWh cost of 0.16€ per kWh this translates to 188312€.

Scenario 2 (LRZ traditional cooling winter operation):
With an inlet temp of 30°C and traditional cooling (chiller-less for the HT-DLC IT and chiller supported for the air cooled and indirect cooled storage racks) the COP is 7.95 (HT-DLC heat capture rate of 87%) . The yearly energy consumption is 1209 MWh, translating to 193511€.

Scenario 3 (LRZ traditional cooling summer operation):
With an inlet temp of 40°C and traditional cooling (chiller-less for the HT-DLC IT and chiller supported for the air cooled and indirect cooled storage racks) the COP is 7.40 (HT-DLC heat capture rate of 79%). The yearly energy consumption is 1238 MWh translating to 198090€.

Scenario 4 (fictive adsorption chiller setup with insulated racks):
Using a better rack design with a HT-DLC heat capture rate of 95% at 45°C inlet temperature the COP would be 16.5, the energy consumption 1136MWh, and the energy cost 181779€.

RoI is 19 years (Scenario 1 compared to Scenario 2).

RoI is 10 years (Scenario 1 compared to Scenario 3).

RoI is 6 years (Scenario 4 compared to Scenario 3).

Conclusion:

Using adsorption chillers to generate needed cold water can replace mechanical chillers. But the heat transfer into air at higher inlet temperatures needs to be reduced and the foot print of adsorption chillers per kW needs to be reduced to make the use of this technology viable for production deployment.

LRZ and Fahrenheit AG (the developer of the adsorption chillers used at LRZ) are working together to address this issues.

6 Session IV - Site Visit to CSCS

The tour of the facility took the participants through all three floors of the datacentre building – comprised of the supply deck (underground), the distribution deck (ground floor) and the datacentre (first floor). The building was planned from the start to reach a PUE of less than 1.25 and the infrastructure to be built out in a modular fashion in order to avoid locking in CAPEX for installations that were not yet needed. The building has a datacentre of 2000m^2 and can accommodate up to 25MW of which 12MW are currently installed. In order to better accommodate humans and machines the facility is divided into an office building and a datacentre building that are connected by an underground tunnel and a bridge on the first floor.

Supply deck (Underground)

The 16kVA electrical supply and the arrival of the lake water cooling pipeline enter the building at this point. The electrical supply comes from the substation that is located 25m from the datacentre, whilst the lake water comes from Lake Lugano that is located at 2.8km from the datacentre. The water is pumped from the lake at a depth of 45m and the underground pumping station, located in the city park, conveys the water to CSCS and also houses the two micro turbines that generate power from the free-falling return water. At 45m depth the lake has a virtually constant temperature of 6°C and on average the water gains 1C° on its journey to CSCS. The return water currently has an average temperature of 14°C and is fed back into the lake at a depth of 15m. A maximum return temperature of 25°C is allowable in order to avoid any adverse impact on the lake.

Inside the building the lake water runs through a series of heat exchangers before returning to the lake. The first set of heat exchangers will supply the cold temperature cooling loop, that accommodates the high-density compute systems that require low cooling temperatures. The return from this loop is fed through a second set of heat exchangers that supply the in-row cooling units with an inlet temperature of 19°C. This allows CSCS to use every litre of water pumped from the lake to cool two separate circuits. The facility also has the ability to add a high-temperature cooling loop and in the event of the lake temperature rising, space is foreseen to add mechanical cooling.

CSCS only provides UPS supply to essential equipment (storage, network, etc.). This allows the number of batteries to be kept low. The batteries are located on this floor. Cold water is stored in large reservoirs to allow for a 15minute ride-through in the event of a short power cut.

Installation deck (Ground Floor)

This floor is essentially the raised floor of CSCS. Within this 5.5m high space we find the 400V electrical distribution to the various PDUs as well as the cooling loops for the various systems. Depending on the requirements of the system, the cooling loop will be connected to the low or the medium temperature cooling distribution with a heat exchanger to separate the main loop from the systems loop. In order to keep power and water separate in the event of leak, the PDUs all stand on a raised walkway 1m above the ground.

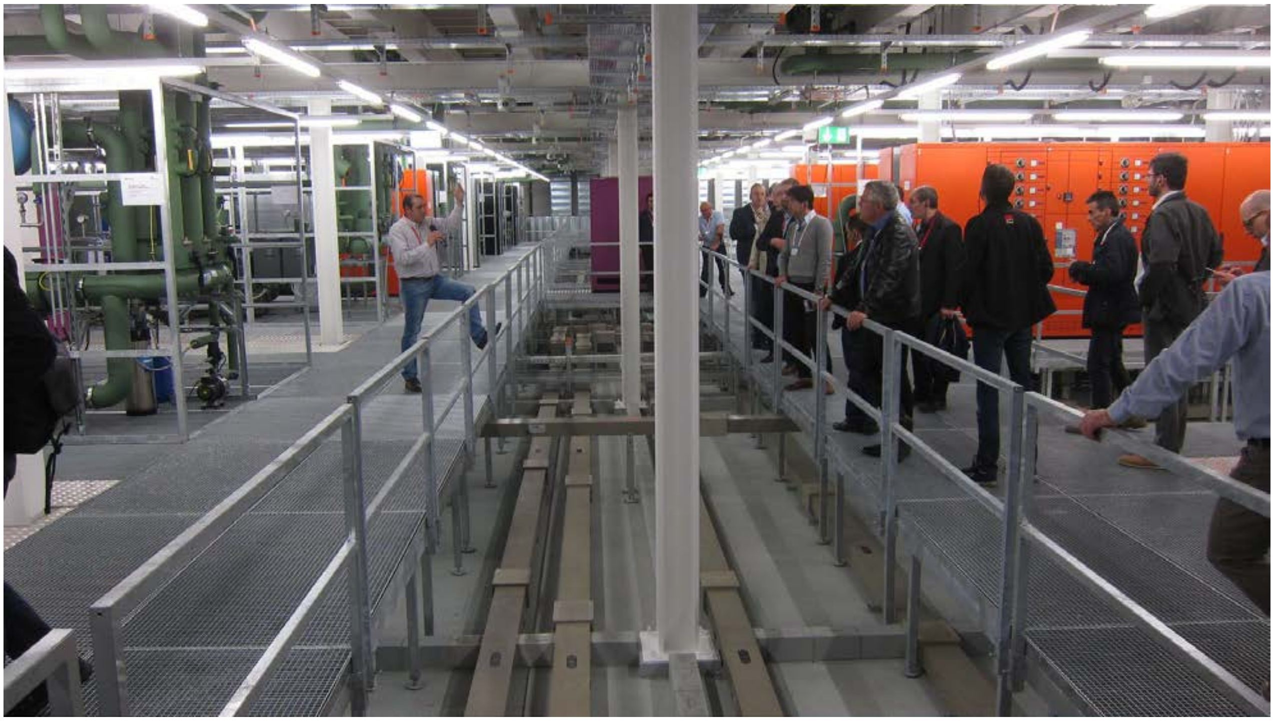


Figure 2: On the installation deck, i.e.: beneath the raised floor

The raised floor structure is based on a 3-tier construction of I-beams on which the very short pedestals sit that support the raised floor tiles. This allows the floor to support a point load of up to 7kN.

The UPS systems are located in the rooms adjacent to the supply deck.

Datacentre (First Floor)

The datacentre consists of 2000m^2 of contiguous raised floor space. 500m^2 are dedicated to hosting the current and future HPC User Lab systems side by side. A further 500^2 are dedicated to all central services such as storage and network and VMs. The last 500m^2 host a number of systems for 3rd party institutional customers.

7 Session V - Design and Procurement

7.1 An approach to control the sharp fluctuations of power consumption and heat load of the HPC system – Shoji Fumyjoshi, RIKEN

RIKEN Advanced Institute for Computational Sciences(AICS) was established on July 1, 2010 and is located in Kobe, Japan. Riken AICS has three main missions:

1. To manage the operations and enhancements of the K computer.
2. To promote collaborative projects with a focus on computational and compute sciences.

3. To plot and develop Japan’s strategy for computational science which includes defining the path to exascale computing called Flagship2020-project.

The K computer has been designed for general purpose computing. It is rich in memory and interconnect bandwidth, but it has no accelerators. On the facility side, conservative but promising technologies like cold water cooling were widely used.

Power and cooling details

Total power consumption for the site is 14-16MW. The base power load, between 3-5 MW, is produced by two gas turbine power generators of about 5MW each, working in an active/standby configuration. Gas turbines have a lower limit of 2MW per unit and an upper limit of 5MW per unit. They have a lead time of 1 hour until they can produce the power output and they also need a 10kW/second lead time for change in the power output. The efficiency of gas turbines gets better the higher the load. For flexibility, to accommodate power fluctuations and high load, an additional local power substation can deliver an additional 11-12MW with upper limit in 12.5MW.

Between September 2012 and May 2016 the average total power used was 14.7MW. Out of that, gas turbines generated 2,4-5MW which is around 17-35% of total power used. PUE values for same time frame were between 1,5 and 1.35.

Around 70% of the equipment is water cooled, and 30% is air cooled. The computer room is constructed to have the K computer in the upper floor and air handlers at the lower floor. The air circulation, from lower floor cool air by the air handlers rise to upper floor computer room and is then circulated back down to be cooled. Computer cabinets with liquid cooling have a heat exchanger that is combined with the air handlers, to use the outer water cooling loop to chillers and cooling towers.

The gas turbine power generators use a co-generation system to achieve higher energy efficiency by re-use of waste heat for cooling and heating. This co-generation system uses heat from the gas turbines so that 30% is used to produce electricity and 45% is used as heat for air conditioning and steam for absorption chillers for child water cooling. Both absorption and centrifugal(electric) chillers are used. Four absorption chillers produce cooling capacity of 1700 US Refrigeration ton (USRt), or 5.98 MW, and need 273kW of to do so. Two plus one centrifugal chillers produce cooling capacity of 1400+700 USRt, or 4.93+2.46MW, and need 901+389kW.

Absorption and electric chillers differ in lead time of activation, in power consumption and in energy efficiency. Absorption chillers are slow to activate and need 2 hours lead time (including 1 hour for generator to activate). Electric chillers can be activated in 10 minutes. But absorption chillers need far less power than electric chillers and energy efficiency is significantly better in absorption chiller. In terms of price per unit cooling capacity the difference between the two types of chillers is very small in this setup: Absorption chillers cost approximately 6M$/5.98MW and electric chillers 5M$/4.93MW.

Riken uses absorption chillers for the base load. Electric chillers are used to handle the fluctuation. They are activated on higher loads. As absorption chillers work more efficiently by richer steam, in the optimal case only one generator should be used near 100% to produce rich steam to two absorption chillers which both would then be enough to deliver child water without electric chillers. In the actual but non-optimal case, one generator in half load produces enough steam for only one absorption chiller to work and then an electric chiller is needed to get properly child water.

Power consumption and fluctuations

Typical power consumption history at site between 2014-2016 divides into large scale job mode and normal mode. Large scale job mode takes about 3 days per month, its job size is typically from 36865 to 82944 nodes and there is usually just one job running. In normal mode job, size varies between 1 to 36854 nodes. The average job size is around 500 nodes and number of jobs is around 100. The power consumption trend of the K computer dominates that of site total. In the normal mode the power consumption varies a little due to mixed power consumption profiles of all jobs. In large scale job mode the one running job can produce greater fluctuations of power consumption. During a large scale job run, large and rapid changes of power consumption changes, from 14 to 18MW and back to 14MW ,can be seen during just 2h30min. A rapid power consumption rise of 4MW raises concerns for power and cooling facility of the site.

Power consumption and cooling

Power consumption during idle times in large scale job mode can be handled with the power supplier (12.5MW) and one gas turbine generator of 5MW but the second power generator needs to be activated for peak power consumption during high load. Power consumption of the job can be roughly predicted by user’s application and the job history database.

Cooling in idle mode is handled with two absorption chillers and one electric chiller which is enough for around 17MW but for the additional peak load above 17MW two cooling options need to be evaluated based on the lead time. If peak load is sudden, the second electric chiller needs to be activated. If peak load can be expected in advance then a third absorption chiller can be used which has about 2h lead time.

If the gap between the idle and peak becomes huge, cooling supply might not be ready in time due to the lead time involved. There is a need for more dynamic cooling solution. One option for this is to use thermal storage for cooling. Thermal storage has a lead time of few minutes, it does not depend it’s efficiency against the output and can be used in minimal supply level. It also has about the same cost range as the electric chillers.

The plan for thermal storage tank consists of water capacity of 210 tons which has a charge time of 200 minutes and discharge time of 40 minutes for 2.9MW or 80 minutes for 1.4MW. The water temperature is between 8-18°C. Thermal storage cost is estimated to be around 2Million USD. Testing the thermal storage for dispersion revealed that after 185 hours from full charge an imperceptible drop in upper zone of the tank occurred. More realistic test cases are now in progress.

Other options to deal with rapid and large power fluctuations

For GPU/MIC based supercomputers rapid fluctuations are expected to be larger. They have a negative effect on energy efficiency. The approach to tackle the large and short power fluctuations is to use capacitor, a DCU unit in between AC power input and the rack PSU unit. This way the height of the peaks will be lower and the width of the fluctuation will be narrower. A capacitor is a faster and a more cost effective way to handle the fluctuations compared to battery type of solution.

A prototype unit is made in collaboration with Riken and Fuji Electric. This prototype replaces 3 of 9 PSU units from one rack to DCU unit with one capacitor each. The capability for these 3 DCU’s are 3kW times 150msec. Preliminary results indicate the height of the peak could be decreased but there was a small overshoot at the beginning of the peak which is now under investigation.

On summary the sharp fluctuation of power consumption is one of the major concerns for facility design and operation. Riken has chosen to handle cooling challenges with thermal storage tank and power fluctuations with capacitor unit.

The constructions have been finished in March 2017 and systems work well for simple tests. Tests with some realistic operation cases are ongoing.

7.2 An HPC experience with Design-Build – Steve Hammond, NREL

The National Renewable Energy Laboratory’s new Energy Systems Integration Facility, ESIF, was procured using design-build as strategy. It is a process of procurement currently not yet possible in Europe. Procurement experts are currently all trying to work this process into our procurement rules and construction contracts. New rules and regulations will make it possible for European sites to procure in this way too.

NREL needed office space, laboratories and a datacentre, all together. They used the design-build process to specify what they wanted. They did not come up with a design first and then tendered for a building company to implement the design. The budget was fixed at 140 M US$, which included the money for an HPC system which was procured separately and the procurement of the HPC system is not in the scope of this presentation.

Just some of the priorities of ESIF project on a large and heterogeneous wish list were:

- Safety in Design
- At least 13 laboratories with specific requirements, and an option for at least two more; some of the labs need to be suitable for handling hydrogen, some for a grid simulator, etc.
- 200 offices, with an option for 50 more
- A LEED platinum facility
- A datacentre capable of providing 2.5 MW on day one, and capable of growing to 10 MW
- The datacentre PUE should be 1.06 or lower

There were numerous site plan drawings.

NREL followed the best practices specified by the Design Build Institute of America(DBIA):

- Do a best value procurement – the budget is fixed and specified up front

- Do a two-phase solicitation – first ask for qualifications of parties potentially interested in doing this. Select a shot-list of the best qualified teams. This resulted in a short-list of three qualified teams. The second phase is to select a design
- All three teams got the full specifications and were asked to make a design
- All three teams were paid stipends for their designs. So NREL paid all three an amount of designs 200,000 US$, thereby owns all designs, not just the winning bid. Thus NREL could choose to integrate a design feature of another design into the selected one.
- During the process there were incentive fees for safety, for progress, and used was made of performance specifications

The first step was a Request for Qualifications (RFQ). In this step the participants needed to explain why they wanted the deal. Participants expertise was evaluated based on successful completion of design-build projects or projects with a similar acquisition approach within original schedule and budget. Also design and construction of research and laboratory facilities was evaluated as well as construction of energy efficient datacentres, preferably HPC. Under evaluation was also participant’s view of developing architectural image suitable for NREL.

The second step was a Requests for Proposals (RFP) which was project specific and had weighted technical criteria, and which was received by three teams selected out of the six that had submitted an RFQ. Interim interviews were conducted. Best value selection was done by evaluation of the bidders proposals against the priorities in the beginning. Weighted technical evaluation criteria was used instead of just cost. NREL had more scope than budget which resulted in competitions being focused on amount of scope that would be provided for the money available. It was a process of iteratively arriving at a prioritisation of items on our wish list.

The specifications were based on performance. What something must do and how it needs to perform, not what is must be. E.g. you specify that you need at least 1000m^2 of contiguous floor space with a raised floor with a minimum height, a minimum point load, etc. The Subcontractor formally substantiates that their design will meet performance requirements during various project phases. The owner “accepts” items as being compliant but does not “approve” submittals/designs/products etc. Ultimately the facility must meet performance requirements.

Award fees

An award fee must be large enough to motivate the design-build team and is typical feature of the design-build process. ESIF had 2.5M US$ (2.5% of total subcontractor value) as award fees. An award fee guarantees owner has a voice during design/constructions. The award fee earned by the subcontractor is based on subjective criteria by owner. There were monthly feedback sessions with design-build team.

The award fee programme awarded portions at six project milestones:

1. preliminary design - 20%
2. design development - 15%
3. construction documents - 15%
4. construction - 35%
5. warranty period - 5-10%

6. final completion - 5%

Only around 100,000 US$ were not used as award fee. Award fees are not always granted but may be passed on to next goals.

Benefits

Design-build approach to project delivery offers the owner advantages like singular responsibility. With both design and construction in the hands of a single entity, there is one point of responsibility for quality, cost and schedule adherence. Measurable benefits include reduction in schedule by approximately 33% and a reduction in cost of about 6%. These figures are based on experience throughout the design-build industry. This meant about 18 months from first site preparations to datacentre commissioning. In this project there were few change orders - less than 0.5%. Innovative design is encouraged by the process and some risks are shifted from the owner.

There were also some unanticipated benefits. Leverage insights and expertise from a variety of trades. The concepts of all three bidders were acquired with stipends. Opportunities and challenges arose during the process that allowed to revisit specifications. The design-bid-build may not have been easy to address. The result was better than anticipated. Design goal datacenter PUE of 1.06 or better and actual trailing twelve month PUE 1.038. Lab of the year 2014 and office space is energy efficient.

Disadvantages

The owner loses some control of the design process. Design is often managed through owner approval of design documents during performance. Typically the contractor is given flexibility in design.

Less competition. Not every company can put together an effective design-build team. The process involves best value approaches to solicitation development, evaluation and award, not always familiar to construction management personnel.

Contract management is more challenging. Contract administration overall requires more collaboration. The absence of effective collaboration may be where the growing pains of design-build are revealed.

7.3 Challenges and opportunities of “slab on grade” datacentre design – Jim Rogers, ORNL

Jim Rodgers points out that, in the history of ORNL as well the history of datacentres at large the raised floor is sort of a constant element of the datacentre. An early supercomputer CDC6600 (1965) was already installed on a raised floor contained 100 miles of wiring. Power was also under the floor. In the years 1990-2012 ORNL followed the traditional design path. ORNL has more than 100,000 ft^2 of traditional raised floor datacentre space. Circa 2002-2012 (most recent construction) space incorporated 24” or 36” raised floors and 150-250 lbs/ft^2.

The Cray XT3 Jaguar system required a large cut-out beneath the cabinet and forced air from below. Cray XK7(Titan) is room-neutral and pulls air in from above the floor. No need for traditional forced air delivered via plenum. Titan has 42F(5.5C) water supply temperature and perimeter CRACs.

In 2014 the mission needs for a new system, to be operational from 2017-2022, were clarified:

- No less than 5 times the performance of Titan.
- Support for power up to 20MW, to accommodate compute, cooling, networks, file systems and management systems.
- 15000ft^2 space, the floor was for dedicated use as single-tenant facility.
- Around 7000 refrigeration tons of additional cooling, improved operational efficiency with the goal to eliminate or reduce dependence on chillers.

ORNL went to "all the usual suspects", HPC chip providers and system integrators, to anticipate what their systems would look like in 2017. The anticipated 2017 packaging requirements were very high density: 50-100kW (or more) per rack. This required the majority of heat dissipation via liquid cooled solution as 30kW is assumed to be the upper limit for air-only. The rough order of magnitude estimate for non-standard WxHxD cabinet weights were as high as 4000kg. This exceeds the floor loading for traditional raised floors. Installation issues include that the rack has a pathway, capable of the point load requirements from the delivery point/loading dock to its home.

New facility site selection considerations

Existing shared facility had insufficient free space and insufficient electrical distribution. The floor loading capability in the existing site was insufficient. Retrofitting the existing facility was more expensive and more disruptive than upfitting of a new undeveloped space 11000ft^2. This allowed a separate electrical distribution system to 20MW and new warm-water mechanical plant. The floor would be "on-slab" (i.e.: no raised floor) with 17' to upper deck.

LIDAR, a light detection and ranging system that uses pulsed laser light to measure ranges, was used to build a high-fidelity map of the existing but unused facility. Exact positions of power cabling, water distribution, air handling ventilations and columns were identified.

The facility height is 5.2m and from floor to electrical cable tray 2.74m. The new system, "Summit", has all water, power and network cables on the top of the rack. Trip trays are needed for water which is close to the power cabling. It was not trivial to fit everything that was needed above the racks. All supporting beams, water loops, air handlers on the ceiling, electrical cables, LED lighting and network cables trays are located between 2.74m to 5.2m of room height.

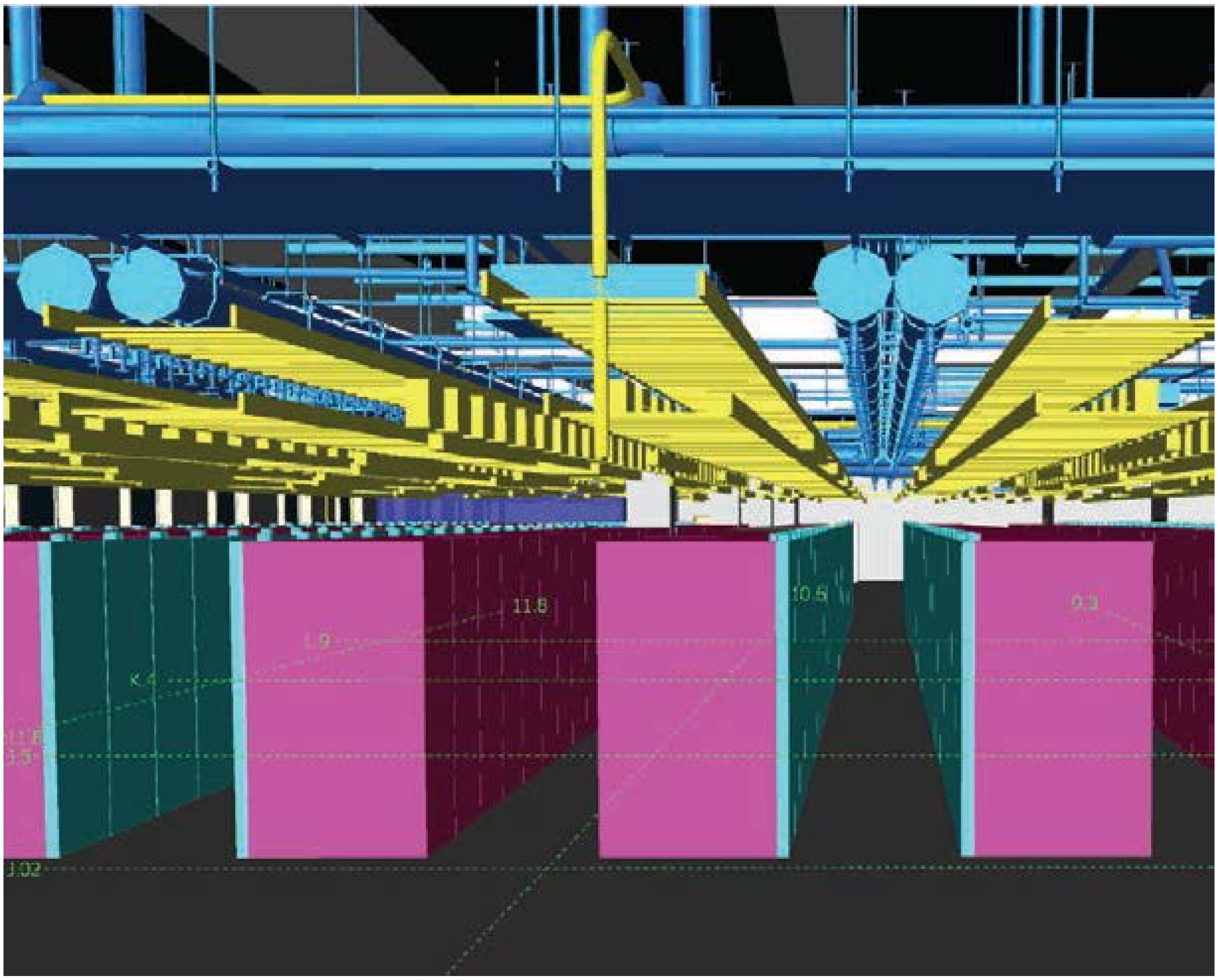


Figure 3: Summit design, with "everything delivered from above"

There is 48 tons cooling for 5300m^2. Summit has 256 compute racks. The racks have a standard size: 600mm Width x 1231mm Depth x 2020mm Height. The weight of a rack is 869kg and its distributed load is 1191kg/m^2. This is a lot less than anticipated in 2014. Airflow of 1100 CFM. Direct water-cooling of 68 liter/minute(18 gpm) at 20C. Power 480VAC of 3 phase circuit with inline 100k fault current protection needs two trays of power cables onto the racks. Cabling with infiniband fiber AOCs(36) and ethernet for management. Cabinets are set back to back to save piping on water cooling with rear door heat exchangers.

There is no light on the floor until maintenance.

Main switch boards(MSBs) are elevated 4" above slab surface on housekeeping pads for extra water protection. Up to 7 MSB's provide up to 20MW. MSB connect via 5000A main through an exterior wall to 3.0/4.0MVA transformer immediately outside the datacentre space. Lower material cost and lower line-loss save money.

An over cabinet cable tray of 24” will hold the individual branch circuits. Each circuit is protected in flex and terminated in a junction box that adds a 100000A series-rated fuse to protect against fault current.

Wet isle has 12”MT branch circuit that provides overhead delivery of water from the secondary loop system. A secondary loop uses Aquatherm with polypropylene pipes. 1” connections for supply/return to each cabinet. Water supply/return rows are separate from the electrical cable trays. Water system has been flooded in March 2017 with temporary loopback pipes in the place of the racks.

Some minor architectural modifications were necessary. Drainage and a hard epoxy floorcovering were added to the concrete slab. Incoming makeup air (HEPA filtered) ensures positive pressure.

In the discussion following his presentation, Jim Rodgers points out that he would have done things differently, if there had been an option to design a brand new facility from scratch. But tendering for a facility is complex, the time schedule for Summit would have become untenable. He would have opted for a ‘hybrid’ design that separates the mechanical from the electrical infrastructure and has them on different floors than the computer itself, similar to what the CSCS datacentre and the new NREL datacentre have.

8 Session VI - Energy Storage

8.1 Lead Acid Battery Life Cycle Management – Tiziano Belotti, CSCS

The CSCS installation

CSCS has ten 400kVA UPS systems with a total of 960 batteries and a floating voltage of 438V (min. 310V). A wireless measurement system has been installed, that allows each battery to be monitored separately in order to collect detailed data on their health. In order to ensure that the batteries are only replaced when this is really necessary, CSCS set out to monitor and test their batteries in great detail.

Battery monitoring system

The battery monitoring system provides data on Voltage, Temperature, Resistance, Current and Boost back (coup de fouet). A study by a CSCS intern showed, that the best way to measure the health of the batteries was to attach a load and observe the behaviour of the batteries under stress.

Norms for leadacid batteries

The main norm for lead acid batteries is EUROBATT that is formulated by the Association of European Automotive and Industrial Battery Manufacturers that all the main battery manufacturers are affiliated to. Batteries are classified according to their life cycle expectation:

- 3 – 5 years Standard Commercial
- 6 – 9 years General Purpose

- 10 – 12 years Long Life
- > 12 years Very Long Life

The EUROBATT norm refers to the IEC 60896-21/22 standard regarding testing methods and float voltage limits.

The 2011 battery procurement

The 2011 procurement document for the CSCS batteries specified nominal capacity, a life cycle of 10 – 12 years and a 5-year warranty with the mandatory requirement of 15 minutes of autonomy for a 400KVA/360KW load. The procurement did not however specify how this performance should be tested following the initial acceptance. The only way to test the system with the full specified load posed a risk for CSCS as it would have meant putting the production load on bypass for the duration of the test.

CSCS battery tests

Between 2012 and 2014 CSCS ran regular tests of their UPS installation that showed that 35 batteries were displaying low voltages. All of these were replaced by the supplier who also analysed 18 of these batteries in detail and reported incorrectly working valves due to assembly errors, short circuits caused during assembly and electrolyte filling faults. This analysis was of great concern to CSCS as all 960 batteries come from just two serial batches.

Thanks to the vendor providing reference tables for discharge currents at various loads, CSCS was able to test their battery park with the existing load. For this, the UPS entry is cut and the load runs on battery power until either the supplier-indicated test duration is reached or the voltage drops below 310V. The first set of tests in December 2014 had to be interrupted half-an hour prior to the indicated duration of two hours due to 62 batteries falling below 9.5V and a further 143 falling below 8V. Extensive discussions lead to the supplier running their own tests on their premises. These showed that most of the batteries still performed well at low load and long duration tests. However, all batteries achieved only 50% of their normal capacity for the critical 15-minute test at full load. The supplier suspected that this was due to micro cycles caused by the UPS.

The second set of tests (after replacement of 93 batteries) was run in April 2015. Once again, the test had to be cut short due to a total of 259 batteries falling below 9V. At this point the vendor started to look for excuses by questioning the validity of the battery monitoring system, the room temperature in the battery rooms and the presence of micro cycles caused by the UPS. The first two excuses could be disproven and CSCS agreed to replace the UPS filters to ensure the micro cycle problem would also be allayed.

In May 2016 the vendor decided to run their own tests on the CSCS facility that consisted of an initial boost charge of each battery to 14.4V (bad idea with old batteries) and subsequent connection of a 65A load to each battery branch. The test was set to run for 2 hours or until the first batteries fall below 4.5V. Once again, the test had to be interrupted early due to 337 batteries falling below 9V. Based on this test, the supplier has now agreed to replace further 192 batteries free of charge. All costs for the testing were carried by the vendor and the new batteries will be place in separate branches and not mixed with the older batteries.

Conclusion

After 5 years of operation the vendor has replaced 324 batteries free of charge. CSCS has gained vast experience on the topic of batteries that will allow them to write a much more precise procurement document next time round. Experience shows that a life expectancy of 10 – 12 years cannot be trusted as the tests confirm that even long life batteries have a life expectancy of 6 – 8 years. CSCS will continue to run regular tests to ascertain the exact point at which they need to run the next battery procurement.

8.2 Lithium Batteries – Morten Stoevering, Schneider Electric

Energy storage parameters

When selecting a high-power energy storage for a UPS system the important parameters are runtime, life expectancy, environment, performance and safety. The typical backup time required by datacentres is 2 – 15 minutes. A number of technology solutions cover this range, amongst which lead acid batteries and Lithium Ion batteries. This talk focused on the comparison of Lithium versus Valve Regulated LeadAcid batteries (VRLA).

Lead Acid batteries come in two types: VRLA or “Open Cells”. VRLA typically has a life span of 3 – 10 years and is the most common choice for UPS applications in the 1 – 2MW range and runtimes of up to a few hours. Extremely large power and runtime applications are still dominated by Open Cells where up to 20-year life spans can be achieved.

Lithium ion battery (Li-ion)

Since the year 2000 this has become the most important storage technology for portable and mobile applications. It shows high gravimetric energy density and has the potential for large cost reductions due to mass production. One of the main challenges to developing large-scale Li-ion batteries is the special packaging required to protect circuits against internal overcharge. These batteries are highly efficient but due to the thermally unstable nature of the metal oxide electrodes safety is a serious concern and for this reason several safety systems are built into these batteries at the cell level as well as the system level.

There are three main families of Lithium batteries (Li metal, Li Ion, Li Polymer) of which only the Li-ion family is compatible with stationary UPS application. The Li-ion family comprises 6 main chemistries for power applications. The Lithium Manganese Oxide (LMO) and Lithium Iron Phosphate (LFP) are the ones compatible with stationary UPS applications or high power and short to medium run time.

Safety

Each Li-ion battery cell comprises six safety mechanisms: a vent for thermal pressure release, an overcharge safety device, a multi-layered separator design, a safety functional layer anode design and an internal fuse to prevent external short circuits or overcharge.

The overall Li-ion UPS system further comprises safety mechanisms to measure temperature and voltage at the module level a current at the rack level. Fuses and MCCB provide current protection and the system will automatically disconnect at 75°C.

Li-ion solutions

With a 15-year design life a Li-ion battery system is typically used for backup between 5 and 20 minutes. It is built as an intelligent and self-protecting system. Each rack will contain its own switch gear and rack BMS as well as SMPS and system BMS alongside the battery cells that are located in modules.

Compared to a VRLA solution the Li-ion solution requires 60% less footprint, has 2-3 times the expected life, weighs 70% less, it provides 10 times the number of cycles. Although the initial CAPEX is double that of a VRLA solution, the TCO calculation shows savings of 10 – 40% over 10 years. Although the solution does not require cooling and can function at an expanded temperature range of 0 – 40°C higher temperatures will reduce the expected battery run time.

Conclusion

Temperature is key to both VRLA and Li-Ion batteries. Low temperatures ensure better life expectancy. Li-Ion solutions are particularly interesting if space is a concern and they provide in-built monitoring as well as requiring less frequent replacement and lower TCO.

8.3 Application of fuel cells for UPS – Felix Büchi, Paul Scherrer Institut

UPS: Power and energy

A UPS system provides backup in the event of a power outage and is located between the grid and the load. It consists of an energy storage unit and a conversion unit. In the case of a battery, the storage and conversion are coupled, whilst for example in a diesel generator the engine (conversion) and the fuel (energy) are decoupled.

Batteries provide low energy density and have no noise vibrations and emissions. They can be placed close to the consumer and have an average life expectancy of 5 years. Diesel generators have a high energy density but cause exhaust gas and heat and thus need to be located at a distance from the user and has a life expectancy of up to 20 years. Whilst a battery UPS system can run for the duration of the available stored energy, a diesel generator can be run for as long as fuel can be supplied.

In comparison, fuel cells provide medium energy density and storage and conversion are decoupled. They display no substantial noise or vibrations and zero emissions. The gas needs to be stored outside. Expected life cycle could be 10 years.

Fuel Cells: Principles and Properties

A Fuel Cell unit can be placed in lieu of the batteries inside a UPS system, whereby a limited battery capacity will be required to start the reaction in the event of a power failure. The fuel cell is fed from a gas storage that needs to be located separately and possible outside the building.

Although you can provide multiple MW of power with a battery system, the duration is limited by the amount of batteries you can store. For longer backup durations diesel generators are better suited. Fuel cells are mostly used in lower power ranges and for medium range backup durations today.

Fuel Cells are built like a battery – with two electrodes and an ionic separator. Whilst in a battery the materials transform to produce energy. The fuel cell on the other hand is fed with hydrogen from one side and oxygen from the other side and their interaction produces energy.

A fuel cell system will comprise an air inlet and a compressor that pushes this air though a humidifier before it reaches the fuel cell. Humidification is required to ensure conductivity. This provides the oxygen. The fuel cell is then connected to a supply of hydrogen. As both gases need to be provided in excess capacity, the system contains a feedback loop so excess gas can return and be reused. It is also possible to supply the oxygen in pure form from gas storage rather than from fresh air. The latter setup shows improved efficiency.

Fuel Cells for UPS: Application and Outlook

Since 2009 Fuel Cells have attracted increasing interest for use in backup power systems. PSI, the Hochschule Luzern, Swiss Hydrogen and the Swiss Confederation have developed a first fuel cell UPS system with a 17KW capacity and an autonomy of 13 hours with 12 bottles of H_2 and 6 bottles of O_2 at 200 bar. The Fuel Cells show very short start-up times.

To be of interest for HPC applications Fuel Cells would need to reach much higher power levels. Development in this direction could come from energy storage that will be necessary in order to stabilize the electrical grid and provide distributed energy storage due to the variability of renewable sources. The idea would be to produce and store H2 and O2 by means of Electrolysis when there is excess supply and then use these to power fuel cells and produce energy when there is insufficient supply.

PSI has a project to engineer such a solution at a 200kW level and gas storage of about 3MWh. The concept stores hydrogen and oxygen in a containerised solution. This concept has been developed with the company Swiss Hydrogen. Each stack has a 20% inbuilt redundancy.

It may be possible to combine energy storage and UPS applications. Swiss Hydrogen has developed a concept for up to 3MW.

Conclusion

Fuel cells are silent and compact energy converters and show good potential for UPS applications. Currently there is a lack of development in this area, but interest in energy storage may lead to developments that will benefit UPS applications.

9 Session VII – Technologies

9.1 Power Management framework: A consistent vision to manage your cluster – Abdelhafid Mazouz, Atos

Abdelhafid Mazouz points out that with the current hardware technology an exascale machine would need about 150 MW of power. However, the power constraint for exascale aimed at is 30 MW. A substantial part of the energy efficiency improvement has to come from the software side.

Idle compute nodes still consume between 30% and 50% of the energy that is used by the same node under full load. This is because only the CPUs go into a deep sleep state, but no other components. There are still a lot of “knobs” that are not utilized. To utilize them, a centralization of knobs is required. Atos proposes to monitor and control all relevant power management values and actions (at job and at cluster level) from a central framework that is useful for:

- Facility managers, who aim at high system utilization with low energy
- HPC users, who aim at high performance

The goals of these principal stakeholders are not the same. The goal for energy management should be to achieve performance/energy tradeoffs that go beyond traditional power capping and utilize all dynamic power management knobs that are available.

HPC energy policies must be defined that satisfy some quality of service:

- Job oriented policies, aiming to improve Flops/Watt
- Cluster-oriented policies, aiming at improving throughput/Watt

Strategies at the job-level are:

- To expose application dynamism (computation patterns)
- To adjust the execution environment to the application context

Strategies at the cluster-level are:

- Hardware consolidation
- Resource sharing: smart job co-location, co-scheduling

Atos is implementing a framework that consists of five modules, two of which are still under development:

1. Data collection: Monitoring and data collection, which is done through out-of-band querying via IPMI or high frequency using RAPL or specific Bull hardware called HDEEM.
2. Data Diagnosis: Store and display time-series of previous metrics, based on BEO (Bull Energy Optimizer)
3. Prediction: (still in development) Perform power/energy predictions for future jobs in the cluster

4. Anticipate: (still in development) Definition of policies and the actions to be done in order to follow them
5. Control: Component that can limit power consumption of parts of the cluster, trigger alerts or reactions following the information from the other components of the framework

At the end of the presentation a more technical and detailed information is shown of the actual design of BEO, showing the detailed hierarchical collection of metrics to a web application with a command line interface and a graphical user interface.

9.2 ABB SynRM motor & drive package –Super premium efficiency for HVAC application Huy Hour Lo, ABB

ABB is a well-known company that provides cooling compressors, pumps and other facility equipment. Huy Hour Lo in his presentation informed about a new type of motor, which introduces several advantages to normal induction motors, which are called low voltage synchronous reluctance motors (SynRM).

The main advantages of a SynRM motor compared to induction ones for HVAC (heating ventilation and air conditioning) applications can be summarized as:

- High energy efficiency
- High power density
- Low bearing temperatures, which has longer bearing lifetime
- Same maintenance as induction motors

A disadvantage is, that can only be used for variable speed drive (VSD).

Reluctance motors exist since 1888, but they have been used mainly for little motors and servomotors. The key difference in this design is in the rotor, which has four magnetic poles.

ABB also presented an analyses of the most common issues that a cause a motor to fail. As shown in Figure 4 below, with 51% bearing problems are the most common cause.

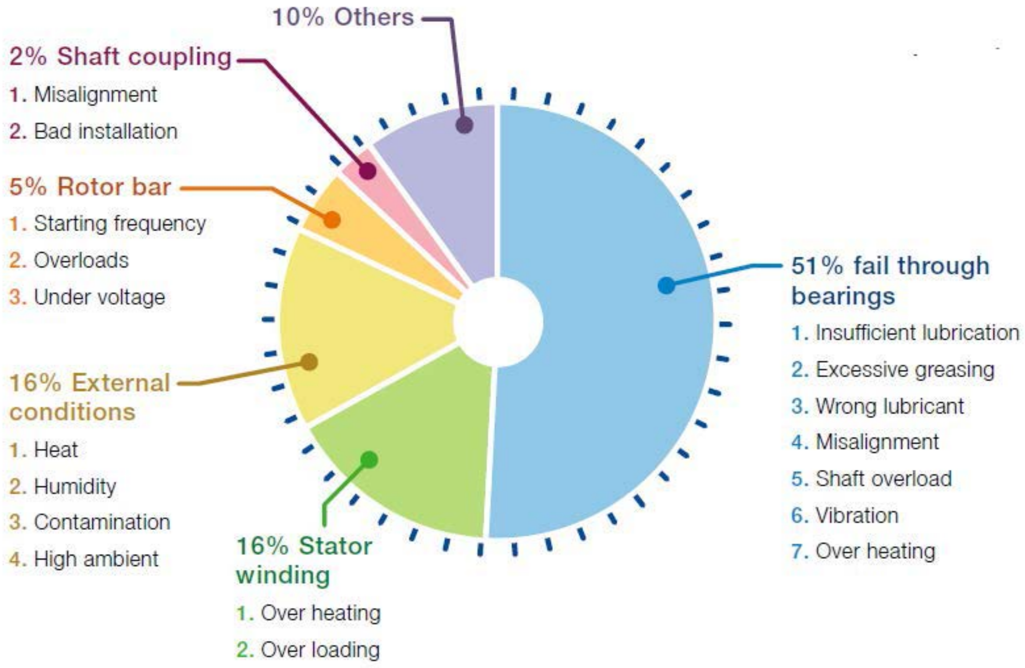


Figure 4: Causes of motor failure

9.3 ABB AbilityTM Smart Sensor - Tom Bertheau, ABB

Tom Bertheau presented an appliance developed by ABB, called SmartSensor, which tries to predict motor failures and tries to avoid the outages caused by such failures.

First, the different factors/conditions that can affect the lifetime of a motor were presented, such as, thermal, electrical, ambiental and mechanical. Though some scheduled maintenance can reduce failures, unpredicted stops can happen at any time. SmartSensor will collect information from the motor and try to predict failures before they occur.

The appliance presented needs a bluetooth connection and through the mobile phone of the service engineer information is uploaded to ABB private cloud. Information can later be visualized in a customer portal.

The process of data collection proved to be not scalable to a high number of motors. A new version is in development in which each SmartSensor sends information to a gateway which takes care to upload information from several devices.

During the question time, concerns were expressed due to the usage of a private cloud and the ownership of the data generated from the sensors.

9.4 Preventing biofilm in cooling installations with the use of ultrasound - Michael Widegren, Creando

Michael Widegren presented a product called Harsonic which tries to mitigate the formation of biofilm in water cooling systems using ultrasounds. Biofilm is a layer of bacteria protected by a coating (slime) which is not solvable. It is therefore difficult to remove. Biofilm can become attached to the hull of ships, but in a datacentre context it is the material that can clog water pipes, heat exchangers and filters.

Simple bacteria can create this biofilm. Open water systems, slowly flowing water of low temperature are typical conditions in which bateria become very productive “film producers”.

An ultrasound solution is proposed (emision of sound waves above 20KHz), which makes more difficult to adhere the bacterias to the surface, and also kills single cell organisms that burst apart due to the high frequency. Sound in water also creates cavitation bubbles that grow in successive cycles. The bubbles reach an unstable size and violently collapse. The collapsing cavitation bubbles bombard the organisms with jet streams of water. Small Barnacles and Cyprids are killed due to bursting cavitation bubbles. Also eggs up to 10 days old are thus killed.

Sound frequencies between 20 KHz and 65 KHz will make the bacteria feel quite uncomfortable. A peculiarity is that the bacteria adapt to the ultrasound. To stay effective, the frequency has to be changed regularly.

The solution is actually not a cleaning technology, but a preventive measure. If a system already has biofilm, it first needs to be cleaned. Once it is cleaned, Harsonic is very effective for keeping it clean.

10 Session VIII – Vendor Panel

10.1 Lenovo - Martin Hiegl

At Lenovo a research project in combining data from the Infrastructure Operational Technology and the Information Technology is underway. It was named as “IoT (Internet of Things) for the Datacentre”. Usually most DCIM systems today only use data from the operational technology. The collection of data is similar to any DCIM system but data is also collected from the IT technology. On top of that Lenovo is planning to build a data analytics stack that could be run locally or in the cloud. It was described that you could work on different levels like:

- Visualization – Descriptive
- Diagnostics
- Predictive
- Prescriptive

The analytics application could have modules for:

- Application Scheduling Optimisation

- Application Runtime Optimisation
- System Reliability Optimisation

Some use cases were mentioned:

- Job scheduling can be different during day or night depending on cooling conditions
- Different scheduling regimes on nodes connected to different cooling loops
- Prediction when a server is not behaving normally
- Controlling the IT system in heat re-use applications so that the outgoing cooling water is within certain limits. This can include taking this into account in the scheduling algorithm and setting the clock frequency of processors to appropriate values.

10.2 IBM – Ronald P. Luijten

Ronald P. Luijten from IBM’s Research in Zurich presented the ideas for a microDataCenter, which he defined as compute, storage, networking, power &cooling integrated into one ultra-compact form factor.

One of the main problems with today’s datacentres is the high power consumption. Combining this with some other facts like that the power usage today mainly comes from moving data (not computing) and that is takes more power to move data longer distances gives the conclusion that the only way to drastically decrease the power consumption is to build a much denser compute centre. To take this to an extreme IBM has prototypes an entire datacentre in a box with microservers and with a full Ethernet interconnect and storage.

The microservers in the microDataCentre are 64-bit server class computers with ECC etc. but at this point not aimed for HPC that requires high floating-point performance. There is also a micro 10 Gb/s Ethernet switch, which can be configured for different Ethernet topologies. With the density of this microDataCentre water-cooling is required. It is possible to build a microDataCentre with air-cooling but around 4 times less dense.

The microDataCentre contains some novel technology like the copper plates that are used both as heat sink and for distribution of power for the micro servers.

The microServers comes with PowerPC or ARM processors with the potential for new additions in the future.

Already today the servers achieve twice the number of operations per Joule as the energy-efficient Xeon E3-1230Lv3 and at the same time are 20 times denser. A novel measure was also introduced: the memory bandwidth density measured in GB/s/m^3. Of course the microDataCenter gives record numbers for this measure.

10.3 HPE - Frank Baetke

The topic of this presentation was the plans for an exascale system in the 2022-2023 time frame. Energy is the main problem of all exascale systems today and it was pointed out that it is not just the MW total that counts but also the picoW and picoJoule for each transaction on the lowest level.

HP has plans for memory-centric system called “the Machine”. Here are some required characteristics of such a system:

- Open – operating system and middleware need to be open
- Balanced between compute capability and capability to transfer data to the memory
- The programming model need to be memory centric
- Energy efficient.

In the current systems the memory bandwidth per flop is much too low. Also the bandwidth per pin on the chip was discussed and found too low. The presented solution to these problems was a new memory protocol called genZ, which is proposed by HP and designed by an industry consortium of all major manufactures of HPC systems and components with the exception of Intel. GenZ opens new possibilities for new memory architectures with higher bandwidth and lower energy consumption and flat common large address space for different types of memories (including disks, SSD, RAM memory, interconnects, etc.). This would mean that all memory and peripherals including memory attached to other nodes would be addressed in the same way. Memory of different forms can be attached to the CPUs or GPUs in a memory fabric. GenZ also allows for photonic components and serial connections. The flexibility of genZ would allow for many different memory architectures in the future. To build an exascale system with today’s technology would require an interconnect that would consume 10 MW only for the interconnect. That needs to be decreased dramatically before a system with true exascale performance can be built.

10.4 Vendor panel discussion: Lenovo - Martin Hiegl, IBM – Ronald P. Luijten, HPE - Frank Baetke, Intel - Mike Patterson.

It was stated that standardization is required also for HPC system of exaflop scale. Here we talk about weight, power, limits for cooling water etc. An important reason is that the vendors are expected to sell a few racks of an exascale system to smaller countries and research institutes that can not afford a full fledge exascale system. Standardization of the architecture was also hoped for so that a program developed for one vendor’s system could run efficiently also on system from other vendors. This is not true today even if all the system use MPI for communication.

For exascale systems air-cooling was not considered an option. Some kind of liquid cooling will be required as water, oil or special liquids (like 3M Novec) cooling.

Moore’s law was considered to be close to the end even if the development could continue for some more years. It can be debated if CMOS would end at 13 nm or 5 nm but at some point the granularity of the atoms will dominate. “Atoms are not scalable”, as IBM put it. IBM also pointed out that it is possible today to get a good performance without going to the very most advanced and latest semiconductor fabrication. A good architecture can beat fast semiconductors.

At some point the limits dictated by physics will kick in. We are talking about things like:

- The temperature at which silicon can operate at are limited

- The speed of light is limited which implies that dimensions of computer system need to be limited
- Atoms don’t scale

This calls for a highly dense liquid-cooled architecture. The difficulty is to find a good architectural balance between size, power and cooling.

Special design versus standardization was also discussed. A certain flexibility in the cooling specification may be needed to accommodate for different climates and use case like heat re-use or lake cooling, etc. It should also be noted that lower temperature of the cooling liquid would increase the performance of the system. But still, classes like ASHRAE w2 and w3 seems to fill a role. No chillers will be needed in the datacentre in the future since the water cooling will open the possibility for “free cooling” in most climates.

It seems that future system will distribute memory and compute capacity to avoid the situation of today with memory on one side and CPU’s on the other. To optimise the design it is important to measure and optimise the correct parameters on the datacentre level instead of for example on the CPU level.

Over-specified power supplies by the vendors were discussed and both vendors and datacentre operators agreed that this is a big problem. There may be good explanations for that for example that short power spikes may occur and the power supplies need to be able to handle that. Also, there is some variability in silicon and the power supplies need to be able to handle that as well. On the other hand, some of the over-specification is due to security factors on all levels up to the datacentre. A system perspective may help here so it would be possible to limit the power to a certain level.

It was discussed if water-cooling would be a commodity in the future and IBM thought the big size of a system was more of a problem than the actual water-cooling. IBM mentioned their plans for a water-cooled microDataCentre to run Watson Health in a doctor’s office.

11 PRACE Session

11.1 BSC (Spain)

If the time of this presentation, the current system of BSCW was MareNostrum 3 (IBM 1.1 Pflops – around 50000 cores), installedin December 2012. After its shutdown (March 2017), MareNostrum 3 will be divided into 8 sub-clusters (from 2 to 8 racks with spare parts) and transferred to different universities/research centers. They will be operated without maintenance contract and open (50%) to Spanish researchers.

The new system MareNostrum4 includes a “big” general purpose cluster, some smaller clusters and some storage. The contract, won by IBM (Lenovo, Fujitsu acting as subcontractors) includes the modification of the chapel needed for the installation of the new hardware. IBM is responsible for the overall operation. The total cost is 34 M€ and includes maintenance until July 2020.

The “big” general purpose, under installation, is provided by Lenovo. The production will start in July 2017. It includes 3456 nodes with 2 Intel Xeon processors interconnected by OPA. The total peak performance exceeds 11 Petaflop/s and the power consumption is less than 1.3 MW. Three smaller clusters are included in the contracts: one provided by Fujitsu and Lenovo based on Intel Many Cores processors (KNL and KNH, with more than 0.5PF), one provided by IBM based on Power9+NVidia (with more than 1.5PF), and one provided by Fujitsu based on Arm V8 processors (with more than 0.5 PF).

Regarding the new BSC-CNS headquarters, the construction work is progressing slowly due to funding problems. However, the closing of the building (facade construction) will finish April 2017 and tenders are being prepared for the next phase (office space and initial datacenter design).

11.2 CINECA (Italy)

CINECA has a long term roadmap for acquiring supercomputers and data storage equipment with a target of 50 Petaflop/s - 50 PB in 2019-2020.

The current supercomputer is Marconi that includes three different partitions: A1 based on Intel Xeon processors – Broadwell generation (installed in June 2016), A2 based on Intel Xeon Phi processors – KNL generation (installed in November 2016) and A3, not yet installed, based on Intel Xeon processors – Skylake generation (planned for July 2017). The interconnection network of Marconi uses the OmniPath Intel technology.

At the time the A3 partition is installed, half of the nodes of the A1 partition will be used to build a new system (Marconi-cloud). The main usage of this system will be for high energy physics and some no-traditional HPC users. It will use 25 Gbits/s Ethernet for the connection to the core switch of the center.

An important project in the Bologna area is the future installation of the ECMWF datacentre after the selection by ECMWF of this area as preferred location. The datacenter will be a building that was previously a tobacco manufacturing site. This building consists in seven halls of 3000 m2 each. Three halls will be used by ECMWF. CINECA is considering using one hall for its own usage since in the current datacenter CINECA is running out of space. A national institute on high energy physics is also considering using one hall.

11.3 PRACE PCP (Pre Commercial Procurement)

The goal of the PRACE PCP is to procure R&D focused on advances in energy efficiency in order to address a major challenge towards exascale systems. The result of the R&D (at least 80% must be performed in Europe) will be assessed through pilot systems scalable to 100 Petaflop/s.

The PRACE PCP involves five procurers (CINECA, CSC, EPCC, FZJ and GENCI) and PRACE aisbl as observer. The total budget is 9 M€ with an EC contribution of 50%. It is organized in three phases: solution design, prototype development and pre-commercial small scale production/service development including field test with pilot systems.

The PRACE PCP is currently in phase 3 with three suppliers: Bull, E4 and Maxeler. The pilot systems will be tested on benchmarks from the UEABS selected for their scalability, active development and portability to hybrid machines. The benchmarks have been executed and measured (including in term of energy usage) on various PRACE systems in order to define a reference point.

The list of pilot systems is the following:

- E4 provides a pilot system based on IBM Power8+ with NVlink connection to NVidia Tesla P100 with advanced features in terms of power monitoring, profiling, management, capping and prediction. The target performance in 1 Petaflop/s. It will be installed in CINECA (Italy).
- Maxeler provides a pilot system based on FPGA. Most of the budget is spent on application porting since Maxeler will port 4 codes to FPGA. It will be installed in FZJ (Germany).
- Bull provides a pilot system based on Intel Xeon Phi processors packaged in a Sequana cell. The target performance is around 0.5 Petaflop/s. It will include high frequency/high resolution energy monitoring using FPGA and smart energy management sub-system. It will be installed in CINES (France).

The pilot systems are available to PRACE (within PRACE-4IP) until the end of December 2017.

One benefit of the PRACE PCP worth noting is the definition of an energy-efficiency measurement methodology that may be useful beyond the PCP for procurements.

11.4 SURFsara (Netherlands)

SURFsara decided in 2014 to rent floor space in a commercial datacentre (TeleCityGroup, now Digital Realty datacentre) rather than building a new datacentre. In 2016, mostly from July to October, all the SURFsara IT equipment moved, including the Cartesius supercomputer. On the one hand, it was decided to reuse as much as possible the cables to reduce cost. On the other hand, to build a bridgehead at the new datacentre, while Cartesius was still running production at the old site, backbone cabling interconnecting several groups of racks was completely re-done, and partly had to be re-done because the system could not possible be housed with the same physical lay-out.

For Cartesius, Atos/Bull was responsible of deconstruction and reconstruction while SURFsara was responsible of the transport. Two “light weight” protocols were defined for the hand-over from Atos/Bull before transportation and to Atos/Bull after transportation. Shock-detectors were used in order to check that no damaged was caused by the transportation.

The move of all systems went on very well within the limits of the planned time frame. Only one major accident occurred: one container with 9 dual node blades fell down. This accident was covered by the insurance company. Some other shock detector were “in the red” on arrival as well, for unknown reasons. But after inspection, by both SURFsara and Atos/Bull, both parties agreed that these were false alarms, and the components involved were accepted by and taken into contractual maintenance again by Atos/Bull. Since the time schedule was the priority, workarounds had to be found during the installation when preconditions were late. At the end, this lead to a "repair/re-do" back log to be handled later. Similarly, a careful inspection made possible to discover mistakes that were done in the rush of the installation and to record in the "re-do" back log corrective actions when needed.

SURFsara is now a tenant of a datacentre "within a datacentre" with 2 dedicated floors (800 m2 total). The facility was not designed for HPC (the design is for air-cooled racks up to 22 kW), therefore one floor was modified in order to provide 1 MW of warm water cooling. At the end, the scarce resource is capacity for heat dissipation by air.

Since the installation of SURFsara IT equipment in the new location, some contractual vagueness had to be sorted out. This included unexpected additional billing for the connection of the new Sequana island, shortly after the move, which in the end was confirmed to be included in the normal billing since the electrical connection of IT equipment up to 1.5 MW was included. For the rest (including floor customization), the service catalogue of the datacentre had to be used.

In terms of "green(er)" IT ad RoI, SURFsara had to pay the adaptation of the infrastructure in order to make HPC best practices possible. The datacentre acknowledged that this lead to a lower overall PUE – at the end, SURFsara gets a 10% discount on every kWh used by a water cooled rack.

11.5 CEA (France)

CEA is a French organization involved in research and development. The computer complex of CEA, installed in the site of Bruyères-le-Châtel, close to Paris, is divided in two parts: the TERA part, for internal use, the TGCC part, open to external users.

CEA is currently installing a new supercomputer for the TERA part: TERA-1000. Using DLC technology, this system will provide up to 3.2 MW of warm water (45°C). This system is cooled by dry adiabatic cooling towers (no chillers).

The installation of this new supercomputer takes place at the same time than a major modernization of the heating system of the CEA site of Bruyères-le-Châtel (more than 2000 employees) so it was decided to reuse the heat of TERA-1000 for heating the site. This needed heat pumps to be installed in order to raise the temperature to the temperature of the heating system of the site. It is expected that next year 82% of the heat needed for the site will be provided by TERA-1000, the remaining part by gas-fired heating system. The latter will be needed in case of very low temperature (below 0°C) and when the supercomputer is stopped.

In this context, TERA, will be in terms of datacentre Performance (DCp) ranked in class B.

11.6 GRNET (Greece)

GRNET (The Greek Research and Technology Network) manages four datacentres in Greece: two in Athens, one in the north-west of Greece (Louros), and one in Crete (Knossos). The aggregated capacity is 135 racks (6 computer rooms) and 1.6 MW maximum power for IT. Currently more than 1800 servers are installed with 4 PB of disks and 7 PB of tapes.

The main HPC resource is located in the main building of the Ministry of Education. It hosts ARIS (phase 1 and phase 2). This system provides a variety of nodes (thin nodes, fat nodes, GPU nodes, Phi nodes) interconnected by IB FDR. The total peak performance is 444 Teraflop/s. The usage of the system is high, except for the Intel Xeon Phi nodes for which the adoption by users is slower.

The Knossos datacenter is mostly dedicated to cloud computing. It hosts 20 racks (possibility to extend to the double of this number) with a total power usage of 190 kW.

The Louros datacenter is the most advanced one in terms of "green IT". It uses the electricity produced by a dam nearby and, for cooling, the water for a river (0,1% of the river flow – maximum temperature 15.5°C). The management of the datacenter is done remotely, including the monitoring and control of the cooling system. The PUE is expected to be around 1.18 at 100% IT load.

Since the start of operation, several issues had to be addressed:

- Filter clogging, still not fully understood.False activation of fire extinguisher because of the environment where the datacenter is installed (dust from outside).
- Low water level due to hydro-electric power plant shutdown.

11.7 WCNS (Poland)

WCNS (Wroclaw Centre for Networking and Supercomputing) belongs to the Wroclaw University of Science and Technology. It was established in 1994.

The IT infrastructure is installed in two different building: D21 (new), D2 (old). The new facility offers 725 m^2 of computer room and 350 m^2 of technical and office area.

The power supply includes two diesel generators for a total capacity of 2600 kVA (fuel tank for 7 hours) and 6 UPS for a total capacity of 1580 kVA. The batteries can bridge a time interval of 10 minutes. The cooling, with a total capacity of 1612 kW is provided by cold water cooling (with free cooling) and air cooling.

The Bem cluster is installed in building D21. It is based on Xeon nodes (Haswell) for a total of 22656 cores with IB-FRD interconnect and a performance of 860 Teraflop/s. This system uses 375 kW peak.

The storage includes:

- Lustre parallel file system (DDN-1000 HDD) for a capacity of 1.1 PB.
- NFS/CIFS/HTTP storage.
- SAN storage for a capacity of 1.2 PB.
- Archive / Backup (2 robotics with LTO6 drives) for a capacity of 5 PB.

The Bem cluster runs PBSPro and serves 607 active users.

The monitoring system of the datacentre monitors power, cooling, server hardware, software, network and the visibility of systems.

12 Main Findings

The presentations given during the 8th European Workshop on HPC Infrastructures, and the plenary discussions following these presentations:

- Reveal important trends in management, design, and procurement methodologies for energy efficient HPC datacentres
- Provide an insight into new developments on the HPC system integration side
- Give hints to assess the situation in Europe with regard to these domains

12.1 Trends

Developments on the HPC system integrator side are important to understand and anticipate, as much as possible, as they will have an impact on the datacentres that in the near future will have to accommodate such systems.

There is a consensus on the vendor' side that systems still can, and have to, get denser to be more energy efficient. Power usage today comes more from moving data than from computing and power usage is definitely reduced if data has to travel shorter distances. The trend towards liquid cooling is sustained, air-cooling is not considered an option for dense exascale systems. But as to what sort of liquid cooling it will have to be, and particularly on the desirable inlet temperature of the coolant, there is now somewhat more debate than in recent years. The energy efficiency argument from the facility side tends to drive towards warmer water cooling as higher outlet temperatures tend to increase their options for productive heat reuse. That has not changed. System integrators and chip builders go along with that, but now, more markedly than before, point out that there is a trade-off: their systems surely 'can take the heat', but the systems generally would perform better and have extended longevity when driven at a somewhat lower temperature.

There is an unequivocal consensus that there is no need for traditional chillers in modern datacentres. Water cooling, whether warm or 'luke warm', opens possibilities for 'free cooling' in most climates and feasible production ready alternatives to compression cooling exist. As Mike Patterson (Intel) succinctly put it: 'Design for the lowest temperature you can still run without the need of chillers'.

Air cooling is not going away for auxiliary 'non-compute' that usually accompanies HPC systems, such as storage servers and disk arrays. Perhaps this is an area that, although less power intensive, needs more scrutiny. An eye-opener that occurred in the presentation on oil immersion cooling, is that up to 20% of the power consumed by air-cooled equipment can be used to drive the fans inside the IT equipment. And since the fans are inside the IT-equipment, this power, used to move air rather than data, does not even show up as "cooling overhead" in metrics like the PUE.

Denser systems result in heavier racks. And if these racks are all direct liquid cooled, they can do without cold air pushing up from a plenum under a raised floor. Jim Rodgers of ORNL has shown how it can be done without a raised floor. Summit is accommodated 'on a concrete slab'. But Jim Rodgers points out that, when you have a green field procurement for a new facility, a hybrid design, with electrical and mechanical utilities on a separate floor – like the CSCS and NREL datacentres – is the preferred option. If everything has to be 'delivered from above', the routing of power be cabling, network wires, and pipes all above the racks becomes non-trivial. That has a negative impact on the ease of maintenance.

With current state of the art energy efficient hardware, an exascale system would need power in the order of 150 MW. By most stakeholders in the field that is not considered an option. There are initiatives on the IT side to let the software do its complementary bit to close the gap currently left on the infrastructure operations side, like Intel's Global Energy Optimisation initiative, presented at last year's workshop, and the power management framework presented by Atos at this workshop. The strategies proposed for the software side are similar. On the one hand a data on the behaviour of applications are collected and analysed: what patterns in variable energy need during application runs are discernible? On the other hand they try to bring far more "knobs" to control the hardware and its power usage under control of the application and/or the job scheduling and resource allocating system. CPUs now have manipulatable states that are directy related to their energy consumption, ranging from 'turbo modes' to sleeps states that vary in depth. But what about the other components, such as memory and I/O devices, including network devices?

The assumption of power management frameworks is that facility manager is striving for maximum system throughput at the lowest power consumption possible. That is not an unreasonable starting point but site updates like the one from RIKEN show that trying to optimize power usage on smaller timeintervals will bring its own problems and does not necessarily imply an improvement energy efficency at the site level. A lot depends on the local constraints on power provisioning. The RIKEN site for example, runs most energy efficent if there is a continuous high load on their gas turbines. Frequent changes in power demand with high amplitude and steep slopes, adversely effect their energy efficiency.

Backup power provisioning when the default power source fails, is important at least for parts of HPC installations. Several new promissin battery and fuel cell technologies for locally storing energy, are under development. Sometimes cost, sometimes maturity could be reasons to not yet adopt them. On the other hand, detailed investigations of conventional lead acid batteries revealed that their life expectancy is generally substantantially overrated by vendor specifications. Even so-called long life batteries after 6 years failed to sustain the delivery of the minimally required voltage long enough, when put to the test.

12.2 Current situation in Europe

To a high degree facility managers of HPC centres in Europe face the same challenges as their collegues and peers elsewhere on the globe. The annual workshop is an important place for sharing best pratices and new findings among HPC centres, and notably the HPC centres of PRACE partners. The workshop contributes to sustaining a high level of expertise and to the spread and implementation of advanced technologies in European HPC centres.

There is a continued strong involvement of European HPC sites in the work of the the EEHPCWG (Energy Efficient HPC Working Group), a US Department of Energy sponsored intiative which, now for the fifth time, was present at the workshop to provide an annual update on their work and underline its interest in keeping a close connection with the workshop.

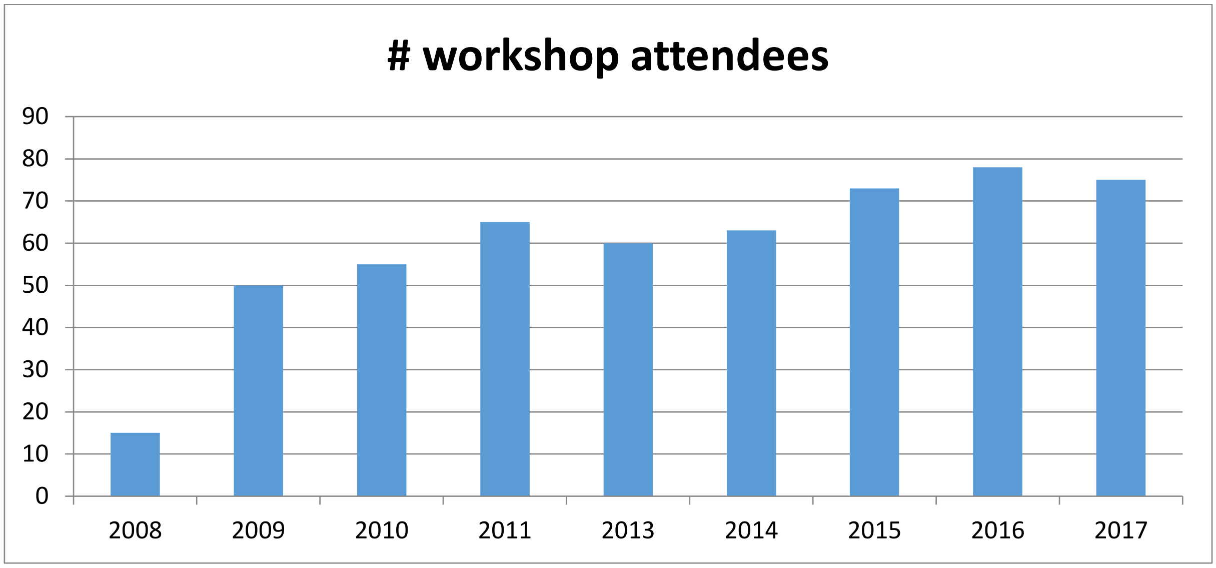
Decisions of datacentres to choose one solution over another are usually driven by cost versus benefit analyses and return on investment considerations. What all the site updates show, is that local context matters tremendously and the local circumstances can differ considerably, even within Europe. The price of electricity differs considerable from site to site and this will obviously have consequences for the outcome of return on investment calculations. The opportunities for providing cooling capacity differ too. European HPC centres are getting ready to accommodate future generations of supercomputers. HPC facility management teams in Europa must be, and generally are, specialized and capable of adopting new technologies that best fit their local context. As the e.g. the presentation of CSCS on lifecycle management of lead acid batteries shows, it is highly benificial if they are capable of conducting detailed monitoring and research, indepent of the vendor, on technolgies they have deployed.

The design-build procurement for a new datacentre, presented at this workshop by a major US HPC datacentre, resembles the competitive dialog procurement process that many sites have succesfully applied to tender for an new HPC machine. European sites however cannot yet, but in the future will be able to, procure in that way for datacentres. It is worth further investigation whether this form of procurement fits the risk assessments and customization choices that need to be made for the long term investment of a new HPC datacentre.

The design-build procurement process that was presented at this workshop, in several ways seems a better fit with the needs of facility managers and the risk assessments they have to make. The process also makes it easier to use all intellectual effort that has been put into the procurement process, not just the effort that was put in by the party of the selected bid.

13 Conclusions

Like its predecessors, the 8th HPC Infrastructures workshop has been very successful in bringing together experts from various disciplines and in various stakeholder roles working on HPC site infrastructures. Though this year there were two unfortunate short term cancellations, Figure 4 shows that the annual workshop, which is attended upon invitation only, has become an institute that is capable of consistently attracting a stable number of experts in the field.



.

Figure 5: Number of workshop participants in a historical perspective

This year's workshop did not focus on common theme that was addressed by all or most presentations, but it addressed a variety of major concerns for the management of HPC infrastructures. Energy efficiency and cooling technologies are of course recurring subjects of prime importance. New but important topics on the agenda were technologies for energy storage and quality control of such devices, and methodologies and best practices for procurement for a new HPC facility.

Regarding trends it is specifically worth noting that:

- Higher density and weight of racks is an unavoidable consequence of striving for more energy efficient of parallel computing with high bandwidth and low latency communications between computing elements.
- Despite the fact that racks are getting denser and heavier, the raised floor is *not* on its way out of HPC facilities. If anything, it should be more raised, rather than replaced by a concrete slab. In a greenfield procurement for a new datacentre a hybrid design with electrical and mechanical utilities on a separate floor are likely to be worth the investment in the long run, as they substantially ease maintenance – and hence maintenance cost – and the transition to new systems.
- Air cooling of high density racks is not an option, but what sort of liquid cooling should be used and at what temperature is not straightforward. Running at the highest temperature possible given the equipment tends the ease the re-use of waste heat. For optimal performance and longevity of the IT equipment, designing for the lowest temperature at which you can run still run with free cooling may be a better rule of thumb.
- New energy storage technologies that are currently under development are likely to make their way into the power provisioning of HPC datacentres when their production scales up and their production costs drop.