



**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

D5.4

HPC Infrastructures Workshop #9
Final

Version: 1.0
Author(s): Huub Stoffers, SURFsara
Date: 26.11.2018

Project and Deliverable Information Sheet

| | | |
|---|--|--|
| PRACE Project | Project Ref. №: EINFRA-730913 | |
| | Project Title: PRACE Fifth Implementation Phase Project | |
| | Project Web Site: http://www.prace-project.eu | |
| | Deliverable ID: < D5.4 > | |
| | Deliverable Nature: < Report > | |
| | Dissemination Level: PU* | Contractual Date of Delivery: 30 / 04 / 2019 |
| | | Actual Date of Delivery: 30 / 11 / 2018 |
| EC Project Officer: Leonardo Flores Añover | | |

* - The dissemination levels 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 #9 | |
| | ID: D5.4 | |
| | Version: <1.0> | Status: <i>Final</i> |
| | Available at: http://www.prace-project.eu | |
| | Software Tool: Microsoft Word 2013 | |
| | File(s): PRACE-5IP-D5.4.docx | |
| Authorship | Written by: | Huub Stoffers, SURFsara |
| | Contributors: | Javier Bartolomé Rodríguez (BSC), Radek Januszewski (PSNC), Andreas Johansson (NSC, Linköping University), François Robin (CEA), Susanna Salminen (CSC, IT Center for Science Ltd.), Filip Stanek (IT4Innovations, VSB-TUO), Gert Svensson (KTH) |
| | Reviewed by: | Kadir Diri, UHEM Veronica Teodor, JUELICH |
| | Approved by: | MB/TB |

Document Status Sheet

| Version | Date | Status | Comments |
|---------|------------|--------|---|
| 0.1 | 04/09/2018 | Draft | Based on latest (09.02.2018) PRACE 5IP template downloaded from BSCW, included several summaries |
| 0.2 | 26/09/2018 | Draft | Corrections, still several summaries missing, layout with illustrations and tables, 1 st draft of introduction, main findings, and conclusions |

| Version | Date | Status | Comments |
|---------|------------|--------|--|
| 0.3 | 30/10/2018 | Draft | Processed feedback from original presenters that responded, finalisation of lay-out, addition of last missing summary. |
| 0.4 | 02/11/2018 | Draft | Minor corrections, finalisation of main findings and conclusions. Version for WP5 internal review |
| 0.5 | 12/11/2018 | Draft | Processed corrections and feedback from Gert Svensson and François Robin, WP5 internal review. Version for PRACE internal (PMO) review |
| 0.6 | 22/11/2018 | Draft | Processed corrections and feedback from Veronica Teodor |
| 1.0 | 26/11/2018 | Final | Processed corrections and feedback from Kadir Diri. |

Document Keywords

| | |
|------------------|---|
| Keywords: | PRACE, HPC, Research Infrastructure, HPC facility, datacentre, Energy efficiency, Cooling, Power management, Power provisioning |
|------------------|---|

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

© 2018 PRACE Consortium Partners. All rights reserved. This document is a project document of the PRACE project. All contents are reserved by default and may not be disclosed to third parties without the written consent of the PRACE partners, except as mandated by the European Commission contract EINFRA-730913 for reviewing and dissemination purposes.

All trademarks and other rights on third party products mentioned in this document are acknowledged as own by the respective holders.

Table of Contents

| | |
|---|-------------|
| Document Control Sheet..... | i |
| Document Status Sheet | i |
| Document Keywords | iii |
| List of Figures | vii |
| List of Tables..... | viii |
| References and Applicable Documents | viii |
| List of Acronyms and Abbreviations..... | ix |
| List of Project Partner Acronyms..... | xii |
| Executive Summary | 1 |
| 1 Introduction..... | 1 |
| 2 Programme, content, and speakers | 2 |
| 3 Session I – Site Updates | 5 |
| 3.1 SuperMUC-NG – LRZ’s next step towards an improved waste-heat recovery – Herbert Huber, LRZ | 5 |
| 3.1.1 <i>SuperMUC-NG Phase 1</i> | <i>5</i> |
| 3.1.2 <i>High-frequency power usage measurement and energy aware scheduling</i> | <i>6</i> |
| 3.1.3 <i>Quantification of benefits</i> | <i>6</i> |
| 3.2 Surprises and challenges with stainless steel welding – Tiziano Belotti, CSCS | 7 |
| 3.3 The ENI Green Datacentre in its early years – Luca Bortot, ENI..... | 9 |
| 3.3.1 <i>A datacentre designed to be a prototype being in production</i> | <i>9</i> |
| 3.3.2 <i>Example: The discovery of a fan “going against the intended current”</i> | <i>10</i> |
| 3.3.3 <i>Monitoring and data collection: “never trust a single source”</i> | <i>11</i> |
| 4 Session II – Standards and Regulations..... | 12 |
| 4.1 Tips, Tricks, and Trends in Liquid Cooling: The State of the Practice – Mike Patterson, Intel..... | 12 |
| 4.1.1 <i>Liquid versus air cooling, a balanced view.....</i> | <i>12</i> |
| 4.1.2 <i>Intel activity pertaining to liquid cooling.....</i> | <i>12</i> |
| 4.1.3 <i>ASHRAE Updates.....</i> | <i>13</i> |
| 4.1.4 <i>Immersion cooling.....</i> | <i>13</i> |
| 4.1.5 <i>System Design Guidance topics</i> | <i>13</i> |
| 4.2 Open Standards for Datacentre Availability – Lex Coors, Interxion..... | 13 |

| | | |
|---------|--|----|
| 4.3 | Update from the Energy Efficient HPC Working Group (EE HPC WG) – Natalie Bates, LLNL | 14 |
| 5 | Session III - Energy to Solution and TCO | 15 |
| 5.1 | Presentations of the results from the three pilot systems of the PRACE-3IP Pre-Commercial Procurement..... | 15 |
| 5.1.1 | CINES, prototype “Frioul” – Jerome Chapelle & Eric Boyer | 15 |
| 5.1.1.1 | PCP goals | 15 |
| 5.1.1.2 | “Frioul” | 16 |
| 5.1.2 | CINECA, prototype “D.A.V.I.D.E.” – Carlo Cavazzoni | 17 |
| 5.1.3 | JSC, prototype “JUMAX” – Dirk Pleiter..... | 19 |
| 5.2 | TCO model for HPC procurement assessment – Eric Boyer, GENCI..... | 23 |
| 6 | Session IV – Bologna Science Park..... | 24 |
| 6.1 | Bologna Science Park: an infrastructure for HPC – Gaetano Moron, INFN/CNAF | 24 |
| 7 | Session V – Heat re-use..... | 25 |
| 7.1 | Making HPC Climate Positive – Eric Rylander, Stockholm Data Parks at Stockholm Exergi..... | 25 |
| 7.2 | The World’s first Climate Positive Datacentre – Lars Schedin, EcoDataCenter | 27 |
| 8 | Session VI – Energy grids and Power provisioning..... | 28 |
| 8.1 | Management of Energy costs at PSNC – Radoslaw Januszewski, PSNC..... | 28 |
| 8.2 | Advanced Power Measurement and Control of the trinity Supercomputer – Andrew Younge, SNL | 29 |
| 8.2.1 | The Trinity system | 29 |
| 8.2.2 | Investigations on Trinity - towards a Power API for measurement and control | 30 |
| 8.2.3 | HPC “mini-apps” | 31 |
| 8.2.4 | Lessons learned, guidance for HPC application power usage profiling | 31 |
| 8.3 | Public (WTO) procurement of electricity: the CSCS approach – Ladina Gilly, CSCS | 32 |
| 8.3.1 | Background of CSCS energy procurement and the Swiss energy market..... | 32 |
| 8.3.2 | Public procurement based on price | 33 |
| 8.3.3 | Public procurement based on a price formula..... | 33 |
| 8.3.4 | Conclusions | 34 |
| 9 | Session VII – Technologies | 34 |
| 9.1 | Direct Liquid Cooling of Datacentre Equipment – Dominik Dziarczykowski, CoolIT | 34 |

| | | |
|-------------|--|-----------|
| 9.1.1 | <i>About CoolIT</i> | 34 |
| 9.1.2 | <i>Datacentre challenges and Direct Liquid Cooling (DLC) Solutions</i> | 35 |
| 9.1.3 | <i>Switching to DLC</i> | 35 |
| 9.1.3.1 | <i>Assessing phase</i> | 35 |
| 9.1.3.2 | <i>Planning phase</i> | 35 |
| 9.1.3.3 | <i>Deploying phase</i> | 36 |
| 9.1.3.4 | <i>Maintenance phase</i> | 36 |
| 9.1.4 | <i>Case studies</i> | 36 |
| 9.2 | Modular Datacentres – Jean-Marc Ducos & Frédéric Souques, CEA | 37 |
| 9.3 | Facility Commissioning Experiences for ORNL’s Summit, The U.S. DOE’s 200PF Supercomputer – Jim Rogers, ORNL | 39 |
| 9.3.1 | <i>Summit compared to Titan</i> | 39 |
| 9.3.2 | <i>Progress of preparations for Summit</i> | 39 |
| 9.3.3 | <i>“12-15 MW, say what?!” - Predicting impact on the local power grid</i> | 40 |
| 9.3.4 | <i>Problems found during preparation and testing</i> | 40 |
| 9.3.4.1 | <i>High fault currents in power distribution units</i> | 40 |
| 9.3.4.2 | <i>Failing fan blades</i> | 41 |
| 9.3.4.3 | <i>Nearly failing circuit breakers</i> | 41 |
| 9.3.5 | <i>Mechanical Facility</i> | 41 |
| 9.3.6 | <i>Challenging GPFS scalability</i> | 42 |
| 10 | Session VIII – Vendor Panel | 42 |
| 10.1 | A flexible approach for Exascale infrastructure providing best electrical and cooling efficiency for all types of nodes and interconnect fabrics – Jean-Phillipe Sibers, Atos/Bull 42 | |
| 10.2 | Preparing datacentres for Exascale – Vincent Pel, Cray | 43 |
| 10.3 | The Future of HPC at HPE – Torsten Wilde – HPE | 44 |
| 10.4 | Liquid Immersion Cooling: A versatile cooling environment – John Wagner, Fujitsu 45 | |
| 10.5 | Behind the Scenes of Efficient Cooling – Axel Auweter, Megware | 46 |
| 10.6 | Water cooling technology for the Exascale computing era – Karsten Kutzner, Lenovo | 47 |
| 10.7 | Vendor Panel Discussion | 48 |
| 11 | PRACE session | 51 |
| 11.1 | PPI4HPC Update | 51 |

| | | |
|--------|--------------------------------------|----|
| 11.2 | SURFsara (Netherlands) | 52 |
| 11.3 | NSC-LiU (Sweden) | 53 |
| 11.4 | IT4Innovation (Czech Republic) | 54 |
| 12 | Main findings | 55 |
| 12.1 | Trends | 55 |
| 12.1.1 | Hosting | 55 |
| 12.1.2 | Exascale or exaFlop/s? | 56 |
| 12.1.3 | Electricity | 56 |
| 12.1.4 | Cooling | 57 |
| 12.1.5 | Monitoring | 57 |
| 12.2 | Current situation in Europe | 58 |
| 13 | Conclusion | 58 |

List of Figures

| | | |
|------------|---|----|
| Figure 1: | 9 th European Workshop on HPC Centre Infrastructures, Bologna, Italy, 1-4 May 2018 | 2 |
| Figure 2: | Workshop session | 4 |
| Figure 3: | SuperMUC-NG | 5 |
| Figure 4: | Welds from the original construction | 7 |
| Figure 5: | State of the art welding, before (left), and after (right), polishing | 8 |
| Figure 6: | ENI Green datacentre, Ferrera Erbognone, Pavia, Italy | 9 |
| Figure 7: | Cold air (blue arrows) pushed up from the plenum below the computer rooms, and the – intended - flow of hot air (red arrows). | 10 |
| Figure 8: | Power provisioning and power usage measurement for specific components of Frioul | 16 |
| Figure 9: | Architecture of the D.A.V.I.D.E. prototype | 17 |
| Figure 10: | Memory organisation of a system with a MAX5 accelerator | 19 |
| Figure 11: | Results of the value Profiling Library used for the SpecFEM3D application | 20 |
| Figure 12: | Interconnectivity between the MAX5 cards and other elements of the system | 21 |
| Figure 13: | Heat recovery in Stockholm | 25 |
| Figure 14: | Containerised datacentre site with network, power, cooling, and waste heat disposal connections | 26 |
| Figure 15: | Electricity, generation by source in 2014 and 1994 | 27 |
| Figure 16: | Impression of a Mobull containerised modular datacentre | 38 |
| Figure 17: | Detection of high fault current PDUs | 40 |
| Figure 18: | Almost failing circuit breaker | 41 |
| Figure 19: | Stacked immersion bath in a rack, sharing a single CDU | 45 |
| Figure 20: | Megware CPU heatsink design with layered honeycomb pattern | 47 |
| Figure 21: | Atos/Bull Quantum Learning Machine | 53 |
| Figure 22: | 864 Solar panels, contributing up to 255 kW of power to the Linköping datacentre | 54 |
| Figure 23: | Number of workshop attendees in historical perspective | 59 |

List of Tables

| | |
|--|----|
| Table 1: Technical data of the D.A.V.I.D.E. prototype | 18 |
| Table 2: Time-to-solution (TTS) and Energy-to-solution (ETS) results for the D.A.V.I.D.E. prototype..... | 18 |
| Table 3: Estimated maximum power consumption of node components | 22 |
| Table 4: Time-to-solution (TTS) and Energy-to-solution results for the JUMAX prototype..... | 22 |
| Table 5: HPC measurement taxonomy..... | 30 |

References and Applicable Documents

- [1] <http://www.prace-project.eu>
- [2] <https://www.euhpcinfrastructureworkshop.org>
- [3] <https://www.thegreengrid.org>
- [4] <https://www.thegreengrid.org/osda-tool/dashboard.html>
- [5] <https://eehpcwg.llnl.gov>
- [6] <https://docs.it4i.cz/anselm/hardware-overview/>
- [7] <https://docs.it4i.cz/salomon/hardware-overview/>

List of Acronyms and Abbreviations

| | |
|---------|--|
| AI | Artificial Intelligence |
| aisbl | Association International Sans But Lucratif (legal form of the PRACE-RI) |
| ASC | US programme for Advanced Simulation and Computing |
| ASHRAE | American Society of Heating, Refrigerating and Air-Conditioning Engineers |
| ATS | Advanced Technology System, of the US ASC programme |
| AVX | Advanced Vector Extensions |
| BCO | Benchmark Code Owner |
| Btu | British thermal unit, non-metric unit of heat. 1 kWh equals about 3412.14 Btu |
| CAPEX | Capital expenditure |
| CDU | Coolant Distribution Unit |
| CMOS | Complementary Metal-Oxide Semi-conductor, integrated circuit technology |
| CoE | Centre of Excellence |
| CPU | Central Processing Unit |
| CRAC | Computer Room Air Conditioner |
| CUDA | Compute Unified Device Architecture (NVIDIA) |
| DARPA | Defense Advanced Research Projects Agency |
| DEISA | Distributed European Infrastructure for Supercomputing Applications, EU project by leading national HPC centres |
| DFE | Dataflow engine |
| DIMM | Dual In-line Memory Module |
| DIN | Deutsches Institut für Normung, German national organisation for standardisation, and German ISO member body |
| DLC | Direct Liquid Cooling |
| DoA | Description of Action (formerly known as DoW, “Description of Work”) |
| DOE | Department of Energy, USA |
| DTU | Denmark Technical University |
| EC | European Commission |
| ECMWF | European Centre for Medium-Range Weather Forecasts |
| EEHPCWG | Energy Efficient HPC Working Group |
| EESI | European Exascale Software Initiative |
| EFlop/s | Exa ($= 10^{18}$) Floating point operations (usually in 64-bit, i.e. double precision) per second, also EF/s |
| EoI | Expression of Interest |
| ESFRI | European Strategy Forum on Research Infrastructures |
| ETAP | Electrical Transient and Analysis Program, software program for electrical engineers |
| FoM | Figure of Merit |
| FPGA | Field Programmable Gate Array |
| GB | Giga ($= 2^{30} \sim 10^9$) 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 |
| GDC | ENI Green datacentre |
| 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. double precision) per second, also GF/s |
| GHz | Giga (= 10^9) Hertz, frequency = 10^9 periods or clock cycles per second |
| GPFS | General Parallel File System |
| GPGPU | General Purpose GPU, a graphics processing unit used for general purpose computation rather than for image rendering computation |
| GPU | Graphics 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 |
| HPDA | High Performance Data Analytics |
| HPL | High Performance LINPACK |
| IFNF | Istituto Nazionale di Fisica Nucleare; Italian national institute for nuclear physics |
| IFNF-CNAF | The main data processing and computer technology research centre of IFNF |
| ISC | International Supercomputing Conference; European equivalent to the US based SCxx conference. Held annually in Germany |
| ISO | International Organisation for Standardisation |
| kA | Kilo (= 2^{10}) Amperes |
| KB | Kilo (= $2^{10} \sim 10^3$) Bytes (= 8 bits), also Kbyte |
| kN | Kilo (= 2^{10}) Newton |
| kW | Kilo (= 2^{10}) Watt |
| kWh | Kilo (= 2^{10}) Watt hour |
| LANL | Los Alamos National Laboratory, USA |
| LIC | Liquid Immersion Cooling |
| LINPACK | Software library for Linear Algebra |
| LLNL | Lawrence Livermore National Laboratory, USA |
| LTL | Low temperature loop |
| MB | Management Board (highest decision making body of the project) |
| MB | Mega (= $2^{20} \sim 10^6$) Bytes (= 8 bits), also MByte |
| MB/s | Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s |
| MFlop/s | Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. double precision) per second, also MF/s |
| MOOC | Massively Open Online Course |
| MoU | Memorandum of Understanding. |
| MPI | Message Passing Interface |
| MTL | Medium temperature loop |
| MW | Mega (= 10^6) Watt |
| MWh | Mega (= 10^6) Watt hour |
| NDA | Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement. |
| OEM | Original Equipment Manufacturer |
| OPEX | Operational expenditure |
| ORNL | Oak Ridge National Laboratory, USA |
| OSDA | Open Standard for Datacentre Availability |

| | |
|-----------|--|
| PA | Preparatory Access (to PRACE resources) |
| PATC | PRACE Advanced Training Centres |
| PCB | Printed Circuit Board |
| PCP | See PRACE-PCP |
| PDU | Power Distribution Unit |
| PFlop/s | Peta ($= 10^{15}$) Floating-point operations (usually in 64-bit, i.e. double precision) per second, also PF/s |
| PPI4HPC | European program for Public Procurement of Innovative Solutions for HPC |
| PRACE | Partnership for Advanced Computing in Europe; Project Acronym |
| PRACE-5IP | Fifth PRACE Implementation Phase project |
| PRACE 2 | The second phase of the PRACE Research Infrastructure following the initial five year period |
| PRACE-PCP | PRACE Pre-Commercial Procurement programme |
| PRIDE | Project Information and Dissemination Event |
| PSU | Power Supply Unit |
| PUE | Power Usage Effectiveness |
| RDHX | Rear Door Heat Exchange |
| RI | Research Infrastructure |
| RfP | Request for Proposals, formal public procurement document |
| SLA | Service Level Agreement |
| SLURM | Simple Linux Utility for Resource Management, a batch system |
| SNL | Sandia National Laboratory, USA |
| SoC | System on Chip |
| TB | Technical Board (group of Work Package leaders) |
| TB | Tera ($= 2^{40} \sim 10^{12}$) 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. double precision) per second, also TF/s |
| TGCC | Très Grand Centre de calcul du CEA; Datacentre of CEA for hosting petascale supercomputers for scientific HPC and Big Data. |
| 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 Applications Benchmark Suite |
| UNICORE | Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources |
| UPS | Uninterruptible Power Supply |
| VESDA | Very Early Smoke Detection Apparatus |
| VR | Voltage regulator |
| W | Watt |
| WP5 | Work Package 5 of PRACE-5IP, HPC Commissioning and Prototyping (also WP5-5IP) |

List of Project Partner Acronyms

| | |
|-------------------|--|
| BADW-LRZ | Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Germany (3 rd Party to GCS) |
| BILKENT | Bilkent University, Turkey (3 rd Party to UYBHM) |
| BSC | Barcelona Supercomputing Centre - Centro Nacional de Supercomputacion, Spain |
| CaSToRC | Computation-based Science and Technology Research Centre, 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, (3 rd 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 (3 rd Party to BSC) |
| CYFRONET | Academic Computing Centre CYFRONET AGH, Poland (3 rd Party to PNSC) |
| EPCC | EPCC at The University of Edinburgh, UK |
| ETH Zürich (CSCS) | Eidgenössische Technische Hochschule Zürich – CSCS, Switzerland |
| FIS | Faculty of Information Studies, Slovenia (3 rd Party to ULFME) |
| GCS | Gauss Centre for Supercomputing e.V., Germany |
| GENCI | Grand Equipement National de Calcul Intensif, France |
| GRNET | Greek Research and Technology Network, Greece |
| INRIA | Institut National de Recherche en Informatique et Automatique, France (3 rd Party to GENCI) |
| IST | Instituto Superior Técnico, Portugal (3 rd Party to UC-LCA) |
| IT4Innovations | IT4Innovations National supercomputing centre at VŠB-Technical University of Ostrava, Czech Republic |
| IUCC | Inter-University Computation Centre, Israel |
| JUELICH | Forschungszentrum Juelich GmbH, Germany |
| KIFÜ (NIIFI) | Governmental Information Technology Development Agency, Hungary |
| KTH | Royal Institute of Technology, Sweden (3 rd Party to SNIC) |
| LiU | Linköping University, Sweden (3 rd Party to SNIC) |
| NCSA | National Centre for Supercomputing Applications, Bulgaria |
| NTNU | The Norwegian University of Science and Technology, Norway (3 rd 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 Centre, Poland |
| RISCSW | RISC Software GmbH |
| RZG | Max Planck Gesellschaft zur Förderung der Wissenschaften e.V., Germany (3 rd 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 (3 rd 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, Laboratório de Computação Avançada, Portugal |
| UCPH | Københavns Universitet, Denmark |
| UHEM | Istanbul Technical University, Ayazaga Campus, Turkey |
| UiO | University of Oslo, Norway (3 rd Party to SIGMA) |
| ULFME | Univerza v Ljubljani, Slovenia |
| UmU | Umea University, Sweden (3 rd Party to SNIC) |
| UnivEvora | Universidade de Évora, Portugal (3 rd Party to UC-LCA) |
| UPC | Universitat Politècnica de Catalunya, Spain (3 rd Party to BSC) |
| UPM/CeSViMa | Madrid Supercomputing and Visualization Centre, Spain (3 rd Party to BSC) |
| USTUTT-HLRS | Universitaet Stuttgart – HLRS, Germany (3 rd Party to GCS) |
| WCNS | Politechnika Wroclawska, Poland (3 rd Party to PNSC) |

Executive Summary

The 9th European Workshop on HPC Centre Infrastructures [2] was hosted by CINECA and held in Hotel Savoia Regency, in Bologna, Italy, from 1 – 4 May 2018. For this workshop, BSC, CEA, CSCS, LRZ, PDC-KTH, and PSNC have invited CINECA to join the committee that organised and assembled the programme for the workshop.

The workshop, upon invitation only, was very successful, with 83 participants, coming from Europe, America, and Australia.

The workshop covered a broad range of topics relevant for HPC Centre Infrastructure management: standards and regulation, energy efficiency strategies, including heat-reuse, total cost of ownership reduction strategies, energy procurement strategies, power provisioning for large infrastructures and its impact on the grid, characteristics, and infrastructural requirements of exascale systems to be expected in the coming years. The workshop brought together experts from the vendor side and experts from the HPC datacentre facility management side. The intention to collectively learn from problems encountered during datacentre operations was fully honoured. Several presentations from datacentre sites gave insightful and unabashed looks into the “kitchen” of their facility management, problems encountered, and the struggles and various solutions tried to overcome these.

The PRACE closed session, held at the end of the workshop, gathered attendees from PRACE Tier-0 and Tier-1 sites. Several site representatives gave an update on specific datacentre infrastructure developments and there was an update on the European programme for Public Procurement of Innovative Solutions for HPC (PPI4HPC) in which four PRACE Tier-0 partners participate. The PRACE closed session gave the opportunity for exchanges between experts from the assembled PRACE sites.

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

1 Introduction

The 9th [European Workshop on HPC Centre Infrastructures](#) [2] was hosted by CINECA and held in Hotel Savoia Regency, in Bologna, Italy, from 1–4 May 2018. BSC, CEA, CSCS, LRZ, PDC-KTH, and PSNC have invited CINECA in the committee organising and assembling the programme for this workshop. Some PRACE-5IP Work Package 5 (WP5) manpower, as well as PRACE sponsorship, was used for the preparation of the workshop. The programme committee consisted of the following members:

- Javier Bartholomé Rodriguez, BSC, Spain
- Titziano Belotti, CSCS, Switzerland
- Carlo Cavazzoni, CINECA, Italy (hosting site representative)
- Ladina Gilly, CSCS, Switzerland
- Herbert Huber, LRZ, Germany

- Norbert Meyer, PSNC, Poland
- Jean-Philippe Nominé, CEA, France
- Michael Ott, LRZ, Germany
- François Robin, CEA, France
- Gert Svensson, PDC-KTH, Sweden

The workshop, upon invitation only, was very successful, with 83 participants, coming from Europe, America, and Australia. Among participating parties, there were HPC system integrators, PRACE-related datacentres, large government funded non-European HPC sites, commercial datacentre sites, and companies aiming to provide greenfield locations for the development of large datacentre “real-estate”.

The proceedings of the workshop, the facts and insights presented during its sessions, are summarised in this document, so as to be accessible to a wider audience of HPC facility management and datacentre management experts. The sites represented are often early adopters of or experimenters with new technologies and procedures. Their evaluations and accounts of problems encountered and solutions experimented are valuable for professional peers.

A chapter highlighting the main trends and important conclusions reached during the workshop is included at the end of this document.



Figure 1: 9th European Workshop on HPC Centre Infrastructures, Bologna, Italy, 1-4 May 2018

2 Programme, content, and speakers

Tuesday, 1 May 2018

- Registration
- Welcome dinner

Wednesday, 2 May 2018

- Session I: Site Updates, chaired by Carlo Cavazzoni, CINECA
 - SuperMUC-NG – LRZ’s next step towards an improved heat recovery – Hebert Huber, LRZ
 - Surprises and challenges in a relatively new datacentre – Titziano Belotti, CSCS
 - ENI Green Datacentre – Luca Bortot, ENI
- Session II: Standards and regulations, chaired by Ladina Gilly, CSCS
 - Tips, tricks, and trends in Liquid cooling, the state of the Practice – Mike Patterson, Intel
 - Update from the Energy Efficient HPC Working Group (EE HPC WG) – Natalie Bates, LLNL
- Session III: Energy to Solution and TCO, chaired by François Robin, CEA
 - Presentation of the results from the 3 pilot systems of the PRACE pre-commercial procurement:
 - CINES – Jerome Chapelle and Eric Boyer
 - CINECA – Carlo Cavazzoni
 - JSC – Dirk Pleiter
 - TCO model for HPC procurement assessment – Eric Boyer, GENCI
- Session IV: Bologna Science Park, chaired by Carlo Cavazzoni, CINECA
 - Bologna Science Park: an infrastructure for HPC – Gaetano Maron, INFN/CNAF
- Social Event and dinner

Thursday, 3 May 2018

- Session V: Heat re-use, chaired by Gert Svensson, PDC-KTH
 - Making HPC Climate Positive – Erik Rylander, Fortum
 - The World’s first Climate Positive Datacentre – Lars Schedin, EcoDataCenter
- Session VI: Energy grids and power provisioning, chaired by Herbert Huber, LZJ
 - Management of Energy costs at PSNC – Radoslaw Januszewski, PSNC
 - Advanced Power Measurement and Control of the Trinity Supercomputer – Andrew Younge, SNL
 - Public (WTO) procurement of electricity, the CSCS approach – Ladina Gilly (CSCS)
- Session VII: Technologies, chaired by Norbert Meyer, PSNC
 - Direct Liquid Cooling of Datacentre Equipment – Dominik Dziarczykowski, CoolIT
 - Modular Datacentres – Jean-Marc Ducos and Frédéric Souques, CEA
 - Facility Commissioning Experiences for ORNL’s Summit, the U.S. DOE’s 200PF Supercomputer – Jim Rogers, ORNL
- Session VIII: Vendor Panel, chaired by Ladina Gilly, CSCS and Herbert Huber, LRZ
 - A flexible approach for exascale infrastructure providing best electrical and cooling efficiency for all types of nodes and interconnect fabrics – Jean-Philippe Sibers, Atos/Bull
 - Preparing datacentres for exascale – Vincent Pel, Cray

- The Future of HPC at HPE – Torsten Wilde, HPE
- Liquid Immersion Cooling – John Wagner, Fujitsu
- Behind the Scenes of Efficient Cooling – Axel Auweter, Megware
- Water cooling technology for the exascale computing era – Karsten Kutzner, Lenovo
- Vendor Panel discussion – Ladina Gilly (moderator)
- Wrap-up, invitation to next year's Infrastructure Workshop to be held in Poznan, Poland
- Dinner

Friday, 4 May 2018

- PRACE closed session and PRACE-5IP WP5 meeting, chaired by François Robin, CEA and Volker Weinberg, LRZ
 - PPI4HPC overview
 - SURFsara (Netherlands)
 - NSC-LiU (Sweden)
 - IT4Innovation (Czech Republic)
 - PRACE-5IP WP5 Meeting



Figure 2: Workshop session

3 Session I – Site Updates

3.1 SuperMUC-NG – LRZ's next step towards an improved waste-heat recovery – Herbert Huber, LRZ

While the LRZ datacentre in Garching is primarily a production facility, from the outset it also has been an environment for controlled experiments that seek to make the compute services provided, and HPC in particular, more energy efficient. The strategy has several aspects, some pertaining to the applications running on an HPC platform, some to the HPC platform itself, and some, such as finding novel ways to use the waste-heat produced by HPC systems, to the datacentre itself.

Since 2011 LRZ is combining the use of direct warm water cooling with the productive reuse of system waste heat at the datacentre itself. LRZ has experimented with adsorption chillers in smaller “CoolMUC” systems. The LRZ history in terms of systems is shown below. The new 2018 system figures as the next step, in which the waste heat of a multi PFlop/s HPC “flagship” system is processed by an adsorption chiller.

- 2011: CoolMUC-1 (178 compute nodes, MEGWARE) + SorTech Adsorption Chiller
- 2012: SuperMUC Phase 1 (9216 compute nodes, IBM, 3.2 PFlop/s)
- 2015: SuperMUC Phase 2 (3072 compute nodes, Lenovo, 3.6 PFlop/s)
- 2015: CoolMUC-2 (384 compute nodes, Lenovo, 466 TFlop/s) + SorTech Adsorption Chiller
- 2017: CoolMUC-3 (148 compute nodes, MEGWARE, 394 TFlop/s)
- 2018: SuperMUC-NG Phase 1 (6480 compute nodes, Intel/Lenovo, 26.9 PFlop/s) + Fahrenheit Adsorption Chiller

3.1.1 SuperMUC-NG Phase 1

SuperMUC-NG Phase 1 (Figure 3) will have 6480 nodes in total that are fairly homogeneous: each node has 2 sockets with Intel SkyLake CPUs, 48 physical cores per node. The system will use an Intel Omni-Path interconnect and has an island structure. There are no nodes with GPUs or other accelerators. There is however differentiation of the nodes with respect to memory. One island will comprise 144 “fat” nodes with 768 GB memory per node. The other 6336 - “thin” - nodes are of the same configuration, except that they contain 96 GB memory per node. It would have been possible to install Apache non-volatile DIMMs in the nodes, but LRZ will refrain from that, as it would go at the expense of memory bandwidth.



Figure 3: SuperMUC-NG

There are no fans in the compute node blades, only in the PSUs are air-cooled and use fans. Water with 45°C - 50°C inlet temperatures circulates through tubes across the 2-node blades, cooling CPUs, memory, voltage regulation devices, and IO devices. With 45°C inlet temperature, the CPUs

still operate at a temperature lower than in an air-cooled system with an air inlet temperature of 20°C. LRZ is however not totally satisfied with the still air-cooled power supplies that are placed at the back of the SuperMUC-NG blades. Rather than fully optimising a direct liquid cooled version, Lenovo have cut some of their design cost and standardised their production by using exactly the same layout as for an air-cooled version of the blade. This decreases the efficiency of the warm water-cooling loop to 75%. In order to ensure that Phase 1 can be operated without the use of mechanical refrigerators, hot aisle containments with in row coolers are implemented and Fahrenheit adsorption chillers are connected to the warm water cooling loop in order to reuse the waste heat of the system to produce cold water, needed to operate the in row coolers. LRZ will demand Phase 2 of SuperMUC-NG to be 100% warm water-cooled.

3.1.2 High-frequency power usage measurement and energy aware scheduling

The blades have a high - 100 Hz - sampling rate of the power usage of different components. The total power consumption of the node, and the effect of changing processor frequency can be monitored per MPI call or code section. The high frequency monitoring data make it possible to study which parts of the application are memory-bound and which parts are CPU bound and to tune for that. Only CPU bound jobs or code sections should be executed at the maximum frequency. For example, studying the power usage of the HPL benchmark on these CPUs with a nominal frequency of 2.7 GHz and a “turbo” peak of 3.7 GHz, reveals that, when AVX instructions are used the application can run optimally at a 2.3 - 2.7 GHz frequency. Higher CPU frequencies are only beneficial for non-AVX instructions.

3.1.3 Quantification of benefits

The value for LRZ of direct warm water cooling and energy-aware scheduling implemented with SuperMUC-NG, is estimated to be about 30% savings in operation cost:

- Reduced server power consumption: Lower processor power consumption (~ 5%) + No fan per node (~ 4%)
- Reducing power consumption for cooling: At 40°C free cooling all year long (> 15%)
- Energy-aware scheduling: Frequency optimisation at run time, only CPU bound jobs get max frequency (~ 5%)

In addition, the noise level in the datacentre is significantly reduced and there are some CAPEX savings in that less conventional chillers are needed.

Adding an adsorption cooler to use the waste heat of the complete system adds at least an extra 5% of cost savings by reducing the power consumed for cooling purposes. This mechanism needs both, hot water of about 50°C on one side, and water coming from a cooling tower on the roof, bringing in water with a temperature up to 25°C to cool the adsorbent, on the other side. Under these conditions the adsorption cooler has an efficiency of about 0.6. For LRZ to be entirely room neutral, this infrastructure needs to produce chilled water with a cooling capacity of 600 kW. Currently this is attained most of the year, but not the whole year through. On hot days, the temperature of the water coming from the cooling tower on the roof goes up, and the efficiency of the adsorption cooling infrastructure goes down rapidly. On days on which the efficiency goes below 0.2 the cooling capacity of the amount of chilled water that can be produced is less than the needed 600 kW and room neutrality cannot be attained.

3.2 Surprises and challenges with stainless steel welding – Tiziano Belotti, CSCS

Whilst executing work to implement a modification on the cooling plant, the CSCS team noted that the original welding work on the stainless steel piping was not so good. Based on the team's experience it was obvious that the welding work during the original construction was not done properly (see Figure 4). As this risked becoming a major insurance case, an external expert was brought in to analyse the situation and write a report. The independent external expert was requested to investigate:

- Whether the welding work, as executed, corresponded to what was requested in the RfP
- Whether the welding work done corresponded to the state-of-the-art quality at the time
- Whether with the current welding there would be a risk of leakages
- Whether the current pipe welding can be expected to last for the normal expected life cycle of such a cooling installation (40 years)

The RfP had originally called for the use of boiler tube steel, meeting the DIN 17100 norm for steels for general purpose structure, for use with cold water and tested for 70 bar pressure. Thus the quality of the material was defined in the RfP. There was no explicit definition of welding quality, which is not at all uncommon. During construction however, the contractor proposed a change of material, viz. to substitute the boiler tube steel by stainless steel at no extra cost. CSCS accepted this proposal and as per standard procedure did not define a quality for the weldings to be executed. Although the engineers had a quality control mandate, the quality of the executed welding somehow went unchecked.

The expert report concluded that based on visual inspection, as well as on radiographic tests, the welding did not meet the state-of-the-art requirements and that a refurbishment would be necessary in order to eliminate the risk of breakage or leakage. The risk was hard to assess precisely but the piping could definitely not be expected to last for the originally envisaged lifetime of 40 years.



Figure 4: Welds from the original construction

A refurbishment plan was therefore drawn up and an elaborate quality control system, to be applied by the contractor executing the refurbishment, was developed by the external expert. The expert was charged with the task of executing regular on-site tests with the following methods:

- A corrosion resistance colour test: a verification by visual inspection that the colour of the - duly polished - welds is golden, which is indicative of the proper formation temperature (200°C - 400°C)
- Internal visual inspection with an endoscope

- A magnetic test: checking ferrite content in the welds to be below 8%
- A penetrant test with a colourant that is able to bring to the fore lack of fusion, porosities, and thin cracks as small as 0.0006 mm
- A radiographic test, defined in the DIN EN 444 standard, dealing with the general principles for the radiographic examination of metallic materials using X-rays and gamma-rays.

Part of the protocol is also that all welds must be numbered, so the welder responsible for a particular weld can be identified. Among the aforementioned methods, the radiographic test is the most complete one. It can be used on pipes up to 400 mm thick. It detects pores, cracks, as well as variation in density and in thickness over very small distances. The DIN EN ISO 5817 standard defines the quality level for the various detectable imperfections, thus enabling very rigorous assessment of quality. However, the test is expensive, time consuming, and requires access to both sides of a pipe. The corrosion resistance colour test and endoscope test are to be performed for all the welds. Other tests, especially the radiographic tests, will be performed only for a limited number of samples.

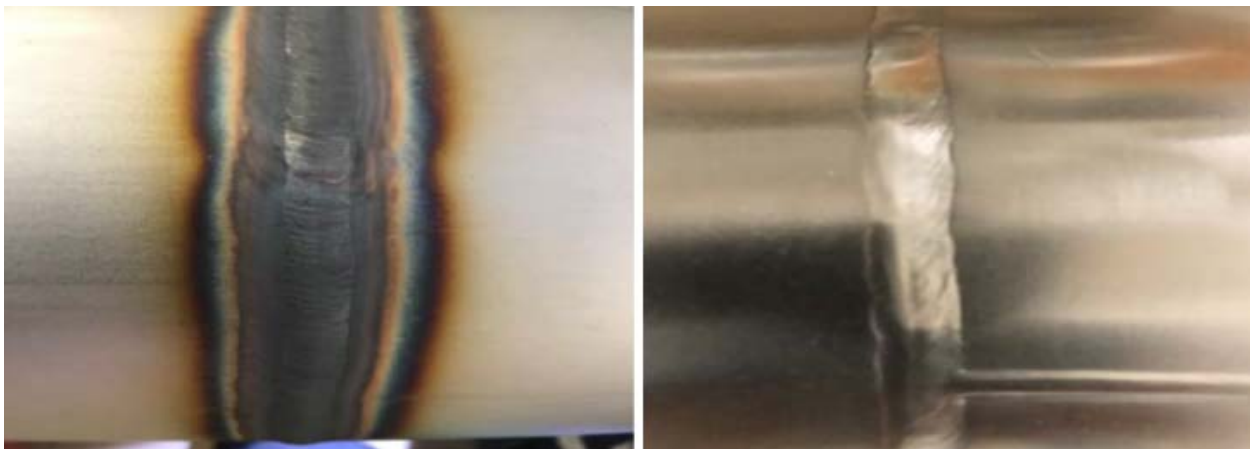


Figure 5: State of the art welding, before (left), and after (right), polishing

Tiziano Belotti stresses that CSCS is not targeting the highest possible quality, but sufficient quality to be confident that the refurbished piping will last for the normally expected life cycle of such installations. The making of good quality welds requires meticulous and lengthy preparation. Tight timelines and cost-cutting can lead to suboptimal results. Therefore, CSCS have acquainted themselves enough with the details of performing state-of-the-art welding (Figure 5) to be able to regularly monitor the welding activity: to measure the composition of the forming gas content in the pipe sections being welded, and to check whether good craftsmanship practices are actually applied.

Thanks to the fact that CSCS has two cooling loops, one for low temperatures (LTL, 9- 17 °C) and one for middle temperatures (MTL, 19 -27 °C), each with their own set of IT systems, CSCS was able to design a refurbishment plan that could be executed without halting datacentre operations:

1. Connect all IT systems to both loops
2. Decrease the temperature for the MTL to 9 °C
3. Switch all the IT systems to MTL
4. Dismantle the LTL and refurbish
5. Switch all the IT systems from the MTL to the new LTL

6. Dismantle and refurbish the MTL
7. Restore original connections of IT systems to their proper loop, and increase the temperature of the MTL to its original temperature

In theory, contractors can be expected to deliver work that meets state-of-the-art quality standards. But, in practice, they do not always do so. Therefore, CSCS suggests that the best practice for a datacentre would be to explicitly define the quality of material and the quality of work by making reference to standards in the RfP as much as possible. Reference to such standards allows for a clear and common understanding of the goals to be achieved with a contractor. Similarly to acceptance tests for HPC machines, that check whether machines do actually meet defined performance criteria, with the welding work, it is worth defining the expected quality of welds and mandating an independent external quality control expert to ensure that standards are met.

3.3 The ENI Green Datacentre in its early years – Luca Bortot, ENI

3.3.1 *A datacentre designed to be a prototype being in production*

The ENI Green Datacentre (GDC, Figure 6) has about 5200 square meters of computer room floor space and can use up to 30 MW of IT power. It is designed to accommodate HPC systems in high density racks, up to 50 kW per rack, but is still a fully air-cooled datacentre. It currently houses two HPC systems: a 12.5 PFlop/s system with a power usage of 1700 kW, and an older 2.8 PFlop/s system with a power usage of about 750 kW. The nodes of the two HPC systems use different generations of Intel CPUs and different generations of Nvidia GPGPUs. The GDC accommodates more than HPC facilities. Different from most HPC datacentres, the ENI GDC is designed for very high availability. It exceeds TIER IV requirements, which aim at an availability of 99.995%. The datacentre can use free cooling for more than 90% of the year. But chillers are available to kick in and help to keep the temperature low on those rare days when it is needed.

From the outset ENI envisaged that in working with their green datacentre, a prototype, there would be much room for learning and improvement. Sound design principles have been used, many techniques have been combined. But determining how they work out, and what must be tuned, necessarily is a matter of empirical research that is conducted continuously whilst in production. The very essence of having a prototype is to acquire knowledge through operation about the many aspects that cannot be predicted from a blueprint.



Figure 6: ENI Green datacentre, Ferrera Erbognone, Pavia, Italy

The datacentre has been designed for automated maintenance and fault prediction. A multitude of sensors was installed to enable monitoring of many components at many levels. What was not expected however was that there would be so many problems with the quality of the data that were collected. Of the more than 300 multimeters installed, over one third had installation problems:

they just produced bad data. At least half a year was spent cross-checking information from different sources to localise those problems. One of the key points to take home is, that a second level of monitoring, viz. the monitoring of the reliability of the data collected from field devices, has to be organised. Monitoring of data reliability is fundamental, as optimisation, improvement attempts, cannot but fail if they are based on low-quality data.

To achieve the four intended goals of the monitoring – datacentre understanding, fault prediction, optimisation, and fault notification – ENI from the outset strived to collect any available data with a high polling frequency, and to subsequently preserve all the data online for a very long time. In the GDC, nearly 100,000 parameters are collected every 10 seconds from all field devices that are connected: 350 multimeters, 1800 probes, 112 UPSs, 4000 switches, etc. At present, after 6 years of data collection, 1012 samples, constituting a volume of around 3.5 TB, are kept online.

Another unexpected but important lesson learned was that having in place automated high availability mechanisms in some cases can lead to fault hiding. Luca Bortot convincingly elucidated this by sharing with us some examples of “embarrassing mistakes”, as he calls them, encountered during the operation of the ENI GDC.

3.3.2 Example: The discovery of a fan “going against the intended current”

The cooling in the ENI GDC is based on a pressure difference that is kept constant by fan activity. A plenum of cold air below the computer room floors is kept at a slight overpressure relative to the floor above it (see Figure 7).

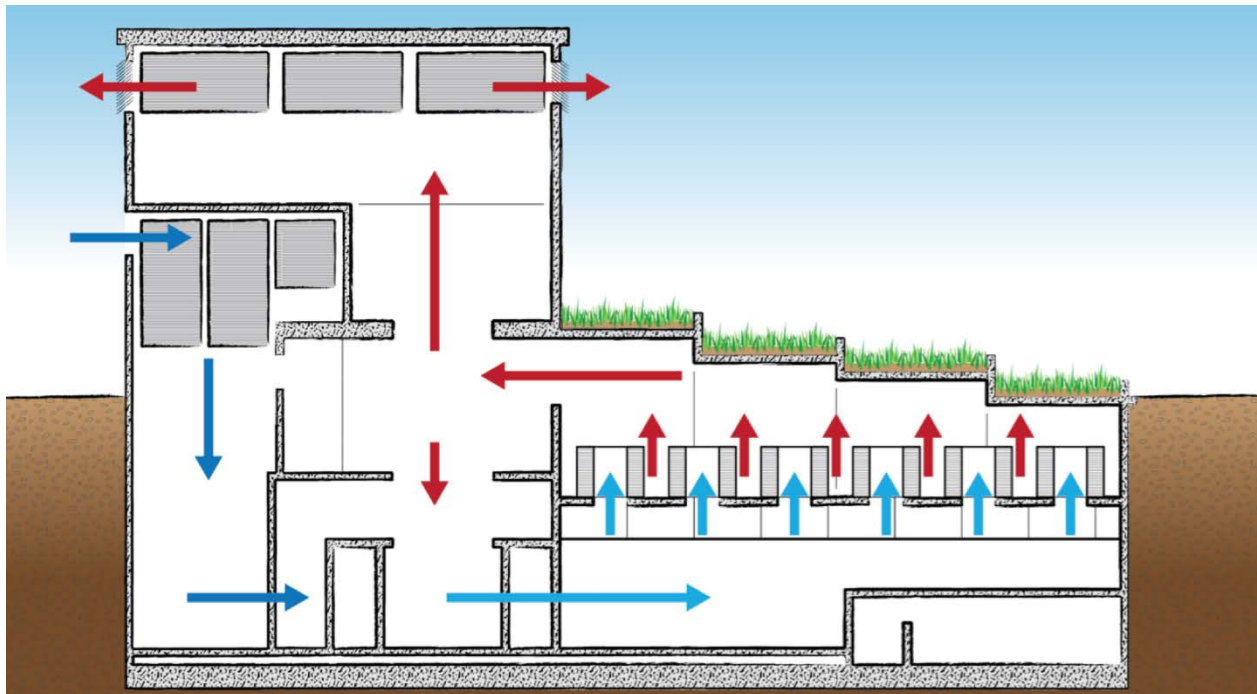


Figure 7: Cold air (blue arrows) pushed up from the plenum below the computer rooms, and the – intended - flow of hot air (red arrows).

There are 8 fans per room, allowing for redundancy. At the time of the incident two units were sufficient to cool the computer room. After a maintenance operation, the datacentre had resumed normal operation successfully. All conditions required for resuming normal operation, the temperatures and pressures, were as needed and expected, and were not different from what they

were before the maintenance. However, one small anomaly was that the power consumed by the fans for a particular computer room to achieve these conditions before the maintenance had been about 22 kW, whereas after the maintenance it levelled at around 38 kW, with four units working, rather than two. As such, this is a small and easily sustainable increase in power since the total power consumption of the datacentre which is in the range of multiple MW. Tracking down the root cause of this eventually revealed that during the maintenance one fan unit had been connected wrongly and was in effect turning backwards, thus effectively working contrary to the intended air flow. So the work of one extra fan was needed to undo the work of the fan blowing in the opposite direction, and the work of yet another one was needed to deliver the pressure at the desired level.

The good thing is that the redundant set of fans kept pressure and temperature at the desired level, as it was designed to do. The bad thing is, that if the details about power consumption of the fans were not available for comparison, or ignored because all environmental conditions were verified to be safe before and after the maintenance, there still is a fault that may go unnoticed for months, or perhaps years. The level of redundancy, the maximum cooling capacity, has actually been reduced. It is no longer at the level it was designed to be and one will only detect this when it is too late. Running at the designed full capacity may actually still be possible. But when a failure occurs in one of the other fans, blowing in the proper direction, while utilising full IT capacity, it will be discovered in a dramatic way that the redundancy level is not what it was expected to be.

Similar non-trivial faults were made during a maintenance operation of the chillers for cold water tanks. Those chillers have two compressors and of one particular chiller only one of the compressors was switched on after maintenance, running actually at its half capacity. Again there was no immediate impact on environmental conditions and operations. This is only detectable if one collects and compares detailed data on the behaviour of the chiller itself. In this case, it will not even use more energy to do its job. To detect what is wrong, one must be able to notice that it takes longer and that the temperature fluctuation has become different from what it used to be and from what one has in the other chillers.

3.3.3 *Monitoring and data collection: “never trust a single source”*

ENI have now developed automated maintenance scripts to aid in the early detection of loss of capacity and / or redundancy. The scripts force components in high availability setups into stress tests. By studying behaviour under a forced run at maximum performance - a situation, because of designed in excess capacity and redundancy, never occurring under normal conditions - hidden faults can be revealed.

Important for the efficient operation of the datacentre is the optimisation of the operating temperature in the computer rooms. The higher that temperature can be, the more days can be benefited from free cooling, and even on days when chillers are needed, the temperature gap they have to bridge is smaller. On the other hand a safety margin is needed. Once again the monitoring system, and the confidence you can have in your data it delivers, determine the room for optimisation.

In the ENI GDC nearly everything is adaptable. It is important that there is a high degree of control over change and over the rate at which changes occur. Spatial uniformity of the temperature and the rate of the temperature fluctuations are important, since the existence of hot spots will have a negative effect on the set point aimed for. Temperature limits, humidity limits, temperature

uniformity and temperature stability are all continuously monitored at a high sample rate and controlled in an automated way. Following the outlined methodology of continuous improvement by gradual controlled changes, over the years, the PUE of the datacentre has been lowered from about 1.5 to 1.12.

A guiding adagio in developing the GDC monitoring, feedback and control system has been to “never trust a single source”. It is not sufficient to collect data and act upon them. Certain relationships should exist and the data should reflect that. But whether they (still) do so must also be checked continuously by comparing measurements for consistency. The monitoring system for the GDC was custom created and is adapted by ENI itself. There is no commercial product involved. The designers of the datacentre in the first years of its operation also became its maintainers. With a reference to software and service developments, ENI could be said to have adopted a “DevOps” approach to datacentre development.

4 Session II – Standards and Regulations

4.1 Tips, Tricks, and Trends in Liquid Cooling: The State of the Practice – Mike Patterson, Intel

4.1.1 *Liquid versus air cooling, a balanced view*

Since liquid cooling of computing units has been used for several years, its benefits are obvious and they are being hyped by different vendors that have this kind of solution in their portfolio. There are however some challenges that disturb the perfect image of direct liquid cooling: air cooling, if done right, may be quite efficient as well. Air cooled PUEs can approach liquid cooled PUEs and air cooling is generally simpler and easier to implement. The right solution depends on the scale. For smaller scales, the economic argument may still suggest air cooling. But the plan is to do recovery of high-grade waste-heat then liquid cooling is the best solution. Therefore the conclusion is that liquid cooling may be mandatory at large and extreme scale, but it is by no means a solution for everyone.

There is no particularly well-grounded metric for system scale, but a reasonable starting point for characterisation around scale and system design considerations is the number of pumps per unit, and the level at which the pumps are situated (CPU, chassis, rack, room). In general the number of pumps/unit goes down as the scale goes up. Some examples of extreme scale are Intel/Lenovo at LRZ, doing SuperMUC-NG with more than 1000 CPUs per pump, and Intel/Cray at Argonne National Lab, planning the future system A21 with more than 5000 CPUs per pump.

4.1.2 *Intel activity pertaining to liquid cooling*

The main goal of Intel’s activities in the liquid cooling area is to move away from proprietary solutions to more standard, plug-and-play products that are vendor agnostic. The Intel Liquid Cooling Workgroup’s objective is to proliferate liquid cooling, by making the ecosystem more ‘friendly’ and lower risk/cost to the OEM. In turn, a more widespread use of liquid cooling by system vendors may be affecting chip makers like Intel. Chip manufacturers can move the TDP higher than the current limits that are more or less enforced by the limitations of air-cooling. Intel is considering forming an end-user advisory group to provide input on what would be the most

important aspects of making liquid cooling more plug-and-play. An interest in joining that group can be expressed by sending an e-mail to michael.k.patterson@intel.com.

4.1.3 ASHRAE Updates

An update of the ASHRAE document “Liquid Cooling for Datacom Equipment Centers” will be available soon, as well as a white paper that will cover some subjects that are critical, from datacentre perspective. Issues that will be covered: water quality (filtration, conductivity, etc.), monitoring and changes to general industry standards imposed by computing device specifics, using CDUs etc.

4.1.4 Immersion cooling

At the moment most of the direct liquid cooled machines use different sorts of heat exchangers on top of chips that need to be cooled. There are however some other ideas that are being experimented with as an alternative. Immersion cooling is one of such ideas. Michel Patterson summarised challenges of this approach, pointing out key areas: cooling efficiency, cost of the solution, environmental hazards imposed by the liquid, chemical reactions between computer components and coolant and serviceability of the machine. For most commonly used coolant (Novec) the key problems are material compatibility and cost. For other, similar fluids either cost is too high or the liquids are toxic. A working group for this kind of cooling is also being considered by Intel.

4.1.5 System Design Guidance topics

The presentation was concluded with several system design remarks or best practices that should be considered when planning to use liquid cooling. First of all, all chips work better, are more reliable, and are more energy efficient at lower temperatures. One can have the same chip temperature with different coolant temperatures depending on the coolant and how close to the chip the coolant is delivered. For most cases, the highest coolant temperature is not the optimal solution. Rather, because chillers represent the most expensive cooling infrastructure to buy and operate, the inlet temperature of the coolant should be as low as possible but still high enough for chiller-less free cooling. This implies that climate data have to be taken into account. It may mean different solutions in different climates.

For warm (ASHRAE W3-W4) coolant loops, biofilm build-up is a real danger, so one should pay attention to areas with low flows and avoid dead legs in the piping. Last but not least, it is important to pay attention to the right quality of materials being used in these loops.

4.2 Open Standards for Datacentre Availability – Lex Coors, Interxion

The presentation of Lex Coors covered the issue of designing, assessing, and defining a “green” datacentre. The “[green grid project](#)” [3] is trying to work out open standards for datacentres that allows them to be more energy efficient without compromising on the basic requirements.

The problem of energy efficiency of datacentres is becoming more important every year as the colocation industry grows 15-20% annually and this trend seems to be stable - the prediction of power consumption by the industry forecasts growth from 16 GW in 2008 to 60 GW in 2020. The sum total of greenhouse gas emissions involved in this level of energy usage is equivalent to more than 300,000,000 metric tons of carbon dioxide.

Facing this problem, the green grid project seeks to develop a feasible plan for decreasing the greenhouse gas footprint of the datacentres by 80% by 2050. The key is a sustainable energy model

based on the massive usage of renewable energy in datacentre industry. Even if the computers are evolving and are becoming more energy efficient, the growth of the datacentres themselves and increasing demand for IT services means that the energy consumption will grow, so using renewable energy is a must. Unfortunately, renewables, like solar and wind, are not a reliable source of energy so solving energy storage issues is a key challenge. Fortunately, the European Union is and will be investing in the development of energy storage.

The next topic was focusing on different aspects of reliability of datacentres and ways for keeping desired levels of service without traditional massive redundancy of infrastructure. A use case was presented: one traditional approach with two synchronous redundant datacentres with diesel generators as the controlled back-up power source, and a second with only single datacentre but with fuel cells as the power source. According to the reliability model the first granted 99.999% reliability while the second yielded 99.9999% with lower costs and energy usage.

An online tool, the Open Standard for Datacentre Availability calculator, also named “osda-tool”, is being prepared for modelling datacentre reliability. The tool can be accessed via the [osda-tool page](#) on the website of the green grid project [4].

4.3 Update from the Energy Efficient HPC Working Group (EE HPC WG) – Natalie Bates, LLNL

The EE HPC WG currently has about 700 members from many countries. About 50% of those are from the user side: particularly individuals working at or involved in all the major HPC sites. About 30% is from the vendor community: all the major system integrators are represented and there are many vendors of hardware and software that is particularly relevant to energy efficiency. 20% of the members are from the academic research communities.

During her presentation Natalie Bates presented topics on which the EE HPC Group has recently focussed. The EE HPC WG primarily focusses on sharing best practices and working with standards defining bodies such as ASHRAE. It also monitors research on energy efficiency.

One of the key characteristics of the HPC datacentre is high power fluctuation. The peak electricity demand for supercomputing is increasing and is, at the same time, becoming more dynamic. While in commercial colocation datacentres the power draw is more or less stable, HPC machines are characterised by big power swings. Starting even a single large job may cause dynamic power conditions that result in power quality events that, as consequence, may propagate outwards as voltage fluctuates on the power distribution feed and cascade to other facilities. A grid with insufficient voltage support resources can become unstable, as the voltage does not quickly return to equilibrium.

The computing power demand forecast of the LANL site was presented as a typical example of an HPC site. LANL is a major electric power consumer in the area, with a demand of 40 MW forecasted for 2020. Currently the power demand may fluctuate by 8-14 MW in short time periods and this already is causing power demand on the grid to become more dynamic. There are concerns about voltage fluctuations in the grid, drop out of equipment, and even equipment damage.

The EE HPC Group tries to predict possible problems of this kind and to explore ways of minimising the impact on external entities. Understanding future conditions requires modelling and time-dependent analysis. Several options to address potential concern are available. The Electrical

Transient and Analysis Program (ETAP) is developing such software under established quality assurance standards. LANL has selected ETAP modelling software for distribution system dynamics. GE PLSF, an energy consulting firm of General Electric, has selected ETAP modelling software for transmission system dynamics.

Within the EE HPC Group there are also smaller sub-teams interested particular aspects of, or related to, energy efficient HPC: The usage of energy efficiency considerations in procurements of new machines, and reliability, availability, serviceability, and maintainability of machines and infrastructures.

One of important aspects of EE HPC Group activities is the optimisation of the energy efficiency of supercomputers by the means of software cluster management and scheduling software. An initial document with a first analysis was released May 2018. The final document was submitted to Supercomputing 2018.

The activities of EE HPC Group are announced and documented on their [website](#) [5], where also their reports can be accessed.

5 Session III - Energy to Solution and TCO

5.1 Presentations of the results from the three pilot systems of the PRACE-3IP Pre-Commercial Procurement

5.1.1 CINES, prototype “Frioul” – Jerome Chapelle & Eric Boyer

5.1.1.1 PCP goals

The presentation started with a quick reminder about what the PRACE-3IP Pre-Commercial Procurement (PCP) was aiming for. The main goals of the PCP was to fund R&D in order to improve the energy efficiency of HPC systems and to procure small scale systems for PRACE with an architecture that:

- Could scale up to 100 PFlop/s
- Demonstrated an energy efficient architecture in real applications
- Provides possibility for energy measurements with a fine granularity

As such, the PCP was aiming at testing the concept of a pre-commercial procurement as an instrument to help the development of new technology that is almost ready for commercial use. This PCP started in 2012 and it was decided to have three steps in the process as recommended in the approach defined by the European Commission:

- Solution design
- Prototype development
- Pre-commercial small-scale production/service development – field test

Partners in the process were the five PRACE-3IP project partners CINECA, CSC, EPCC, JUELICH, GENCI together with the PRACE aisbl as an observer. During the process, it was decided to procure three different prototypes to be installed in CINECA, JUELICH and GENCI/CINES. The total budget for the project was 9 million Euro, where half was provided by the partners and half by the European Commission.

5.1.1.2 “Frioul”

Below the prototype from Bull/Atos at CINES, called “Frioul”, is described. The hardware consists of a Bull Sequana cell with the following characteristics:

- 511 TFlop/s
- 168 Intel® Xeon Phi™ 7250
- 56 Bull Sequana X1210 blade
- IB EDR interconnect

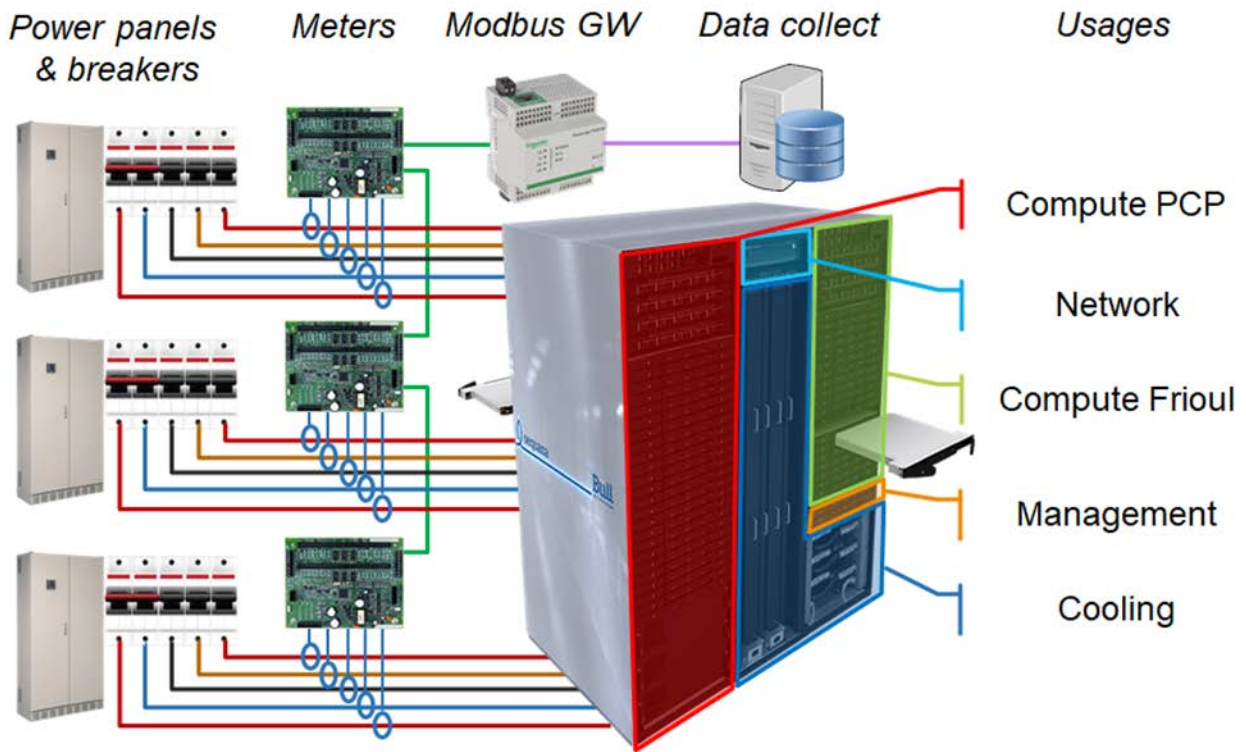


Figure 8: Power provisioning and power usage measurement for specific components of Frioul

The prototype had many improvements compared with existing products:

- The hardware was 100 % water-cooled including the power supplies
- Dynamic Resource Reconfiguration for energy efficiency
- Energy-aware system. Fine-grain FPGA based energy sensors, embedded on each compute blade, measuring energy use with 1 kHz and 2 % accuracy
- Software tools to visualise and optimise the energy use from the energy sensors
- Improved network management of the Mellanox InfiniBand interconnect

CINES also developed a system to measure the energy at each power distribution circuit and breaker. Each breaker is equipped with a power meter and data is collected to a database via a Modbus network. With this second measuring system, power used in addition to the nodes, such as power used by the interconnect and cooling etc., can be accounted for. Figure 8, shows how the provisioning and the various components are organised and how the power usage of each component can be measured.

CINES also developed the PowerMonitor software that is modular with libraries and plugins supporting many sensors types (Modbus, SNMP, etc.) and brands together with different data connection (databases, files, etc.).

The water-cooled power supplies, as well as the software tools developed by Atos, will soon be part of the standard offering by Atos.

The Frioul system and the detailed power usage monitoring capabilities have been used to investigate the application performance and energy usage of (components of) the system using four HPC codes selected from the Unified European Applications Benchmark Suite (UEABS): NEMO, BQCD, Quantum Espresso, and SPECfem3D. Time-to-Solution (seconds) and Energy-to-Solution (kWh) reference values for running the selected codes with specified problem sizes could also be based on the UEABS.

5.1.2 CINECA, prototype “D.A.V.I.D.E.” – Carlo Cavazzoni

Carlo Cavazzoni from CINECA presented the D.A.V.I.D.E. (Development of an Added Value Infrastructure Designed in Europe) prototype built by E4 Computer Engineering. Currently the systems is no longer a prototype. It has been taken into production.

The system is based on OpenPower P8 CPUs and NVIDIA P100 GPUs. This gives a peak performance around 1 PFlop/s. Figure 9 and Table 1, below, summarise some technical data of the prototype.

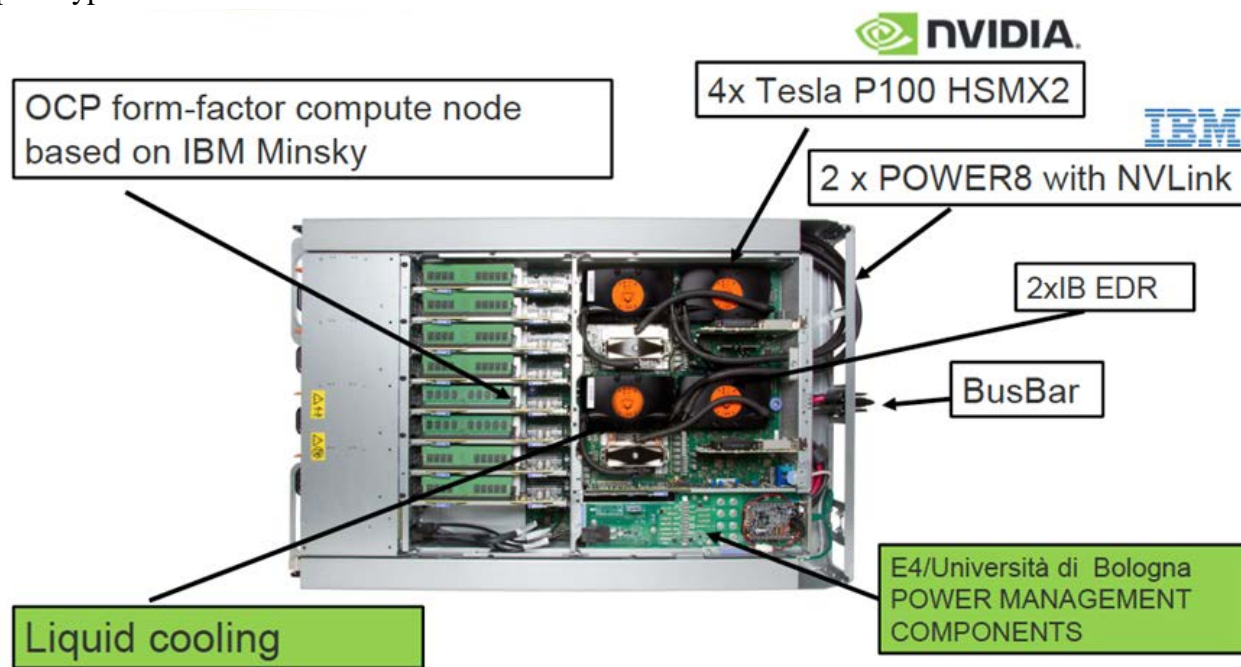


Figure 9: Architecture of the D.A.V.I.D.E. prototype

| | |
|-----------------|------------------|
| Number of nodes | 45 |
| Form factor | 2U |
| SoC | 2x Power8 NVlink |
| GPU | 4x P100 Nvidia |

| | |
|---------------------|-----------------------------------|
| Network | 2x IB EDR, 1x 1GbE |
| Cooling | SoC and GPU with direct hot water |
| Maximum performance | 22 TFlop/s per node |
| Storage | 1 × SSD SATA, 1 × NVMe |
| Power | DC power distribution |

Table 1: Technical data of the D.A.V.I.D.E. prototype

In the project, tools for fine grain power monitoring and power capping have been developed together by the University of Bologna and E4 Computer Engineering. The fine grain power and performance measurement can be used to:

- Verify and classify node performance
- Plan predictive maintenance
- Perform per-user - Energy / Performance – accounting

Together with the scheduler - via a scheduler plugin - the power capping can control the power consumption during the execution so that a maximum power limit is enforced. The power measurement that is developed in the project is based on a PPBB (Power and Performance “Black Box”) on each node. This measures the performance without affecting the execution on the node with a time-resolution of 1 ms. The monitoring tools can display the power consumption on a coarse scale and fine scale down to ms resolution.

Table 2 below, shows the time-to-solution and energy-to-solution results for several HPC applications obtained on the D.A.V.I.D.E prototype.

| Application | TTS [s] | | | ETS [kWh] | | |
|------------------|-----------|--------|----------|-----------|--------|----------|
| | Reference | Target | Measured | Reference | Target | Measured |
| BQCD | 1,584 | 1,584 | 1,6291 | 170 | 74 | 159 |
| NEMO | 1,942 | 1,942 | 84,2670 | 365 | 50 | 1003 |
| Quantum Espresso | 3,216 | 3,216 | 1,063 | 343 | 127 | 14 |
| SPECFEM3D | 595 | -- | 991 | 23,7 | -- | 6 |

Table 2: Time-to-solution (TTS) and Energy-to-solution (ETS) results for the D.A.V.I.D.E. prototype

Except for SPECFEM3D, the reference values, to compare the measured results on the prototype with, are directly based on the PRACE Unified European Applications Benchmark Suite (UEABS). For SPECFEM3D however, a new reference value for a smaller dataset, had to be obtained, using

the PRACE Tier-0 system Curie, as the prototype had not enough memory to accommodate the original problem size and dataset used by the UEABS. The results obtained on the prototype vary considerably from code to code. The codes have been run “as is”. A code like Quantum Espresso, is very able to exploit the GPUs, whereas other some other codes clearly would need adaptation to be able to do so.

5.1.3 JSC, prototype “JUMAX” – Dirk Pleiter

Dirk Pleiter, from Juelich Supercomputing Centre, presented the third PRACE-3IP PCP system, “JUMAX”, which is based on technology from Maxeler utilising FPGAs and dataflow concepts for acceleration. The accelerator is called “MAX5” and is based on one to two commodity CPUs and at least one FPGA-based dataflow engine (DFE), in this case, the Xilinx UltraScale+ VU9P.

The MAX5 card has a memory architecture with three different kinds of memory:

- Host memory
- MAX5 LMEM = large memory: DDR4 DIMMs
- MAX5 FMEM = fast memory: FPGA internal memory

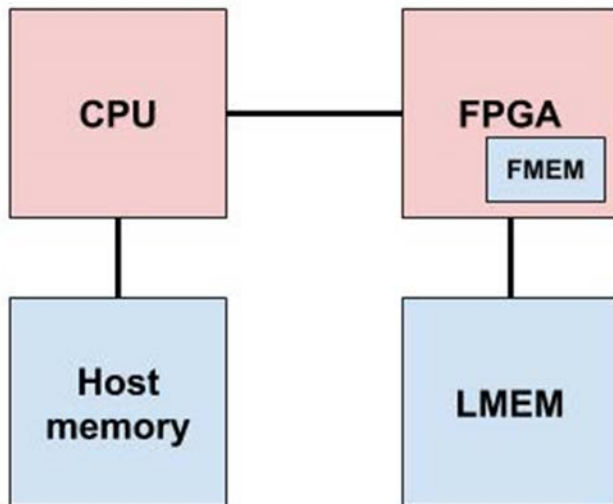


Figure 10: Memory organisation of a system with a MAX5 accelerator

The accelerator is programmed in a high-level language for FPGA programming: MaxJ, which is in its syntax based on Java, but in concept quite different. The language has been developed by Maxeler and is specifically optimised for implementing a data-flow paradigm. This means that data operations are executed based on the availability of data. The application is split up in, on the one hand, a host application, written in C, C++, Fortran, Python, etc., and on the other hand, a dataflow kernel, written in MaxJ. Because of this split, only the most compute-intensive parts of the application need to be redesigned in a dataflow manner.

An example of a computation expressed in MaxJ was given. The simple computational task of:

$$C_i \leftarrow A_i + B_i \text{ for } i = 0, \dots, N-1$$

can be translated into the following MaxJ program:

```

DFEVar A = io.input("A", dfeFloat(8, 24));
DFEVar B = io.input("B", dfeFloat(8, 24));
  
```

```
DFEVar C = A + B;

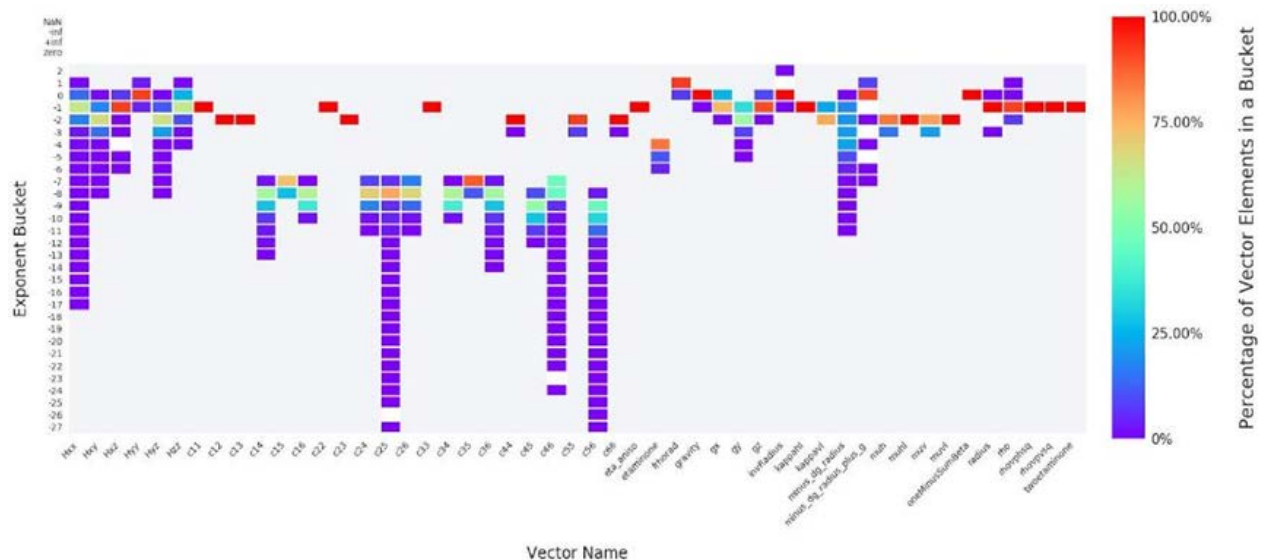
io.output("C" , C , dfeFloat(8, 24));
```

The FPGA can operate with floating-point numbers of different lengths in addition to standard IEEE floating-point numbers. MaxJ allows the definition of floating-point variables with a different number of bits for mantissa and exponent. This can save considerable time and energy for floating-point calculations.

The prototype work led to several improvements in the Maxeler technology:

- Improved balance between CPU and dataflow components
- MAX5 bitstreams for the benchmark applications
- Improvements to the Maxeler monitoring infrastructure and firmware
- Power monitoring and (coarse-grained) energy measurement
- Value Profiling Library to support floating-point to fixed-point conversion process. Floating-point operations can be implemented using a different number of bits. Profiling of the value range of variables can help to guide the decision.
- Execution Analysis and visualisation tool

Figure 11 below, exemplifies the usage of the Value Profiling Library.



The MAX5 cards are equipped with:

- Xilinx VU9P FPGA
- 42 MByte of “fast memory” (FMEM)
- 48 GByte of “large memory” (LMEM)
- Custom ARM-based control board

Figure 12 below, shows the architecture, the interconnectivity between the MAX5 cards and other elements of the system in JUMAX, the pilot system at JSC.

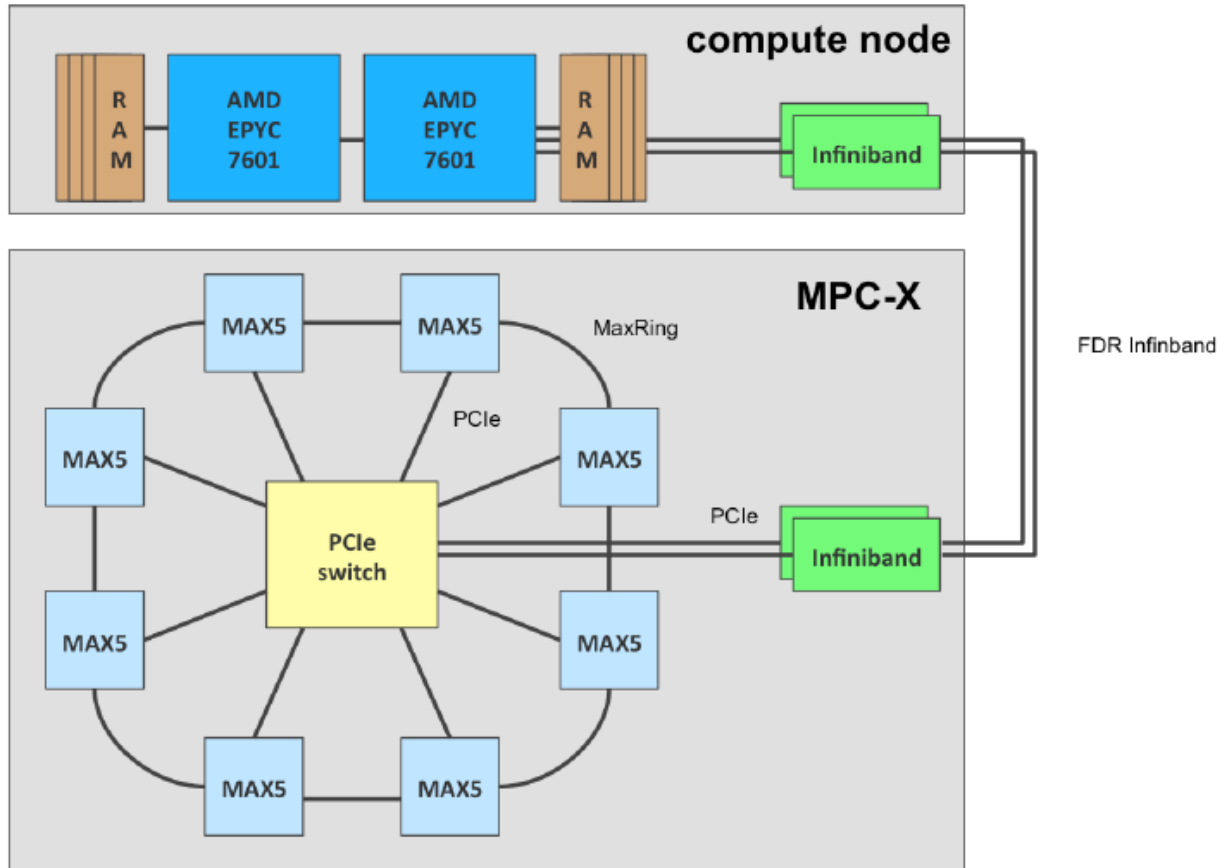


Figure 12: Interconnectivity between the MAX5 cards and other elements of the system

Table 3 below, shows the estimated maximum power consumption of system components.

| Component | Power [W] |
|----------------------|------------|
| Host server | ~500 |
| MPC-X server (empty) | ~80 |
| MAX5 card: | ≤116 |
| <i>FPGA</i> | <i>70%</i> |

| Component | Power [W] |
|-----------------------------|------------|
| <i>transceivers</i> | <i>11%</i> |
| <i>power infrastructure</i> | <i>10%</i> |
| <i>memory</i> | <i>10%</i> |
| MPC-X server (full) | ~1000 |

Table 3: Estimated maximum power consumption of node components

The challenges to adapt the applications to the architecture were discussed, that is: the work to reformulate the kernel of the application to follow the data flow paradigm. One important point is that increased concurrency is needed. It should be noted that the FPGA operates at 150- 250 MHz, which is 6 to 10 times slower than an Nvidia V100.

Several other points need to be taken into account:

- There is a need for clever kernel merging
- Do not foresee reprogramming of FPGA during job execution
- Also, reconsider the need for high floating-point precision. Reduced precision gives a significant saving on FPGA resources

| Application | Without DFE Acceleration | | With DFE Acceleration | |
|------------------|--------------------------|---------------------|-----------------------|-----------|
| | TTS [s] | ETS [kWh] | TTS [s] | ETS [kWh] |
| BQCD | 8507 | 0.56 (estimated) | ~2500 | 0.40 |
| NEMO | 773 | 0.11 (estimated) | 1099 | 0.16 |
| Quantum Espresso | 447 | 0.06 | 195 | 0.04 |
| SPECFEM3D | - | - | 158 | 0.02 |

Table 4: Time-to-solution (TTS) and Energy-to-solution results for the JUMAX prototype

Table 4 shows the time-to-solution and energy-to-solution results for several HPC applications, selected from the Unified European Applications Benchmark Suite (UEABS), obtained on the JUMAX prototype, with and without DFE acceleration.

The experience of integration into the datacentre environment was described. The power density was not considered as a challenge with up to 750W/U. Hot-water cooling could possibly be more challenging as standard FPGAs used to have tighter temperature limits. The experience of the management of reconfiguration in a datacentre environment at large scale is also lacking.

In summary, the prototype was considered a successful demonstration of porting relevant scientific HPC applications to an FPGA-accelerated architecture. But there are still many challenges to solve and more experience is required for large-scale operation especially regarding reconfiguration. A potential for improving energy efficiency was indicated but more efforts are needed to fully establish the benefit. Integration into a normal data centre environment was considered a moderate challenge.

5.2 TCO model for HPC procurement assessment – Eric Boyer, GENCI

Eric Boyer, GENCI presented a TCO model that has been used in several procurements in France. There are several good reasons to use the Total Cost of Ownership as a key factor in procurements. It encompasses CAPEX and OPEX of the proposed system during its lifetime. Traditionally only the CAPEX has been used in procurements but that is not sufficient anymore since the OPEX has increased (especially the energy costs). In the model developed by GENCI, the TCO is calculated for a specific workload. The reason for this is that the energy consumption is dependent on the workload profile. The workload should be selected to be as close as possible to the expected production load during machine lifetime. The workload consists of a set of codes that run simultaneously on a given number of nodes. The TCO model takes into account the cost to install and operate the infrastructure around the system.

The overall goal of this kind of procurements is to find the most cost efficient system for the defined workload. You select the system that can run the workload the largest number of times for a given budget. It gives a strong incentive for the vendors to provide a balanced system with low energy consumption.

The costs in this TCO model include (for the planned lifetime):

- Machine acquisition cost
- Maintenance costs
- Additional staff costs
- Energy costs for:
 - Compute nodes and interconnect
 - Storage and service nodes
 - Cooling components (internal)
- Costs for infrastructure specific add-ons

The actual energy consumption is measured during the execution of the workload and the cost can be calculated. The vendors can change the order of runs in the workload to get the best result. Optimisations of the codes are also allowed as well as frequency adjustments during the execution to optimise the energy use.

The presented TCO based procurement method has been used in several procurements by GENCI with good results, for example, the Occigen system at CINES that was purchased in two steps and the Joliot-Curie system at TGCC.

6 Session IV – Bologna Science Park

6.1 Bologna Science Park: an infrastructure for HPC – Gaetano Moron, INFN/CNAF

Plans for a new datacentre, with focus on HPC and aiming to be one of the largest in Europe, were introduced during the presentation of the developed Bio and Science Park in Bologna, Italy. This is a common effort between the regional government of Emilia-Romagna region, the Italian government, INFN-CNAF, and CINECA, that started already in 2011. It now becomes a reality to accommodate all the current and the future HPC/HPDA/HTC workload, not only for Italy and its research but also for the ECMWF systems already assigned to Bologna.

The whole science park partially reuses an old tobacco factory. Thus it is not wasting any green fields but recycles an already developed land (brown field). The science park in a sense rises “from the ashes of the tobacco industry”. Also, as Bologna already houses 70% of the Italian computational power (not including ECMWF resources) in INFN-CNAF and CINECA, the network connection to this region is well sized.

The overall size of the park will be 100,000 square meters, but only approximately 18,000 square meters will be used for the datacentres. While the ECMWF part is already well-defined, the INFN/CINECA part is still in design consideration phase. This is also reflecting the fact that ECMWF part should start the operations in 2019, while INFN/CINECA will move their equipment at the earliest in 2021.

The ECMWF datacentre will start with two computer rooms and two rooms for power and cooling infrastructure with one extra computer room for expansion (each compute room is approximately 3000 square meters).

In total, the ECMWF part of the park will have 10+10 MW power lines (with the possible extension to 20MW) from a high voltage substation of the electricity providing company (Enel), located next to the park and 5 x 2 MW Diesel rotary UPS devices for backup. To lower the expenses for the power, an array of photovoltaic cells will cover the roof with the possibility of getting 500 MWh per year. The cooling will be provided through dry cooling towers (4x 1850kW capacity) and chillers (5x 1400 kW capacity). The aim is to utilize wells providing chilled water to further lower the needs for mechanical chillers. Expected PUEs are 1.35 peak and 1.18 maximum annualised.

CINECA and INFN will have both one datacentre with the size of approximately 3000 square meters, up to 20 MW of power (in a 1+1 redundancy setup). The decision on the level of redundancy has not yet been taken. The use of tri-co-generation (electricity, heat and cold), using fuel cell technology based on natural gas, should help to minimise the PUE. This goes also for the cooling technologies to be used. Targeted PUE is around 1.2 to 1.3.

Overall a good strategy was developed for at least the next 10 years to accommodate the major Italian HPC resources and it will be interesting to follow this project especially with the aim on considerations for the INFN/CINECA part where some novel design and new experiences are to be expected.

7 Session V – Heat re-use

7.1 Making HPC Climate Positive – Eric Rylander, Stockholm Data Parks at Stockholm Exergi

Stockholm Exergi is a local energy company that produces resource-efficient heating, cooling and electricity for the citizens of Stockholm. The company, formerly known as Fortum Värme, also runs the district heating system infrastructure that brings the heat to its consumers. Erik Rylander is head of Stockholm Data Parks at Stockholm Exergi. The goal and vision of Stockholm Data Parks are to develop a truly green large scale datacentre industry in the Stockholm area where no heat is wasted. Stockholm Data Parks develops greenfield sites as well as conversion sites for datacentre activity, to provide not only space, power and network, but also heat recovery and cooling as a service (CAAS).

The company has ample experience with the Stockholm area district heating system that has been gradually expanded since the 1950s, and into which heat recovery from heterogeneous sources has already been successfully integrated. Heat is already recovered from various industrial processes, from waste burning, from sewage water, and already from some datacentres as well. Due to the fairly cold climate in Stockholm, heat has a value. Stockholm Exergi can offer a price for the heat of companies that want to get rid of their waste-heat, and a better price if the waste-heat is a high volume of high-grade heat that can be easily processed. They can do so, because 95% of central Stockholm buildings are connected to the district heating network. Currently, these collectively represent a demand for heat supplies of more than 12 TWh per year.

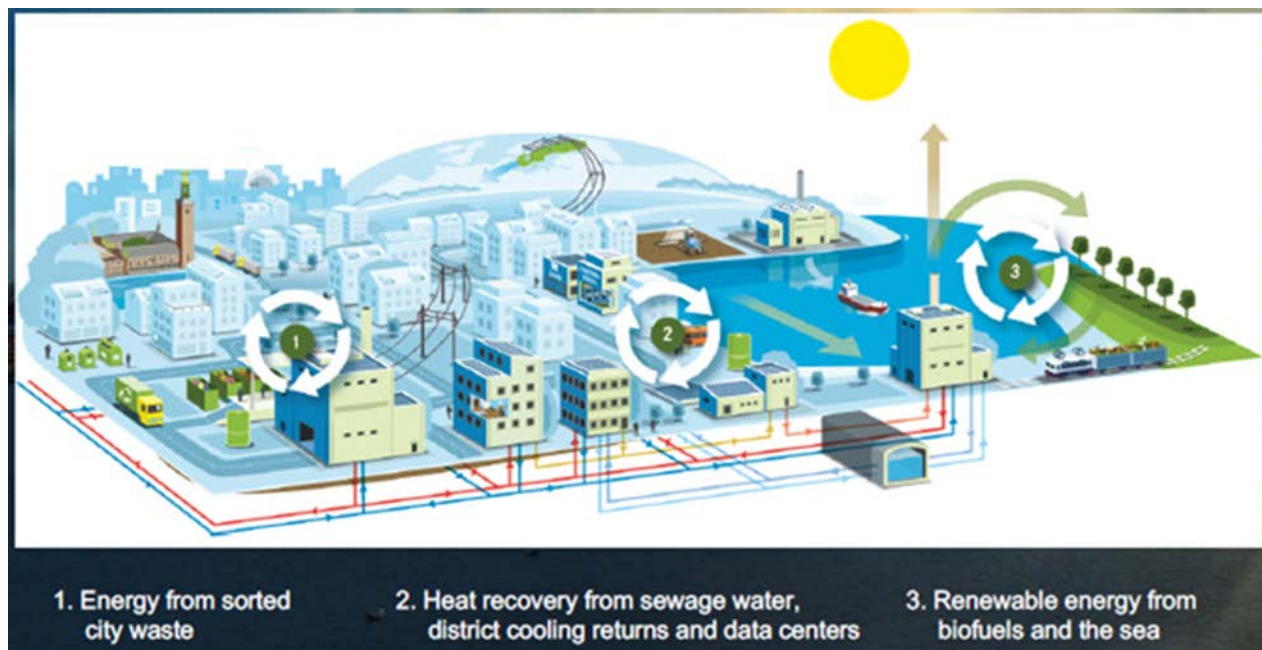


Figure 13: Heat recovery in Stockholm

Figure 13 shows how district heating (and cooling) of Stockholm is integrated with sustainable and locally recovered energy resources. At present, energy plants based on fossil fuels are still part of the energy provisioning scheme. Stockholm Exergi and Stockholm Data Parks plan to have gradually phased out fossil fuels by 2040, and their vision is that datacentres can and should play

a major role in this transition. The waste-heat produced by a 10 MW datacentre is estimated to roughly meet the heat demand of about 20,000 modern residential flats.

Stockholm Data Parks closely collaborates with a power grid operator (Ellevio) and a dark fibre provider (Stokab) to bring together and prepare all necessary infrastructure in various locations around Stockholm that also are well-connected to the district heating and cooling infrastructure.

They seek to foster more large-scale datacentre activity by offering the following options to large datacentres:

- In the “great exchange” option, cooling of the datacentre as a service (i.e.: integral management of the chilled water supply needed by the datacentre) is offered for free, in exchange for the excess heat when the datacentre load reaches above 10 MW
- In an alternative option, the cooling is managed by the customer itself and Stockholm Exergi purchases excess heat at a price reflecting its alternative heat production cost – which is currently about \$200.000 per MW heat, per year

When the customer manages the cooling, Stockholm Exergi contractually obliges itself to buy the datacentre’s heat whenever the outdoor temperature is below 12°C, which, in Stockholm, is about 6,000 hours per year. The price at which heat is bought has a capacity component and an energy component that varies with outdoor temperature.

The current price levels are as follows:

- Capacity payment: 83 EURO/kW, year
- Energy payment:
 - $\leq 5.0^{\circ}\text{C}$: 16.7 EURO/MWh
 - $5.1^{\circ}\text{C} - 12.0^{\circ}\text{C}$: 11.1 EURO/MWh
 - $\geq 12.1^{\circ}\text{C}$: 11.1 EURO/MWh (when called for by Stockholm Exergi)

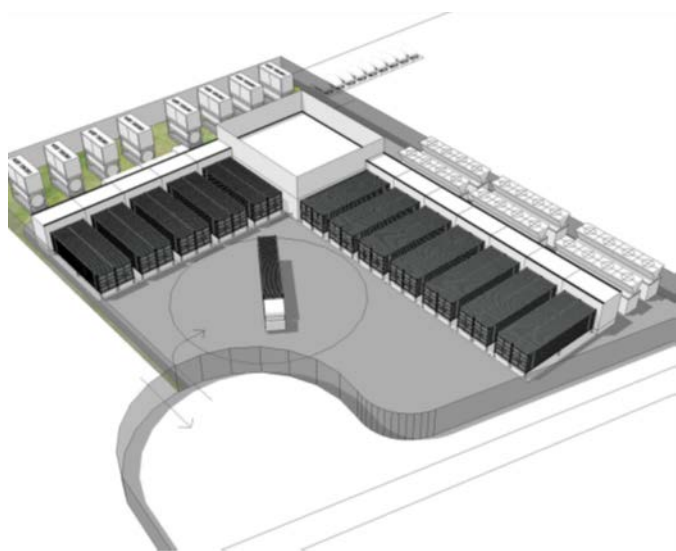


Figure 14: Containerised datacentre site with network, power, cooling, and waste heat disposal connections

In addition, the electricity price in Sweden is very low compared to other countries: for consumers between 150,000 MWh and 700,000 MWh per year, in 2017, 0.040 EURO per kWh, excluding VAT and other recoverable taxes and levies.

Stockholm Data Parks datacentres are sustainable and can become net climate positive, as their heat re-use makes a significant contribution to reducing global CO₂ emissions, by replacing more CO₂ intensive means of heating the city.

Stockholm Data Parks is currently developing a modular, containerised, form of datacentre sites, where clients can choose, per client, the type of power redundancy they want, whether and how they want cooling and heat recovery. Figure 14 shows an impression of a fully containerised datacentre site, as envisaged by Stockholm Data Parks.

7.2 The World's first Climate Positive Datacentre – Lars Schedin, EcoDataCenter

Lars Schedin, economist and CEO of EcoDataCenter, presented a generic argument on the energy usage of the datacentre industry and the need and possibility to turn around the industry into a more climate friendly practice.

Steaming a YouTube video of a song with a video clip is estimated to cost on the average 0.17 kWh of energy consumption. That includes “everything”: the servers, the datacentre overhead, the network in between, the client, etc. This does not seem like a lot. However, this implies that a single popular video alone, viewed 5 million times, is costing 850,000 kWh, which is the equivalent of 400,000 tons of CO₂ emission. This is, for example, the equivalent of what 220,000 average Swedish taxi cabs emit in a year.

IT services are consuming more electricity each year, and most of the electricity is produced using fossil based sources. From 1994 to 2014 the volume has grown from about 83%, from 12 PWh to 22 PWh¹, but as shown in Figure 15, there has been little to no progress in using other, renewable, energy sources.

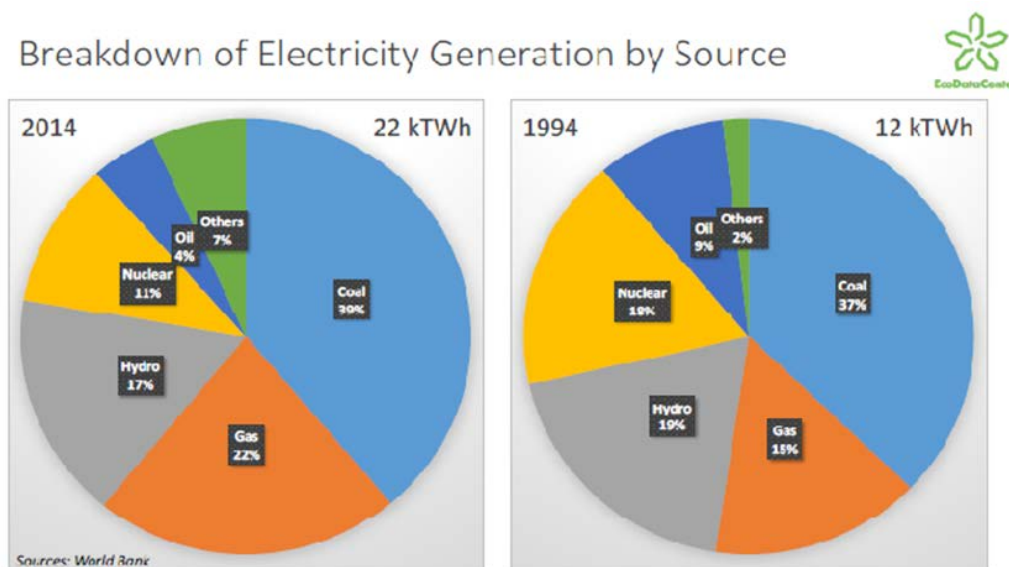


Figure 15: Electricity, generation by source in 2014 and 1994

As Lars Schedin puts it: “The data tsunami is about to absorb 20% of the world’s energy”. This is not sustainable. Three cornerstones for the design of new datacentres need to be taken in consideration:

¹ PWh is indicated as “kTWh” in the graphs of Figure 15

- Low PUE
- Use of renewable energy
- Heat re-use

Of course, climate conditions differ and the possibilities for using renewable energy and re-using waste heat are location dependent. But it should be seriously considered not to put datacentres in particularly unfavourable climates, since the cost for transmitting digital information is much lower than the cost of transporting energy.

EcoDataCenter, in Falun, Sweden, has been designed and is operating along these lines. While electrical and cooling systems have been designed to meet the high requirements of a Tier IV uptime facility, this is achieved with a PUE of 1.15. The electricity available in Falun is 100% renewable. No fossil or nuclear power is available. The datacentre is connected to a district heating system and a pellets production plant that needs a lot of heat. In Falun, heat has value. Heat re-use from the datacentre creates a climate positive datacentre. The datacentre plans to expand its power to 60 MW, all on the basis of renewable energy.

8 Session VI – Energy grids and Power provisioning

8.1 Management of Energy costs at PSNC – Radoslaw Januszewski, PSNC

The first solution envisioned by PSNC to reduce the global cost of energy is simply to reduce the energy bills. To this end, PSNC and 31 other public entities have pooled their purchasing of energy in "purchasing group UM Poznan" that tenders agreements with energy distributors for them. PSNC used to pay the regular tariff. Through the agreement obtained by the purchasing group with energy company Enea, PSNC achieved a price reduction of 38%.

After obtaining a better price, energy costs can be further optimised by reducing energy usage. PSNC added a synchronous condenser, which reduces voltage ampere reactive power. Reactive power is important for ensuring voltage stability, but it does not do any real work. Higher reactive power means grater line losses. The investment paid itself back within a year.

PSNC emphasize that oversizing may reduce some risks, but rightsizing is generally better for cost reduction. There is a real trade-off between cost on the one hand, and features and safety on the other hand. They aim for an electricity contract with limits that are as close as possible to their actual energy consumption. PSNC views decisions about energy contracts to be essentially a form of risk assessment and risk management. "The closer you can actually be to the limit of your contract, the cheaper the electricity is for you, but beware not to go over the contractual limit, because then you probably pay a lot extra, and you lose". Irrespective of the technical ability of the power infrastructure to handle them, large fluctuations in load, causing sudden and steep rises in power demand, are therefore undesirable from a financial point of view.

PSNC has an old and a new datacentre, and plans to keep both in production. The old datacentre has 300 square meters of computer floor space and about 0.5 MW of power. It has a PUE of 1.8. The new one has 1600 square meters of computer floor space and 16 MW of power. It has a PUE of 1.3. A modular infrastructure that grows and is activated when actual needs for it have risen, is best for cost management. This works for the new datacentre but is hard to apply in practice to the old datacentre, where there is literally no room to move and no space to bring in advanced

technology, and where the return on investments that would improve operational cost is often debatable or uncertain at best.

Following recommended practices for managing air flow PSNC has applied hot-aisle / cold-aisle containment in both datacentres. In the old datacentre however, with quite limited success. Both datacentres have an ambient temperature of 22°C. In the new datacentre, that is the cold part. The temperature in the contained hot aisles can go up to 40°C. By contrast, in the old datacentre, the cold aisles are contained and have a temperature of 17°C for the air that goes in to the racks. Raising the ambient temperature in the room would be quite bearable for human operators and in theory would be more energy-efficient, but in practice it caused the CRACs in the room to malfunction.

In the new datacentre, there are no CRACs. DLC configurations are used. However, in practice only about 70% of the heat is dissipated by DLC. There is some over pressure in the hot aisle corridors that appear to be not well enough isolated. Consequently, a substantial amount of heat is leaking out into the computer room and an additional loop with chilled water is needed. For the next machine, PSNC plan to make well-insulated walls, doors, and ceilings into a requirement.

The adoption of oil-free compressors (Danfoss Turbocor compressors), that (re)move ten times more heat than they consume power, was cost effective. And so was the use of humidity control by means of humidifiers with steam cylinders in the old datacentre.

PSNC also have introduced power management into their batch environment. Compute nodes are powered off by SLURM, the batch system, after having been idle for 30 minutes. They are powered on again, when new jobs have been submitted in such quantities that there is a demand for nodes. Whether this is beneficial for energy cost management remains to be confirmed. It depends very much on the (non-)volatility of the production load and demand for compute capacity. Certainly, it has the tendency to increase the number of sudden and steep rises and falls in power demand, which is not desirable.

PSNC has limited options for heat re-use. Some of the heat is used for office heating, but the production is largely exceeding what is needed. The generated waste-heat is too low grade. It is currently not feasible to directly use it in a district heating system. However, the Technical University's Campus is expanding and there may be an option to re-use heat there in the future.

One of the next challenges for PSNC would be to do without chillers. Several options are being studied, such as adsorption chillers in addition to solar panels, storage with buffers tanks for cold water, or - like CSCS does in Lugano - using cold water coming from a nearby river or lake.

8.2 Advanced Power Measurement and Control of the trinity Supercomputer – Andrew Younge, SNL

8.2.1 The Trinity system

The Trinity system is the first instantiation of a series of Advanced Technology Systems (ATS) of the US programme for Advanced Simulation and Computing (ASC). The Trinity system was delivered in 2016 and is a result of the collaboration between Los Alamos National Laboratory (LANL) and Sandia National Laboratories (SNL). The next ATS system is called “Sierra” and is currently being deployed at Lawrence Livermore National Laboratory (LLNL).

Trinity is a single Cray XC40 system with an Aries interconnect that contains two different Intel node architectures:

- A partition with dual socket Haswell (Xeon E5-2698v3) nodes: 32 cores with 128 GB RAM per node
- A partition with single socket Knights Landing (Xeon Phi 7260) nodes: 72 cores with 96 GB RAM plus 16 GB High Bandwidth Memory, per node

The Haswell nodes are well-suited for many existing codes and therefore are well-suited to satisfy mission needs in the 2016 - 2017 timeframe. The Knights Landing partition is significantly more capable and provides application developers with an attractive new target machine. In the Top500 list of Nov. 2017, Trinity was number 7, with only the Xeon Phi part participating in the HPL benchmark run. In addition to the Intel Xeon Phi CPU architecture, Trinity also introduces burst buffer storage nodes and advanced power management (APM) system software enhancements.

8.2.2 Investigations on Trinity - towards a Power API for measurement and control

The power draw of ATS class systems is expected to increase over time. Trinity is using 8 to 10 MW. Exascale ATS class systems are anticipated to be operational in 2023-2024 and to stay within the 20 to 30 MW envelope. While Trinity is NOT power constrained, future systems are anticipated to be, and therefore Trinity is now used to investigate how to best use and operate future Department of Energy (DOE) platforms in a constrained power budget.

As a modern HPC system, Trinity is equipped with power usage measurement and power control capabilities at various levels. The Trinity team focussed on energy efficiency from the user perspective, the perspective of specific HPC applications, rather than that of the overall machine or the datacentre. What was needed first and foremost was a framework for understanding power measurement capabilities for HPC systems. The Trinity team have developed an HPC measurement taxonomy of four levels, which is described by Table 5. Each level has trade-offs related to the amount of information and effort required in terms of:

- Data collection
- Detailed code knowledge
- Overhead
- Data sampling rates

The higher levels require more effort. HPC users should decide what effort is desirable for their application.

| Level | Description | In Trinity, provided / implemented, by: |
|-------|--------------------------------------|--|
| 1 | Job-wide aggregate information | Cray's resource utilisation Reporting (RUR) tool, which records aggregate statistics about each job, uses fully out of band node level sampling, without any impact on application performance |
| 2 | Period sampling of component sensors | Cray's out-of-band power monitoring infrastructure. Per-node power is sampled at 5 Hz and stored in a central database |
| 3 | Application instrumentation | KokkosP runtime profiling interface, and a Power API plugin for Kokkos. |
| 4 | Multi-level correlation | Kokkos is a layered collection of C++ libraries enabling performance portability of C++ applications and libraries across many-core architectures |

Table 5: HPC measurement taxonomy

8.2.3 HPC “mini-apps”

Rather than to apply the framework to real applications, three open source HPC mini applications - “mini apps” - were developed as examples of production workloads. They were designed to maintain computation and communication patterns similar to those of specific classes of real production applications:

- MiniMD: A molecular dynamics mini app representing LAMMPS. It supports Lennard-Jones (LJ) pair interactions, equivalent to LJ liquid simulations
- LULESH: An unstructured mesh Lagrangian explicit shock hydrodynamics mini app, representing DOE hydrodynamics apps, especially ALE3D
- MiniFE: An implicit finite element condition simulation, using a conjugate gradient solver on a rectangular problem. It represents abroad class of finite element applications

To evaluate power and energy measurement techniques, the three mini apps were run on both Trinity partitions, and additionally, on an older Cray XC30 system with an Aries interconnect and dual socket Ivy Bridge (Xeon E5-2595-v2) nodes, 24 cores with 64 GiB RAM per node. In a case study varying P-state (CPU frequency) on all platforms, all mini app workloads were run on 32 nodes from each platform, utilizing all cores on each node. The purpose was to establish the highest performing, lowest overall power consumption, or the least total energy consumed per Figure of Merit (FoM). The FoM is not simply the application run wall clock time, rather it is a figure that quantifies the performance of those application aspects that each mini-app was specifically designed for.

Several detailed examples presented show that the level 2 out-of-band high-frequency sampling of power usage of components is capable of showing different periods and phases of power usage, fluctuating around the aggregate level 1 averages. Level 3 application instrumentation requires more effort and knowledge. Functional regions of code must be specified by domain experts. Instrumentation implies inserting code to enable timing of region duration. The level 4 correlation effort tries to correlate and make sense of the structure in terms of code regions to patterns of varying power usage.

After gathering, per mini app, the power profiles that emerged from the testbed scale runs and analysing the data, the same three mini-apps were scaled to run on thousands of nodes on the Trinity Knights Landing partition. Power usage per node was evaluated as both P-states and number of nodes were varied. The mini apps constitute well-defined workloads for which weak scaling can be achieved. Most important for the Trinity team were not the outcomes as such, but the confirmation that the power profiles that were obtained at testbed-scale for the three mini-apps proved to match those obtained when scaling is increased to thousands of nodes. The impacts on FoM, of varying frequency and a varying number of nodes are different for different mini-apps. But the different impacts at large scale are expected, predictable, behaviour, in the light of the power profiles established during the testbed runs.

8.2.4 Lessons learned, guidance for HPC application power usage profiling

- Decide what scope is necessary: use the 4 level taxonomy
- If only rough average power is required, use aggregate information
- If application code knowledge is limited, out-of-band data collection is preferred (as it avoids overhead for the application), while timed in-band periodic counter sampling is

second best. Phases are readily discernible if adjacent regions of code happen to have power profiles that are not similar. On the other hand, access to out-of-band data may require root privilege.

- If application code region info is necessary and possible, then in-band profiling should follow. Small regions creating in-band overhead should be avoided. It is best to combine code regions with timestamps and correlate those with out-of-band data power usage data, however, this requires the most effort, code expertise and specialised hardware.
- The investigation should start at a testbed-scale first because this minimizes energy used in power profiling itself and simplifies experimentation. Small-scale power profiling can be directly related to extreme-scale runs, if weak scaling is achieved and confirmed.

8.3 Public (WTO) procurement of electricity: the CSCS approach – Ladina Gilly, CSCS

8.3.1 Background of CSCS energy procurement and the Swiss energy market

Switzerland has been a member of the World Trade Organisation (WTO) since July 1995 and it has been bound by the Government Procurement Act since January 1996. Public procurement of energy on an open market is possible since the liberalisation of the Swiss energy market for large consumers (contracts of 100,000 kWh or more) in January 2009. For several years however, the CSCS energy acquisition kept being arranged by successive standard contracts or privately negotiated contracts with regional energy providers. In 2017, the management board of ETH Zurich decided that energy henceforth should be acquired through public procurements following the rules of the World Trade Organisation (WTO) procurement process. Since then, CSCS did two WTO compliant procurements:

- 2018: a tender based on price, for a one-year contract concerning a single purchase of energy at a predetermined date
- 2019 - 2023: a tender based on a price formula, for a five-year contract concerning a multi-lot purchase of energy within the time frame

The basis for public procurement is the value of the contract in Swiss Francs (CHF). Procurements exceeding the following threshold values (excluding VAT) are subject to public procurement:

- Goods exceeding a value of 230,000 CHF
- Services exceeding a value of 230,000 CHF
- Building constructions exceeding a value of 8,000,000 CHF

In this context, the purchase of electricity is considered the purchase of a “Service”. CSCS procures electricity for a value around 3,000,000 CHF. The price of the “electricity service” comprises three components: the cost of the energy itself, transport costs (i.e.: the fees paid for network usage), and taxes. The relative share of these components in the price varies over time. In 2018, energy cost was 41%, network cost was 35%, and taxes were 24%. The taxes component includes all service specific taxes, such as contributions to federal and cantonal renewable energy funds, but excludes value added tax (VAT).

Independently of which WTO procurement process you select - Public, Selective, or Direct – they involve publication and an appeal period of 20 days. This appeal period brings with it a risk of

market price fluctuations between offer and contract signature that need to be managed. Who carries the risk of these price fluctuations, or how can they be better managed?

8.3.2 *Public procurement based on price*

For the year 2018 CSCS bought its energy in one single purchase on a predetermined date through a public procurement based exclusively on price. There are several advantages to this approach:

- The evaluation criterion is very straightforward: price
- Once the contract is signed, both buyer and seller are protected from future price fluctuations and have a firm basis for budget planning
- With a one-time purchase there is no additional administrative work, once the contract is signed

The main disadvantage is the risk of price fluctuations during the 20 days appeal period.

The seller may purchase electricity right away, at the time of the offer, in which case, they expose themselves to the risk of an appeal: if the appeal is granted, the seller could get stuck with energy that he may not be able to sign a contract for and has to resell at a potential loss. Alternatively, the seller may decide to avoid the appeal risk by purchasing once the contract is signed. If a price increases during the period, the seller cannot pass on the increase to the buyer, as he must guarantee the offered price. If prices have decreased, the buyer cannot benefit from the decrease as he must sign for the originally offered price. Whatever the strategy the seller chooses, any risks on his part will be assessed and factored into the offered price.

8.3.3 *Public procurement based on a price formula*

For the period 2019 -2023 CSCS bought its energy in a public procurement based on a price formula that allows it to buy energy in multiple lots distributed over time at a price determined by the formula. There are more advantages to this approach:

- The evaluation criterion is clear as well: price formula
- The buyer can take advantage of market fluctuations and reduce their risk of having to buy at a high point in the market by purchasing energy in lots distributed over time
- The seller does not have to bear the risk of price fluctuations or appeals between the award and contract signature
- The seller will not have to factor price fluctuation risk into their offer

There are some disadvantages as well. This approach requires the tendering party to define a price formula and to provide competing vendors with an energy usage profile of the last year. Furthermore, there is more work involved in observing the market and the buyer may wish to contract with a third party to ensure market observation in order to take advantage of price fluctuations.

The price formula used by CSCS contains **market-given components**, **components defined by the bidder**, and **calculated components**. In the price formula, $P_{j,t}$ denotes procurement price for delivery year j , calculated at time t :

$$P_{j,t} = \left(P_{j,t}^{baseload} \times \alpha + P_{j,t}^{peakload} \times \beta + P_{j,t}^{nordgrenze} \right) \times FX_t + \gamma_1 + \gamma_2$$

Market-given components:

- $P_{j,t}^{baseload}$ Price at moment (t) of the product “Phelix-DE Base load Year Future” for year (j) on the German EEX Market. Best Ask. Offer must be valid for 15 minutes
- $P_{j,t}^{peakload}$ Price at moment (t) of the product “Phelix-DE Peak load Year Future” for year (j) on the German EEX Market. Best Ask. Offer must be valid for 15 minutes

Calculated components:

- $P_{j,t}^{nordgrenze}$ Transfer fee, for transferring energy from Germany into Switzerland, for delivery year (j), at time (t)
- FX_t Exchange rate EURO / CHF

Components defined by the bidder:

- α Profile factor for base load
- β Profile factor for peak load
- γ_1 Premium for required quality of energy (e.g. green energy), in CHF/MWh
- γ_2 Premium for handling costs, balancing energy, risk, margin, in CHF/MWh

8.3.4 Conclusions

The main messages to take home from the CSCS procurement experiences are that, in case of a procurement based on a price formula, the provider is not exposed to price fluctuation risks between the times of awarding and signing the contract. It removes the risk of getting locked into a high-price purchase moment. Furthermore, it allows the buyer to take advantage of market fluctuations and even these out over time, thereby reducing the risk they pose. Finally, it provides a good budgeting basis as purchases for the year ahead can be completed in time for budget discussions.

9 Session VII – Technologies**9.1 Direct Liquid Cooling of Datacentre Equipment – Dominik Dziarczykowski, CoolIT****9.1.1 About CoolIT**

CoolIT is a company that manufactures and delivers liquid cooling solutions for HPC, hyperscale and enterprise datacentres. The goal of the company is to provide energy efficient cooling solutions at a reasonable cost. CoolIT's products are designed to capture the heat very close to where it is produced, i.e.: near the processor and the memory.

The Company, based in Calgary, Canada, has manufacturing warehouses in both Canada and China. It has sales representatives in Europe, North America, and Asia. In partnership with Stulz, a worldwide service network is maintained. The company has a broad selection of coolant distribution units (CDUs) on the market. CoolIT has many patents issued on liquid cooling technology and has deployed over 2.5 million units globally. Many of its products are directly sold from Dell and HPE but also other vendor customisations are available. CoolIT is a member of The Green Grid Association and HPC Advisory Council.

9.1.2 Datacentre challenges and Direct Liquid Cooling (DLC) Solutions

CoolIT identifies four main data centre problems that need to be addressed: energy, density, heat, and noise.

- **Energy:** Datacentres consume 1.5% of the world's electricity, currently about 272 billion kWh.
- **Density:** Most of the current processors have a power usage that is still OK for air cooling. Most of today's GPUs however, have such high power that they benefit - or would benefit - significantly from liquid cooling. The current trend predicts that tomorrow's processors and GPUs are all in the power range above 200 W, where liquid cooling is beneficial. In some cases, air-cooling will be very difficult or outright impossible. Also, the expensive real estate is driving datacentres to increase computing power per square meter.
- **Heat:** Datacentres waste about 400 trillion British thermal units (Btu) of heat every year, and even more heat comes from multiple CPUs and thin servers.
- **Noise:** AC power and server fans, used in air cooled solutions, because unsafe noise levels for workers.

The use of DLC facilitates better performance, better efficiency, and higher density. DLC can facilitate peak performance for higher powered or overclocked processors of 80 kW per cabinet. It enables significant reduction in total data centre energy consumed due to the use of more passive components - the CDU being the only active component left. CoolIT estimates cost savings from 20%-40% and 12-18 months in return investment. DLC can also enable 100% utilisation of rack and datacentre space.

DLC has been used so far mostly in supercomputers and for government, research and universities. Mainstream commercial datacentres have remained predominantly air-cooled. New users of DLC include private HPC users enabling high-end HPC applications for Artificial Intelligence (AI), Big Data and Deep Learning.

9.1.3 Switching to DLC

Switching to liquid cooling in a datacentre following the CoolIT approach consists of four phases: assessing, planning, deploying, and maintaining.

9.1.3.1 Assessing phase

In the assessing phase, the cooling method and the location specific temperature values need to be selected. Free cooling and adiabatic cooling are the most efficient ways to cool. The choice between a cooling tower, a dry cooler or an evaporative cooler must be made based on the water temperature needs and the specific temperature limits.

Checking of the room layout, under floor spaces and the collection of environmental data, like humidity and temperature gradients are included in the assessment phase.

9.1.3.2 Planning phase

In the planning phase server models, CPU types and DIMMs are selected for liquid cooling purposes. For the target heat capture percentage CoolIT offers the choice between three options:

- DLC for the CPUs only, resulting in a 60-70% heat capture
- DLC for CPUs + voltage regulators (VRs), resulting in 65-75% heat capture
- DLC for CPUs + VRs + memory, resulting in 80-90% heat capture

The serviceability of I/O expansion cards or hard drives is difficult to combine with DLC, since these devices need to be hot swappable – therefore such devices are air-cooled.

Planning includes system design and component selection. System design is quite comprehensive and includes the definition of fluid connection requirements, the determinations of the number of nodes required, chip-to-atmosphere system design, thermal and flow analysis, rack layout. Component selection is modular and deals with the choice of cold plates, the choice of rack manifolds for fluid transport, and the choice of the coolant distribution unit - the only active cooling component in a CoolIT system.

Cold plates are mounted onto the CPUs. They are made of copper with micro channels that increase the cooling surface and have patented Split-Flow technology for high-performance cooling. Tube connections on cold plates have multiple positions for versatile fitting onto motherboards. The CPU mounted cold plates can be extended to capture VR heat. There are 130 fins on each cold plate and the thermal resistance as low as 0.035°C per W.

Addition of cold plates for DIMMs allows up to 80%+ server heat capture. The DIMM cold plates maintain serviceability of DIMMs without needing to remove cooling components.

Manifolds connect server level components to CDUs with manual or automatic stainless steel and dripless connections. Manifolds can be horizontal or vertical and are mounted in the rack like a PDU. Sizing is optimised to flow and pressure drop.

CDUs provide pumps to generate the flow of the coolant. The CDUs have sensors for temperature and pressure. Collected data can be moved to a monitoring system. Sizing ranges from a 2U rack mount model of 40 kW/40 nodes and 4U of 80 kW/100 nodes, up to a row based 700 kW+/800+ nodes model which has 3 pumps and power redundancy.

Last but not least, the planning phase includes checking the logistics of installation and commissioning, and establishing an SLA and a preventive maintenance plan.

9.1.3.3 Deploying phase

The deploying phase consists of the installation process and involves Stulz and possibly local partners.

9.1.3.4 Maintenance phase

Planned Preventive Maintenance (PPM) consists of two check-ups for the first year - three months and eight months after the installation - and subsequently twice annually, for the lifetime of the system. The quality of liquid coolant is analysed and system checks are performed to verify proper operation.

9.1.4 Case studies

CoolIT presented two case studies: an installation at Poznan PSNC and one at the Centre for Biological Sequence Analysis in Denmark.

The Poznan installation has 1200+ servers from Huawei with CPU+VR+DIMM DLC. Of the total heat produced under full IT load, 85% is cooled by liquid, running with 75% fan speed reduction. The rack density is about 30 kW per rack. The inlet temperature of the water is about 40°C. Heat recovered of the returned water is used to heat office space.

The Centre for Biological Sequence Analysis of Denmark Technical University (DTU) has a containerised datacentre that is part of “Computerome”, which ranked #121 on November 2014 Top500. The required inlet water is over 40°C for maximum heat re-use. Adjacent buildings as well as the nearby town of Roskilde use the waste heat of the system. The cluster consists of HP ProLiant servers with dry quick disconnects for hot swappable servicing.

9.2 Modular Datacentres – Jean-Marc Ducos & Frédéric Souques, CEA

The CEA Supercomputing complex contains two supercomputing facilities. The presentation focused on the evolution of one of them: the “TERA” facility. The TERA facility was designed for three generations of supercomputers foreseen in 1999. The design was based on the expected needs of these systems: 6 MW of power supply, 1200 m² floor space for supercomputers of two generations + 750 m² floor space for storage, 1 to 2 tons of floor load per square meter, and 3 MW air cooling capacity.

The building and infrastructure delivered proved adequate for housing the TERA-1 system, which was installed in 2001 and built by HP (Dec/Compaq). The Tera-1 had a compute performance of about of 5 TFlop/s peak. It entered the Top500 at rank 4. At full load 0.6 MW of heat was cooled by air, coming up from under the raised floor.

In 2005 the TERA-10, a system of about 60 TFlop/s peak, using about 1.5 MW of power, was delivered. The 3 MW of available air cooling capacity turned out to be insufficient for accommodating both the TERA-1 and TERA-10 system. A new small auxiliary building was needed for power supply. In addition, a new 2 MW chiller and a cooling tower had to be built. Since the TERA-10 used cold doors in addition to air cooling, a water loop had to be built too.

The TERA-100 system was delivered in 2010, with a compute capacity of about 1250 TFlop/s peak, using about 5 MW of power. But the 6 MW turned out to be not enough for operating the TERA-10 plus the TERA-100, plus the facilities. For TERA-10 the power density was less than 10 kW per rack available, for the TERA-100 it was about 40 kW per rack. The subsequent facility upgrades included a new power distribution and a chiller, and an upgraded chilled water distribution with cooling towers.

Since 2016-2017 the TERA facility is hosting the 25 PFlop/s TERA-1000 system, which uses about 5 MW of power and is cooled by Direct Liquid Cooling (DLC). This time the facility upgrades included a warm water loop with new cooling towers and a distribution system to enable the reuse of the system’s waste heat for the CEA site. In addition, the weight of the racks was above 2 tons per square meter, which implied that the raised floor as well as the underlying concrete floor needed to be reinforced.

The evolution portrayed above all shows that it is hard to predict the specifications of the future computers. Each step in the evolution takes time and money spent on infrastructure upgrades. Many times the upgrade work has to be started before there is exact knowledge of the new computer and its specifications. The upgrade work on the facilities is further thwarted by the fact that the former computer is still in production and has to be kept available. This situation results in a lot of constraints on the way in which the facility upgrades can be executed.

The above listed difficulties have led CEA to look for a modular hosting infrastructure for the next generation EXA-1 system. The principles of modular construction involve building modular blocks

for IT, cooling, power and logistics. This flexible and scalable solution is booming in the world of very large scale - though not geared towards HPC - datacentres deployed by service providers like Google, Apple, Facebook, and Amazon. The modules are standardised and therefore quickly installable. Design can be done per slice, integrating both IT equipment and technical infrastructure. This optimises energy efficiency since the sizing of the infrastructure can be built to match the computer needs and allows easier upgrades or extensions. Each block is independent of the others.

Modular solutions for HPC are for examples available from ATOS/Bull, which have developed “Mobull” containers - based, at the beginning, on the recycling of sea containers. Modules can be built, quite literally, on top of each other, forming an independent slice around the computer module. The module can be tailored to the needs of the computer module. One module takes about 18 months from order to operation.

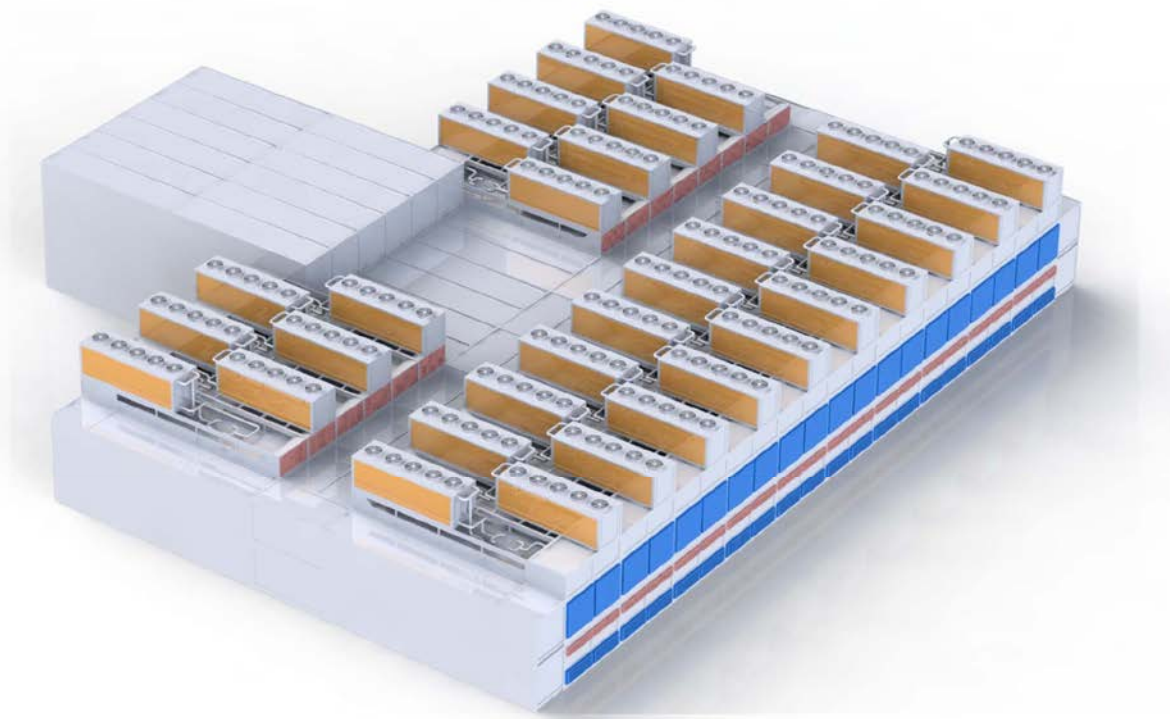


Figure 16: Impression of a Mobull containerised modular datacentre

A basement module is available for power distribution and uninterruptible power supplies (UPS). On top of the power module is the IT module for the computer hardware and warm water cooling, and cooling tower module can be stacked on that. Free cooling modules can also be on the side of the 3-module stack. Warm water cooling is used for the supercomputer. Service racks can use air cooling.

The power of the current slice, consisting of two containers plus a cooling tower module is about 2MW. Slices can be connected together and changed based on needs for, e.g., different row topologies. A logistics section can be added as well. Figure 16 shows a possible combination of multi MW HPC datacentre, consisting of many slices and an integrated logistics section.

As systems become denser, old buildings become less suited for HPC systems. The renovations are costly and complex to perform. CEA regards modular design solutions as a very interesting, cost-effective, flexible and easily upgradable option for implementing an HPC datacentre.

9.3 Facility Commissioning Experiences for ORNL's Summit, The U.S. DOE's 200PF Supercomputer – Jim Rogers, ORNL

9.3.1 Summit compared to Titan

Oak Ridge National Lab's (ORNL) large new supercomputer installation is called "Summit" and it will replace "Titan" as the Oak Ridge Leadership Computing Facility's (OLCF) flagship system. Summit is going to have more memory than Titan, per node as well as in total, and fewer, but much more powerful, nodes. As a hybrid CPU/GPU system, about 90% of the Flop/s come from Summit's GPUs, with two CPUs (IBM Power9) and six GPUs (NVIDIA Volta) per node. Application performance is supposed to range between five and ten times the application performance that Titan could deliver. The interconnection network is going to be faster, with a much higher bandwidth between CPU and GPUs, and with a different topology: where Titan had a 3D torus, Summit comes with a non-blocking fat tree. The GPFS scratch filesystem is going to be faster and larger compared to Titan's Lustre file system. The peak power consumption for Summit can be as high as 15 MW, compared to the 9 MW of Titan.

9.3.2 Progress of preparations for Summit

Summit will be in full production by 1 January 2019. The installation has already been completed for most of the facility systems and computer cabinets, storage, and interconnect. The actual installation brings together all the needed bits and pieces to form the entire complexity of a water cooled computer room on a concrete slab. The installation of the hardware, from the first cabinet to the last part of the storage system, has taken 9 months to complete. Of the first part of the system, about 25% was installed in December 2017, which has forced early on testing of electricity and the cooling systems. Also, much of the software stack testing had to be done early. "Sleep is highly overrated at ORNL", as still in progress are full system load testing with cooling optimisation, storage and filesystem acceptance along with system acceptance of 75% and 100% of the compute capacity before full production can start.

Last year's talk for this workshop on the preparations for Summit was about installation "on slab", without underfloor space, and on building all of the infrastructure for Summit. The transition, from last year's talk with 2D space drawings, to the actual installation, with 3D drawings, has revealed much more layers into the facility. Last year there were 5 transformers, 256 computer racks (+1 test system rack), 18 InfiniBand switch racks, 44 storage racks (250 PB, 2.5 TB/s aggregate bandwidth) and 5 infrastructure racks in the 2D drawings. Today's 3D drawings include 15 layers of equipment, from ceiling level thermostat heating (RH) control cabling and sprinkler systems, to slab level leak detection and drainage systems. Since there is no raised floor space, all infrastructure connections have to come from above the computer racks. Above the slab there is 5.5 meters of space to fit in all the needed infrastructure, along with the computer system itself. The water cooling and electricity are kept apart in rigorously separated wet and dry aisles. The system has about 80,000 water connections in it, and a leak detection system on the floor as well as under the cable trays. There is a wet sprinkler (that hopefully never will have to go off) and a Very Early Smoke Detection Apparatus (VESDA) installation.

9.3.3 “12-15 MW, say what?!” - Predicting impact on the local power grid

The ORNL site in total has a peak power signature of about 85MW. The power provider, Tennessee Valley Authority (TVA), needs to adjust to the power fluctuations and forward knowledge of the impact helps in preparation. Since Summit’s peak power can be 12-15 MW, the impact of this system alone for the local power grid can be big. While Summit is being tested, Titan still is running production and ORNL has also additional labs that need a lot of power. ORNL used its own Dynamic Protection-Planning Simulation Model to estimate the impact on the branch power flows and the apparent impedance sensed by the step relays in the TVA system. The disturbances simulated were in the 20-30 MW range. Branch power relays seemed to be able to handle the load swings and smooth out the ripples of power fluctuations.

One key components of the power infrastructure are the transformers. ORNL uses K-rated transformers with a K-rating of 4 for their HPC loads. These transformers have a larger core cross section to compensate for increased flux density and they use a heavier conductor and a double sized neutral bar to deal with increased heating. The goal of using these transformers is to protect against non-linear loads and give a better management of winding temperatures, which leads to increased service life.

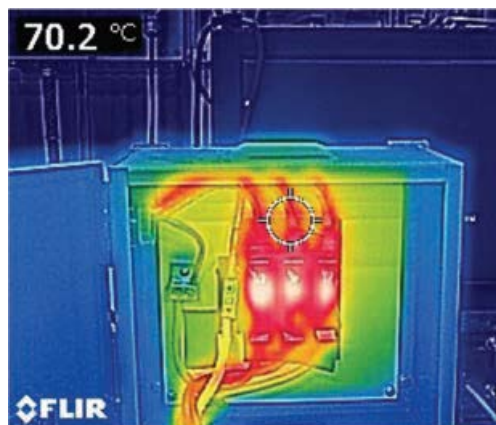
The best practices of ORNL for electrical distribution can be summarised in five key points:

1. The usage of the already mentioned K-rated transformers
2. The usage of 3.0/4.0 MVA oil-filled unit substations. These provide higher efficiencies, quieter operation, more overload capacity and more transformer per square foot
3. The usage of aluminium windings instead of copper. Aluminium windings have a better thermal capacity, their tolerance to surge and overload current is better and they have a lower weight
4. The usage of FM-approved transformer fluid to minimize the potential damage from a fire;
5. The installation of transformers as close to the computer systems as possible This minimizes losses, copper cost and helps to regulate voltage

9.3.4 Problems found during preparation and testing

A test revealed some problems with the quality of the materials used. Some failed to meet the desired specifications.

9.3.4.1 High fault currents in power distribution units



Infrared cameras were used to identify potential installation problems where elevated heat signature can reveal incorrect or under-performing terminations (see Figure 17). ORNL calculated fault current significantly higher than the 10 kA provided in the manufacturer’s PDUs. Depending on rack location, and distance from the transformer, fault current could be as much as 55 kA. ORNL decided to use ORNL custom designed, NRTL tested T-series rated fuses that provided 100 kA fault current protection. Tests for fault current with 9000 amps fuses showed that they reacted in the first millisecond.

Figure 17: Detection of high fault current PDUs

9.3.4.2 Failing fan blades

ORNL bought transformers with fans and runs these fans 100% of the time. Fan failures have occurred with blades breaking off and flying around inside the transformer. There have been no arcing incidents yet, but this is a major concern. The vendor comment was that the failures should not have happened, but were “statistically not significant” given the number of fans and the number of hours they ran. But ORNL have had several different fans with the same problem and hold that it should not be unreasonable to expect the fans to operate full time. Apparently, the fan blades vibrate, which causes stress to the blade connecting to the hub which is fixed to the motor shaft. The complete fans are stamped from a single piece of aluminium. The original design, with a steel hub fixed to the motor shaft and aluminium fan blades riveted to it, was much more robust.

9.3.4.3 Nearly failing circuit breakers

4000 A 480 V circuit breakers running for several years with 2700 A and 100% load had a maintenance break. After the break contact resistance in one of the breakers appeared to have creased from 30 to 180 micro-Ohms. When the contacts were examined, a substantial portion of the non-metallic parts in the contact space had degraded so much that they broke up and fell to the bottom of the breaker (see Figure 18).

The breaker did not fail, but would have probably failed soon. Theories as to why this happened include that the vendor used materials not suited for temperatures inside the breakers, that there was insufficient ventilation in the breaker compartment, that there was a bad connection in the breaker, overheating, or a combination of the above.



Figure 18: Almost failing circuit breaker

To avoid learning about such problems too late, ORNL recommends installing temperature and infrared monitoring equipment in the breaker compartments and to add exhaust fans in top of the breaker to remove heat.

9.3.5 Mechanical Facility

The mechanical facility for Summit is built for 20MW and an inlet temperature of 20°C for the main cooling loop. It has PLC (Programmable Logic) for infrastructure automation control. The facility has 7700 tons of evaporative cooling. The pumps and heat exchangers have N+1 redundancy and the system can circulate 17,000 litres of water per minute. ORNL is located in the southern part of the US, where 100% of evaporative cooling works about 70% of hours per year.

During the spring of 2018 the PUE has been mostly between 1.03 and 1.1. Infrared monitoring has been used to monitor heat management. The goal is to feedback monitoring into the regulatory system and provide variable flow and constant temperature. Thus, monitoring helps pump

optimization. Infrared monitoring has also revealed leakage of heat into the computer room caused by insufficient air management in the Mellanox InfiniBand director switches. Monitoring also shows how the RDHXs (Rear Door Heat eXchanges) adequately capture residual and parasitic heat load from compute cabinets, DIMMs etc. Since the system is on a concrete slab, there is no additional cooling to help with parasitic heat, only humidity control.

Every device in the room has a rear door heat exchanger to capture all possible heat. CPUs and GPUs use cold plates to capture 73-74% of the heat load direct to water. Fans blow through power supplies and DIMMs to the rear door heat exchanger, which receives water at 20°C, and thus deals with the rest of the heat. Very little heat escapes into the room.

The tight specifications of the system's CPUs and GPUs are tested with the help of runs of the HPL benchmark. CPUs and GPUs that are not within the standard deviations of the contract specifications are sorted out.

9.3.6 Challenging GPFS scalability

The GPFS filesystem is huge and uses a new enclosure from Seagate, named Geo4, which includes both spindles and burst buffers. GPFS has been a very reliable filesystem, much more robust than Titan's Lustre, but not at this scale. It appears to be one of the most difficult things to get right. IBM has proposed to split the storage capacity into two file systems, but ORNL insisted on having a single scratch file system. A lot of co-operation between GPFS development staff and ORNL is needed to build the largest GPFS installation, with several modifications to the software stack to enhance scaling.

10 Session VIII – Vendor Panel

10.1 A flexible approach for Exascale infrastructure providing best electrical and cooling efficiency for all types of nodes and interconnect fabrics – Jean-Phillipe Sibers, Atos/Bull

Atos has developed the Sequana X1000 DLC system with the sight set on exascale. The basic building block is a cell that embeds the management network and hosts several different types of compute elements. About 75 cells (approx. 20,000 nodes) have been shipped so far, to multiple customers in Europe. These are equipped with configurations of Xeon, KNL, and V100 GPUs for compute and with InfiniBand EDR or Bull BXI for interconnect.

Experiences from current systems in the field are that:

- The open and modular architecture is good. The number of node types is increasing and the cell concept improves scalability
- More flexibility is needed in interconnect fabric topology. Currently, only a two-layer fat tree is possible, while for some applications a generalised hypercube is preferred
- Power flexibility is needed to accommodate a wide range of compute nodes with compute blade TDPs ranging from 1000W to 2200W
- Cell weight is at its maximum for current datacentres, some deployments have needed to reinforce floors

Plans for a future version of Sequana (S2) include:

- More interconnect topologies, including generalised hypercubes. These will be realised by using pluggable QSFP connectors instead of large backplanes and allowing horizontal cabling between racks
- The use of scalable 3kW N+1 redundant power units from 30 kW up to 90 kW per cabinet, with up to 63 A per incoming power cable
- A fully hot water cooled fan-less design. Compute nodes, power supplies, interconnect and ethernet management switches will all be liquid cooled. 97% of the heat will be captured using water. Even an insulated rack will radiate some heat though and acts as a heater for the room. Some room air conditioning will still be needed, not for the IT equipment but for humans working in the datacentre
- Heat exchanger integrated into the rack
- No separate networking and compute racks, everything will be integrated into one single type of rack. One consequence is that all racks have the same weight, lowering the maximum weight per rack. Target is 2,096 kg/m² on average per cell with a maximum of 2,100 kg/m² for a single rack within the cell

For very large datacentres the modular container based Mobull system is a possible option. This packaging can be used with both air-cooling (<40 kW/container) and DLC (<90 kW/container), even with both types of cooling mixed within a single container. For HPC the DLC version can be used to densely pack compute nodes since it removes the need for air in- and outlets. While the lifetime of the container itself is limited to the lifetime of the system within, it does not require large changes in the datacentre building and infrastructure when replaced and thus the Mobull system aims at lowering capital expenditure.

10.2 Preparing datacentres for Exascale – Vincent Pel, Cray

Floating point performance has been steadily increasing over time, Cray systems were the first to reach 1 GF (1988), 1 TF (1998) and 1 PF (2008). However, the plan to reach 1 EF in 2018 was a bit too optimistic. Now the talk in the trade is about “exascale”. The exaflop systems will come later and require substantially more power than the 20 MW that were initially expected, or rather, hoped for. Currently there are only a few applications that will be able to efficiently utilise the floating point compute capacity of the biggest systems. Most applications need a lot of work on the software side before they can reach exascale performance.

Exascale systems target a broader range of applications, independent of floating point capabilities. Improvements in many areas are needed for exascale systems. The price of memory is becoming more important since the drive for higher memory bandwidth requires an increased amount of memory channels which in turn leads to more DIMMs. High bandwidth memory modules are both smaller and more expensive. Combined with higher core counts this requires applications to use memory in a different way.

Comparing, as an example, Blue Waters (11.5 PFlop/s peak) and Sunway TaihuLight (125.4 PFlop/s peak) shows that. The difference in the number of cores and in floating point peak performance is huge, but both systems have the same total amount of memory: 1.3 PB. To use the larger system effectively the applications must be able to cope with a lot less memory per core.

Interconnect latencies have not improved that much since the Cray T3E days when looking at bandwidth per core. Software changes are needed in order to reach exascale. Hardware alone cannot solve this. In effect systems are getting wider, not faster.

Denser racks have as a side effect that cables are becoming shorter and cheaper. Racks are getting heavier though, so new data centres should plan for 2000+ kg/m². An “entry level” exascale system will require around 10 MW power, and, combined with dense racks, will require DLC. Not only the processor TDP needs to be taken into account but also the chassis temperature. Using warmer water may restrict which CPUs can be used.

Energy efficiency requires constant monitoring and control to optimize for the requested power and time budgets. Re-use of the energy is easier with high return temperatures, but return temperatures may be lower than possible and desirable for optimal heat re-use due to other optimisations.

10.3 The Future of HPC at HPE – Torsten Wilde – HPE

Torsten is on sabbatical from LRZ to work with exascale systems at HPE, and can compare how it is on both the buyer and vendor side. More manpower is needed to design and build systems than most buyers realise. His presentation focused on the way power and cooling work.

HPE, in collaboration with the National Renewable Energy Lab (NREL) and Power Innovations, have built a fuel cell test system. This system will be installed at NREL in May 2018 with long-term testing beginning in June. NREL will be using renewable energy to power an electrolyzer that generates hydrogen to store the energy. A fuel cell can then use this hydrogen to power a cluster.

Torsten compares the Tsubame 3 system with the first SuperMUC system; many changes have been made to improve the GFlop/s per Watt ratio. Technology in this space has been moving fast, but there is still room for improvement.

Denser racks with 100% DLC will reduce the footprint of future systems since no hot or cold aisles will be needed. HPE is investigating fan-less systems that are optimized for chiller-less cooling and with potential for heat re-use. Using high voltage direct current to power systems is also likely. How systems are placed in the datacentre in the future will be decided by cabling, power and cooling.

Monitoring of exascale systems will generate large amounts of data per second, mostly from the fabric. Raw data is important, but needs to be processed to be manageable. Data that is operationally useful from both the IT and the facilities sides needs to be merged into one management system for analysis and control.

An overview of HPE system development work was made, noting the technologies demonstrated by Apollo 8000 (dry disconnect and leak detection) and The Machine (memory centric computing).

Near-term work is focused on collaborations for producing an ARM based system. These ARM projects are not only about the hardware, but also making sure that the software ecosystems exist when the hardware hits the market. Longer-term development focuses on the Gen-Z fabric, aiming at replacing many of the interconnect networks used today. Gen-Z also allows memory to be directly attached to the fabric and accessible by all devices. HPE is working on Gen-Z chip development, both on the silicon and optic side.

10.4 Liquid Immersion Cooling: A versatile cooling environment – John Wagner, Fujitsu

Denser node packaging combined with higher CPU TDP leads to higher requirements for cooling, driving the need for developing new cooling techniques. Fujitsu is exploring liquid immersion cooling (LIC) for systems in the 30-70 kW/rack range. The technology was first shown in 2016 and is now a Fujitsu product that is available on the market.

Supercomputers using LIC are not new. The Cray-2 already used LIC in 1985. Development of LIC stalled afterwards, but the pace has picked up again in the 2010s with systems developed by multiple vendors. Fujitsu uses the coolant Fluorinert from 3M that is non-conductive, non-flammable, chemically stable, and non-toxic. Its ozone depletion potential (ODP) is zero. Pricing for the liquid is around 120 USD/litre.

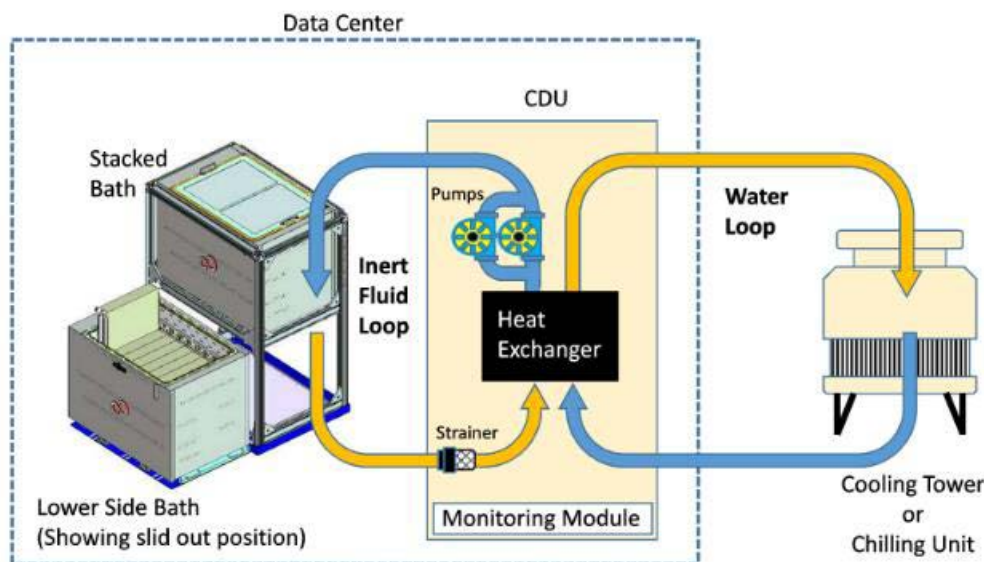


Figure 19: Stacked immersion bath in a rack, sharing a single CDU

The Fluorinert immersed nodes are almost standard Fujitsu Primergy x86 servers. The differences are that the chassis-fans have been removed and that BIOS fan monitoring has been turned off. Nodes are totally immersed in the liquid, including cable mounts. Maintenance is the same as for air-cooled systems after the node has been removed from the bath. Nodes are normally mounted top down, except for disk chassis that are mounted top up. To remove the heat from the bath, a coolant distribution unit (CDU) is used, this has pumps and a heat exchanger connected to a water loop. The inert fluid in the bath captures the heat that is generated by the submerged equipment and transfers it from the bath to the CDU through an inert fluid loop. Subsequently, the heat exchanger within the CDU exchanges the heat with the water loop and expels it outside the datacentre, via a cooling tower, a chilling unit, or whatever cooling mechanism is used at the datacentre level. A rack equivalent is being developed that consists of two baths, stacked on top of each other for a total capacity of 40 U combined with a single CDU (See Figure 19). Current configurations have an upper limit of 45 kW/bath and 60 kW/CDU.

Power savings come from a combination of not running any fans inside the equipment and CDUs being more effective than computer room air handling units (CRAHs or CRACs), because the heat

transfer capacity of the liquid is much higher than that of air. For an enterprise workload there is space savings expected due to higher node density, but HPC nodes are normally very dense already.

As a side benefit of the immersion, the nodes are well protected from the environment. Dust and other contaminants in harsh environments are caught by the liquid. The datacentre itself does not need not to be cooled for the sake of the machines and can instead be heated by the systems. Having a good floor that can sustain the weight is more important than the rest of the building structure, in principle the baths can be placed in a garage. Due to the absence of fans, the system also produces little noise.

This currently marketed system is version one. Improvements planned for future versions are:

- Raising the cooling capacity of a single CDU from 60 kW to 200 kW, thus enabling the sharing of a single CDUs between up to three baths
- Validation of more system components, such as Nvidia V100 GPUs, servers with Cascade Lake CPUs, and Fujitsu DLU AI servers
- Racking system, enabling two baths per rack

40 systems have been sold so far. Operational experiences are that with Fluorinert, different from other coolants, cabling has no problems being immersed. They do not stiffen over time, as has been reported for some oil immersed cables. The baths have the possibility of attaching a network switch externally. Not all disk drives are 100% sealed, so only validated drives can be used, but with the current trend of replacement of spinning drives by SSDs this is a temporary issue.

10.5 Behind the Scenes of Efficient Cooling – Axel Auweter, Megware

Megware is a small company, employing around 50 people at the moment. Since 1990 it has delivered 1200 systems. The company is not only a system vendor, but also operates a small datacentre that is used for staging new systems before delivery, benchmarking of systems and also for co-location of customer systems. While this datacentre is much smaller than most HPC sites, it still enables Megware to test cooling solutions in a realistic setting and see many of the same problems. Currently the datacentre has 200 kW power and a cooling capacity of 80 kW, which is being expanded to 160 kW.

Air cooled machines are still requested by customers and need to be tested, so the datacentre is equipped with computer room air conditioning (CRAC) units. At the moment this requires Megware to run the main water loop at 10°C, which is not optimal. The CRAC units have valves regulating the flow, and these can get stuck in their 100% position. Running the water loop at 10°C ensures that they do never reach the 100% level. Work is ongoing with the vendor to fix this, but it is an example of problems only found in real-life conditions.

ColdCon is the name of the Megware DLC solution, which has been evolving since 2011. The current generation was introduced in 2017 and is the first to cool the entire nodes and network switches with liquid. The Xeon Phi based CoolMUC-3 system at LRZ is a reference installation for Megware. A high-frequency (100 Hz) in-band power measurement system was developed to integrate with the node management.

Axel went through the basic physics behind heat transfer. Different materials have different thermal conductivity, with copper being an example of a material that is good at transferring heat. This is only one of the factors. Surface area, transfer distance, and delta-T also matter. Copper has a low

thermal capacity, however, so it is not suitable for storing heat. Water has the opposite properties, so combining them gives a good system.

Heat sinks can be designed in different ways to increase the surface area. Megware started with basically an empty copper box that water flows through, but the immediate surface area is bounded by the CPU size. Later Megware designs introduced fins to increase the surface area, which is easy to manufacture using a computer numerical control (CNC) machine. Having too many fins, however, result in narrow channels that can clog up if there are impurities in the water. Currently, a layered honeycomb pattern is used; this 3D inner structure gives a large surface area while retaining wider channels for water flow and keeping the weight down (see Figure 20). This type of structure would be ideal for 3D printing, but unfortunately copper as a material is not ideal for that. Instead, a chemical process is used during manufacturing to create a solid piece of copper by melting the layers and fusing them together.

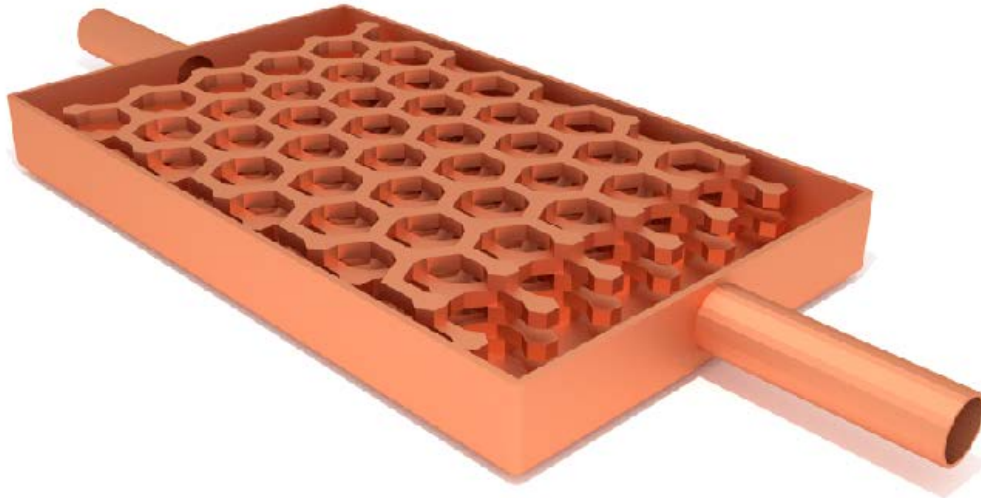


Figure 20: Megware CPU heatsink design with layered honeycomb pattern

DIMMs have separate copper plates, and a cold plate covers other components on the board. The thermal conductivity of the printed circuit board (PCB) itself contributes to dissipating the heat due to internal copper layers, and should not be underestimated. Heat sinks are designed to be large enough to cope with hotspots on CPUs and GPUs.

While the focus of this talk was component cooling, many other things need to be taken into account in DLC design: water quality, materials used, designing the system for maintenance and integrating with the datacentre infrastructure are a few examples.

10.6 Water cooling technology for the Exascale computing era – Karsten Kutzner, Lenovo

Karsten started by reminding the audience that using liquid to cool systems is not new. His first job at IBM included dismantling a water cooled mainframe. Plenty of plumbing was involved in that system due to the power density of the bipolar components. The advent of CMOS technology reduced the need for liquid cooling, because of its lower power consumption, enabling air cooling. But now, due to several factors leading to more concentration of heat, the industry is back to liquid cooling again.

There are many reasons why it is important to care about power and cooling:

- datacentre limitations
- electricity cost
- relation between performance and power consumption, increased TDP per socket
- application diversity, CPU bound and memory bound applications have different power requirements
- re-use of heat, either directly or indirectly, is important and benefits from high output temperature

Lenovo divides the work into Hardware, Software and Infrastructure with the focus of this presentation being on hardware. The infrastructure work is done together with partners.

The Lenovo ThinkSystem range is based on dense 42 U racks for both air and water cooling configurations with up to 72 nodes and six switches per rack. The design supports 66 kW/rack with three non-redundant PDUs. It aims at supporting 300 W per CPU socket. Potentially the rack can be equipped with four PDUs with 88 kW capacity. A fully loaded rack will be quite heavy, 1800 kg/m² and a point load for the feet of 6.6 kN when placing racks in a row. Floor tiles must be able to cope with this load; tiles with cut-outs have to be taken into consideration.

The return on investment for DLC versus rear door heat exchange (RDHX) differs substantially between a brand new purpose built datacentre and an older one that is retrofitted. For a new datacentre the investment in DLC will pay for itself within a year. In other cases the electricity price is the main parameter.

Cost of liquid cooling has come down with current designs. The extra capital expenditure for DLC compared to an RDHX solution is estimated to have shrunk to about 2.5%. Lenovo offers both technologies. However, if the power density is above 30 kW/rack, RDHX is not feasible and DLC is needed. DLC is also the only solution when higher water return temperature (40°C and above) is desired for more efficient heat re-use. On the current Lenovo DLC node design (SD650) the only remaining air cooled part is the power supply unit on the chassis (PSU). The message that the PSU must be liquid cooled as well, has been received loud and clear, most notably from their current customer LRZ.

Having water return at a temperature of 60°C is usually only achievable for HPL runs. Typical production workloads will not reach that. To increase the delta-T between the inlet and the outlet, and stabilise it at a desired temperature, the flow must be regulated to adapt to the actual heat produced by the computational workload on the system. Water quality is an issue. Definitely, filtered water is needed. A flow of about 50 litres/minute flow for a rack can be expected.

Accurate power readings of the hardware, on both node and chassis level, are needed to enable fine-grained job statistics. These are required both for reporting and control, since controlling the energy usage of jobs at runtime is getting integrated into the software stack.

10.7 Vendor Panel Discussion

The vendor panel consisted of Torsten Wilde (HPE), Karsten Kutzer (Lenovo), Jean-Philippe Sibers (Atos/Bull), Axel Auweter (Megware), John Wagner (Fujitsu), and Vincent Pel (Cray). Ladina Gilly (CSCS) was the moderator and to start the discussion gave the panel two questions:

1. What is “exascale” going to bring, and what should customers, responsible for hosting your future systems, be particularly worried about and prepare for?
2. What will the power consumption of the first and second generation EFlop/s systems be?

HPE: The first generation system will be in the 35-40 MW range with the second generation at 25 MW. System weight will be probably an issue for most current datacentres. Possible solutions are being explored by HPE, but a raised floor may no longer be feasible for all systems.

Lenovo: What will be the definition of an exascale system? Will a conventional peak performance measurement, such as Linpack still be used as a decisive criterion, even if real production applications will never reach that level utilising the total floating point capacity of the system? Does it make sense to keep this definition? If so, then a first generation system will need power in the 60-80 MW range.

Megware: Being a small company nobody will expect us to come with an exascale system soon. But we will probably be building 100 PFlop/s systems in the same time frame as large system integrators will build exascale systems. The question for us is whether 100 PFlop/s systems will simply be a tenth of an exascale system, or a totally different system that employs different technologies. Vendors need to be flexible, customer requirements will be driving this. New architectures and platforms are coming.

Fujitsu: What is an exascale system? Must it be based on standard general purpose CPUs? Is a GPGPU based system acceptable? How much of the compute power comes from accelerators? The answer to these questions will have a significant impact on power consumption. A first generation system built entirely from general purpose CPUs might indeed need 80 MW.

Comments from the audience:

- But who can afford 80 MW? GPGPUs and other accelerators must be acceptable.
- Accelerators can provide exascale performance earlier for some but not all applications. Many artificial intelligence workloads, for example, can use half precision. It is important to be able to have different options in the same system.

Atos/Bull: Systems need to be as flexible as possible, with regards to which nodes you can fit. Fat nodes and nodes including more and more accelerators in a single node are requested by more and more customers. This will also influence current programming models. Exceeding a rack weight resulting in more than 2000 kg/m² for most customers is not possible.

Cray: GPGPUs are needed, otherwise exascale will indeed be pushed much further into the future. Many promising technologies exist, systems need to be open to different options. Customers have different demands, some want a small footprint and dense systems and others can have large systems in which the compute power is more spread out. The available floor space may differ. The price of electricity varies in different countries. Suitable cooling depends on location and climate. Flexibility is key to making systems usable for all these use cases.

Remark from the moderator: You indeed will need to be able to cater to a plethora of different needs, not just because of differences in climate, price of electricity, or datacentre characteristics. Even in this audience we have quite a heterogeneous representation of sites with different needs for their future machines. All of them do HPC, but some can be specialised more than others.

National or regional centres often also have to accommodate the very “long tail” of HPC applications somehow. Many of those applications cannot readily be adapted, or even do not want to be adapted, to GPGPUs and other accelerators.

Question from the audience: *Will DLC lead to fewer opportunities for system upgrades?*

Megware: No. The cooling architecture is more complex than air-cooled systems and rear door heat exchange systems, but serviceability remains a requirement. Components in DLC systems will still have a failure rate, so access to, and replaceability of, components must and will be available.

HPE: No, but CPU upgrades sometimes require large changes to the surrounding system. Gen-Z is designed to be fast enough for multiple generations and should make it easier to upgrade parts.

Lenovo: No, not so much the DLC architecture, but rather the increased power consumption may lead to large parts being replaced, since installed power supply units may not be able to cope with the power demand of new cards.

Question from the audience: *What will be the requirement on site power outlets?*

Lenovo: We will try to stay below a maximum of 63A. It will depend on the number of power distribution units, and ultimately on the density of the system you want.

HPE: There is a trade-off between power consumption and floor space. HPE will work on different options and tries to fit within given limits.

Atos/Bull: We are reaching configurations with 100 kW/rack or more. Power delivery to the systems by means of power cables is becoming more awkward. We see many centres switch to bus bars with a higher amperage. This may become required.

Cray: In some datacentres the rack density may need to be decreased below what is optimal from a power efficiency point of view, to fit with existing power configurations.

Comment from the audience: *It used to be that it was hard to fit the cooling in a given space, the increasing power density may shift the difficult part to fitting the power distribution.*

Question from the audience: *What about quantum computers? Is that an unrealistic dream or something we should start preparing for?*

Atos/Bull: It is not an unrealistic dream, but it is not considered to be a candidate for an exascale system soon. A programming model needs to be developed first. Progress on this topic can be made on other systems simulating the working of a quantum computer. Having some selected applications up and running in 5 to 10 years from now may be realistic. A very different infrastructure is needed to accommodate quantum computers, due to the cooling requirements of near 0°K. The cooling infrastructure is much larger than the compute system. It is a very small chip requiring 20-30 kW for its cooling infrastructure. Quantum computer systems are very sensitive in several other respects. They need to be rigorously shielded from various types of noise: vibrations, and electromagnetic influences.

Lenovo: The first exascale machine will not be a quantum computer. Applications cannot just be ported, the programming paradigm is totally different. Therefore, new applications need to be written for quantum computers. Throwing away the current code base, and experience that programmers have, will be hard.

11 PRACE session

11.1 PPI4HPC Update

Within the work programme 2016-2017 of Horizon 2020, the European Commission launched the call EINFRA-21-2017 in order to provide support to a public procurement of innovative (PPI) HPC systems.

A public procurement of innovative solution is an approach defined by the European Commission (EC) in which several public procurers procure jointly innovative solutions and as such act as early adopters of innovative products or services that are newly arriving on the market (not widely commercially available yet). This approach aims at speeding up the availability of solutions meeting public sector requirements for large scale deployments.

The PPI4HPC project, prepared by four PRACE partners (CINECA, BSC, GENCI/CEA and JUELICH) in answer to this call, was selected by the EC in December 2016 and the project itself started in April 2017.

The goals of the project are to:

- Foster science and engineering applications in Europe by providing more computing and/or storage resources through PRACE
- Promote R&I on HPC architectures and technologies in Europe
- Have a greater weight and more impact on common topics of innovation and on the design of the solutions according to the need of scientists and engineers in Europe by a coordinated approach

During the project, each partner involved will procure an innovative IT equipment to be integrated into its computer centre and to be used as a production system. Among the four partners, one (BSC) will procure a storage and computing infrastructure for high-performance data analytics while the others will procure compute equipment:

- CINECA: an evolution of the Tier-0 system towards exascale architecture
- GENCI/CEA: a high-performance and energy-efficient computing system towards the convergence of HPC, HPDA and AI
- JUELICH: a supercomputer with innovative support for data-intensive applications

The project started with a market consultation in order to inform the market about the project and to get input from the market both on technical and legal aspects. The consultation was widely announced by a prior information notice, by articles in the major HPC electronic magazines and by the national contact points.

Then the procurement organised according to the competitive dialogue procedure was prepared and the procurement documents were written. The procurement documents include a common part mostly based on a set of topics of common interest and four local parts related to site specific context and needs. Consequently, it was decided to divide the procurement procedures into two steps: a first step conducted jointly until the selection of candidates for each site, a second step conducted separately until the selection of the winner for each site.

The topics of common interest include:

- Energy efficiency and power management: Control, optimization, efficiency of energy and power are major goals of future systems
- Data management: Data storage and management is needed to keep pace and make supercomputer usable
- Programming environment and productivity: Usability and flexibility are key aspects of future computing centres
- Datacentre integration: Enable links between systems and facility to optimise efficiency and operations
- Maintenance and support: Reliability, online maintenance are also a way to be more efficient
- System and application monitoring: Towards unified full-system monitoring capabilities that include production workload performance data
- Security : Security is a strategic aspect for HPC facilities, and its level will be determined according to legal regulation
- Total cost of ownership: Assess all the cost items related to an HPC system (OPEX & CAPEX) to ensure that the best value for money is reached taking into account the future activity on the system

In the end, the publication of the contract notice, the formal start of the procurement, occurred in May 2018.

11.2 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 to the new datacentre.

After the move the “to do backlog” (corrections were needed for the cabling and for the air flow) took much longer than expected. However, all was fixed without any major impact on the production. In addition, the installation of an extension of one Sequana cell in Q1 2017, to the national supercomputer Cartesius, went on well without any cost in terms of infrastructure work. After one and a half year of datacentre operation, the experience is very positive with more than 10% savings on just the exploitation side, compared to the old datacentre.

More recently, 23 nodes with 4 Nvidia consumer grade GPUs each were added to the Lisa national compute cluster. Those nodes are very cost effective for machine learning as they are as powerful in single precision as server grade GPUs.²

² SURFsara is allowed to experiment with this setup in a non-commercial context. Nvidia however does not support the use of consumer grade CPUs for this purpose.



Figure 21: Atos/Bull Quantum Learning Machine

A tender for a successor to Cartesius is now ongoing. Digital Realty actively participates in the dialogue rounds where the datacentre is concerned. The new system, with a target performance of 7-8 PFlop/s for the first phase, is expected to be delivered and in acceptance by the end of 2019.

In the SURF Open Innovation Lab (SOIL) an Atos/Bull “Quantum Learning Machine” - a conventional computer with ample memory, enough CPU capacity, and software to simulate the workings of a 30 qubit quantum computer – has been deployed. Several parties in the Netherlands are interested in furthering the development of quantum computing and collaborate to experiment with and explore what are suitable programming concepts for quantum computing.

The presentation also summarised an inquiry into the aspects of hosting a real quantum machine important for facility management, based on the site preparation manual of D-WAVE, a company who currently markets quantum annealing computers.

11.3 NSC-LiU (Sweden)

NSC-LiU (National Supercomputer Centre-Linköping University) provides HPC services to academic institutions in the country as well as to partners SMHI (weather forecast) and SAAB, since 1989, with a staff of circa 35 people. The talk focused on major evolutions since the presentation given during the 2014 workshop (#5).

In terms of systems for national academic research, “Triolith” (2012) is being replaced by “Tetralith” (4 PFlop/s based on Skylake CPUs only). The installation is organized in two phases, the second phase will take place once the previous system is removed. The target date for the availability of the whole new system is the end of 2018. Some GPU nodes will be added by the end of 2018 for development and production. ClusterVision, the system integrator, provides the new racks for the new system, which have the same dimensions as the racks of the old system. So only minor adjustments are needed to the surroundings of the racks. The new system will reuse the hot aisle air containment structure that was built for the previous system.

When the datacentre was built, the roof was prepared to receive solar panels. The installation took place recently, with 864 panels for a peak capacity of 255 kW (Figure 22). The location is good, with an unobstructed path to the south. However, the quantity of sun is limited in winter (6.5 hours of sun). The production starts more slowly than expected in the morning, but continues longer than expected in the evening. The electricity from the solar panels is sent to the outside electricity feed through 5 inverters which limit the maximum quantity of energy that can be sent to the electrical network.



Figure 22: 864 Solar panels, contributing up to 255 kW of power to the Linköping datacentre

Finally, two recent electrical problems that were difficult to detect were reported:

- A badly tightened screw in the power rails of a substation caused electrical arcs. This incident was difficult to identify and fix, since the access to the part of the substation where the problem was located was very difficult. Designing a power substation for serviceability is important.
- A UPS failure that caused batteries to be overcharged and inflated. This problem was undetected by temperature monitoring as a door, that inadvertently was left open, disturbed the monitoring.

11.4 IT4Innovation (Czech Republic)

Located in Ostrava, the mission of IT4Innovation is to carry out excellent research in the fields of high-performance computing and advanced data analysis, and to operate the leading national supercomputing infrastructure, mediating its effective use in order to increase the competitiveness and innovation of Czech science and industry. IT4Innovations aspires to be a leading supercomputing centre that provides professional services and conducts excellent research in the field of high performance computing and processing of advanced data sets for the benefit of science, industry and society.

Currently two HPC systems are operated at IT4Innovation:

- The [Anselm cluster](#) [6] with 209 nodes, 94 TFlop/s peak, 70 kW power consumption.
- The [Salomon cluster](#) [7] 1008 nodes, 2 PFlop/s peak, 700 kW power consumption (DLC cooling). This system was installed in 2016.

The talk focused on the experience of running the infrastructure described in the presentation given during the 2016 workshop (#7) [2]. The infrastructure works as expected, however, several points can be noted:

- The power supply is fully redundant, except for the power line. A second power line was planned but not installed so far
- The voltage supplied by the utility (23 kV) was not exactly the one expected (22 kV)
- Procedural measures had to be implemented for dealing with the OxyReduct system
- Diesel fuel needs to be checked annually for degradation and overall quality
- New investments were decided in order to:
 - Improve the monitoring of the power consumption of the cooling system
 - Provide power supplied from public network protected by the DC DRUPS to selected management offices, critical infrastructure components (telephone, network, etc.) and meeting rooms

Several issues occurred since the infrastructure is running:

- Corrosion of bearings inside of cooling units even if these cooling units were supposed to be designed for outside usage
- Overheating of flywheel UPS bearings. The operation mode had to be changed in order to prevent such problems leading to higher operation cost
- Overheating of OxyReduct generators. The air flow had to be improved
- Problem of oxygen concentration in the computer room, due to the difference of pressure caused by polluted air filters for the building. The filters had to be changed
- Issues with the quality of components of the OxyReduct system
- Large leakage of water in the computer room due to a disconnected hose and a switch to a secondary supply of water leading to the same leak. This incident caused one week of outage that was needed for cleaning (glycol is sticky) and for checking the electrical distribution

Next year, a new building will be built next to IT4Innovation. The effect of vibrations that may occur is unknown.

12 Main findings

The presentations that were given during the 9th 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 insight into developments on the HPC system integration side and the challenges these pose in terms of infrastructural facility requirements
- 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.

12.1.1 Hosting

Such anticipation is inherently difficult and fallible. And when it fails, adaptation is often costly. This motivates the quest from the facility management side for flexibility, modularity, multiple

options for cooling, variation in the density of racks. Vendors, and certainly the palette of vendors as a whole, do offer a wide span of options, not only to be able to cater to localities with different climates and different power provisioning conditions, but because they realise that for many of their customers the envisaged lifetime of a datacentre is usually five or six generations of machines. One new development from the vendor side is to offer modularity and flexibility by expanding the scope of system integration and take substantial chunks of infrastructure into the product they offer: plug-and-play containerised datacentre extensions. The containerisation of IT equipment is not new, but it is quite new for (warm-)water cooled HPC equipment.

In Sweden, where, because of its climate, heat has commercial value most of the year, real estate development companies plan to develop modularly expandable datacentre sites built from HPC containers. The sites aim to plug-and-play “sockets” for power, network, electricity, air and water cooling, and waste heat disposal.

The raised floor remains to be debated. For dense heavyweight racks, that cannot be air-cooled anyhow, it is not needed and a construction that can carry the weight is a costly challenge. On the other hand, serviceability and safety – e.g. the separation of electricity from water – are challenging issues when everything has to be fitted above the racks.

12.1.2 Exascale or exaFlop/s?

During last year’s workshop system integrators estimated that a general purpose CPU based exaFlop/s system, using current chip technology, would need an unaffordable and unsustainable 150 MW of power. This year, the estimate is that between 40 MW to 80 MW will be needed, depending on how much of the compute power is allowed to come from GPGPUs and other accelerators. The wording has shifted from “exaflop” to ‘exascale’, the focus from more generic Flop count to the performance of specific applications. Exascale, may be attainable and affordable for a few applications, but for some time to come, not for the “long tail” of HPC applications.

12.1.3 Electricity

The quest for highly potent but power efficient systems has led to systems that, being efficiently (liquid) cooled, contain many densely packed chips with a high TDP and many DIMMs. Vendors tend to design racks in which the custom designed distribution of power within the rack is also highly optimised, using direct current (DC). The consequence is that the rack has to be provided with high amperage AC/DC by the datacentre, that is taken in by the rack, transformed and distributed inside the rack. The bus bars, circuit breakers, fuses, fault current detection systems, etc. that are needed to provide the racks with this amount of power effectively and safely, are not off the shelf commodity goods. This has an upward effect on the cost of the datacentre’s power provisioning infrastructure. Safety precautions for dealing with high currents are also demanding (see also 12.1.5, Monitoring).

The price of electricity determines a large part of the operational costs of datacentres. Looking at buying electricity from the angle of risk management, and the advantages and disadvantages of various ways of procuring electricity, have been important topics at this workshop. From a cost perspective, a stable predictable power draw is an asset, because electricity tends to be at its cheapest when the actual usage is as close as possible to the limit of one’s contract.

For the most “powerful” – to be taken quite literally – centres, their growth in absolute power usage may have non-negligible effects on the regional power grid that have to be investigated and taken into account.

Rotary, usually diesel motor driven, UPSs are present in most datacentres, but their energy efficiency is debatable, while the recurring maintenance that is required fairly often is laborious and costly. Off-line static UPSs are becoming an interesting alternative.

12.1.4 Cooling

Multiple cooling loops give the flexibility to avoid downtime when maintenance of the mechanical infrastructure is required. Air cooling needs a large plenum, a “raised floor”, but is generally simpler to implement than liquid cooling. Air cooling limits rack density, makes it more difficult to re-use waste heat, and is vulnerable to pollution that may increase corrosion - especially in the case of “free cooling” – but it still needs to be taken into account.

Direct liquid cooling is more costly to implement, but when done at higher inlet temperatures (warm water cooling) produces high-grade waste heat that is easier to re-use. The removal of fans also leads to noise reduction. However, heat re-use cannot be the dominant consideration with respect to operating temperature. Chips generally perform better and have extended longevity when not run at their maximum temperature. The best rule of thumb is probably to choose an operating temperature that is as low as possible, while still allowing you to run without traditional chillers (at least for the bulk of the year). In addition, the requirements for new generations of processors may change: high power consumption processors of the future are quite possibly to require a cooler inlet temperature.

Immersion cooling has been around for a number of years and has not quite caught on, but seems to be maturing slowly. Better coolants, which avoid the undesirable longer term side effects associated with oil, are available. Immersion cooling has better noise reduction and better protection against pollution as desirable side effects. It can operate at temperatures as high as 75°C, which makes profitable heat re-use easier. The cost of the cooling fluids may be an issue.

12.1.5 Monitoring

Practice shows that monitoring needs to “understand” the technical design of the datacentre. Just monitoring key parameters such as temperature, humidity, etc., will not suffice. Redundancy and fault tolerance without monitoring the details of the way in which redundancy and fault tolerance are organised, leads to fault hiding, and ultimately to delaying catastrophic impacts rather than preventing them. Redundancy and/or overcapacity as such are an effective means of reducing interrupts of production. Monitoring systems must be triggering alarms, and preferably send insightful diagnostics whenever redundancy is reduced and whenever quantitative relations between parameters within a fault tolerant system change.

Adequate monitoring needs a large number of sensors. Practice shows that not all of them are reliable and that they have to be calibrated, and possibly checked periodically for the need for recalibration. Regular inspection of the datacentre to detect degradation of components, fault currents, and excessive heat, are also recommended.

12.2 Current situation in Europe

Facility managers of HPC centres in Europe face the same challenges as their colleagues and peers elsewhere on the globe. There is a continued strong involvement of European HPC sites in the work of the EEHPCWG (Energy Efficient HPC Working Group), a US Department of Energy sponsored initiative which, now for the sixth 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.

The workshop is important for sharing best practices and contributes to sustaining a high level of expertise and to the spread and implementation of advanced technologies in European HPC centres, most notably HPC centres of PRACE partners. It confirms that HPC centres in Europe are able to host and operate large supercomputers in a reliable and energy efficient way.

The largest machines and facilities are becoming so large that their operational behaviour has non-negligible effects on power grids. Whether power grids in Europe are more suited to accommodate quick changes of power consumption than US grids is to a large degree still an open question.

Large Tier-0 sites and smaller sites alike, must all take into account that a significant growth of application performance will imply that more power is needed. To some extent this can be mitigated by including accelerators to provide additional “raw” Flop/s capacity. However, smaller sites, that have to cater to the “long tail” of HPC applications, tend to be more limited in their options to invest in more power efficient but less generic GPGPU and other accelerator technologies, as a considerable number of the codes targeted need significant adaptation.

The price of electricity differs considerably over Europe, but the affordability of power is a growing concern for many centres. With electricity being a major part of their operational expenditure, centres are exploring the cost efficiency and risks involved in alternative ways of procuring for power at their disposal.

The PRACE Pre-Commercial Procurement process (PCP) and the more recent Public Procurement of Innovative HPC systems (PPI4HPC) foster the development of energy efficient HPC technologies and architectures, with, for the PCP, R&D for a significant part conducted in Europe. The PCP round that started in 2012 has shown promising results for advanced technologies that are capable of scaling to 100 PFlop/s or more and that can be demonstrated to be highly energy efficient running real applications.

13 Conclusion

The 9th Workshop on HPC Infrastructures has been very successful in bringing together experts from various disciplines and in various stakeholder roles working on HPC site infrastructures. Figure 23 shows that the annual workshop, which is attended upon invitation only, has become an institution that is capable of consistently attracting a stable number of experts in the field.

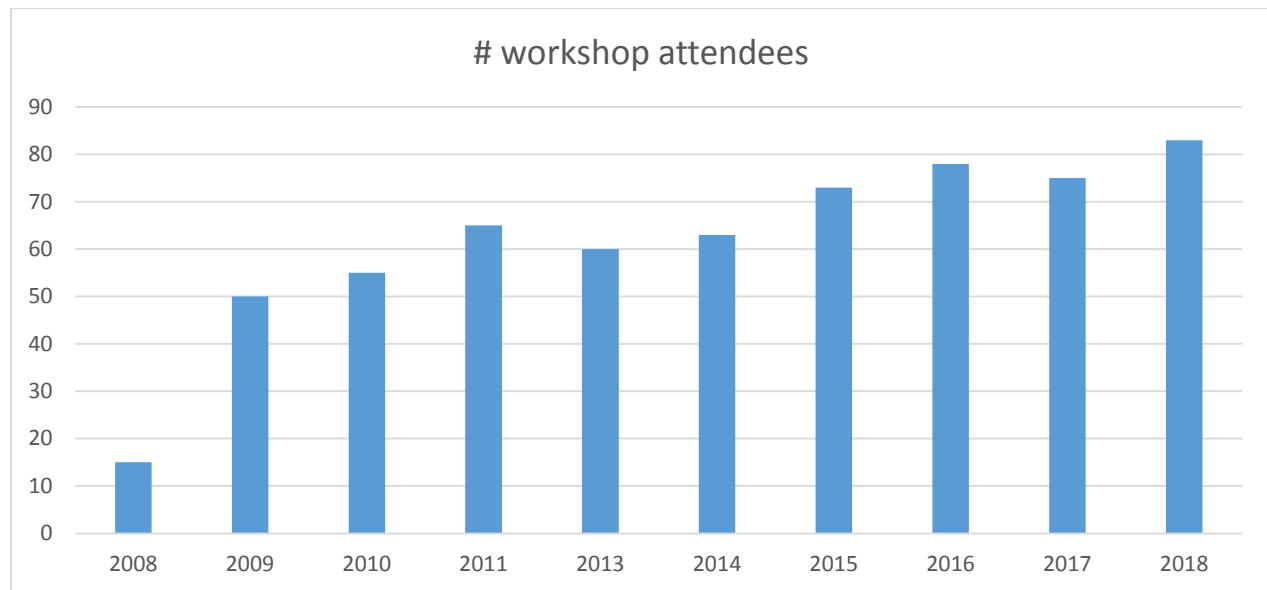


Figure 23: Number of workshop attendees in historical perspective

Energy efficiency and cooling technologies are of course areas of prime importance that also recurred this year, as did strategies for managing total cost of ownership of HPC facilities and machines.

New on the agenda were real-estate development projects that showed an explicit interest in attracting HPC datacentres to sites that offer not only space, network, and power, but also cooling and waste-heat removal as an integral service. Their business case is based on selling services on the one hand, and buying high-grade heat from HPC sites and other possibly industrial processes as a by-product, to integrate them into large, regional scale heating systems, on the other hand.

The provisioning of power, the possible impact of machines on the regional electricity grids, getting high amperage efficiently and safely to very dense racks, and cost-effective ways of procuring for power contract figured more prominently on the agenda than before. Since the application of several advanced cooling technologies and heat integration with heat re-use have progressed – as contributions to the workshop document – the balance of attention in coming years might shift even more to aspects of power provisioning.

During the workshop discussions, system integrators and chip builders dared to estimate that a first generation exaFlop/s machine would need at least 40 MW to 80 MW of power, depending on how much of the compute capacity would require to be implemented by general purpose CPUs, rather than GPGPUs and other accelerator technologies. Exascale application performance, if defined in terms of acceleration compared to performance measured on current large systems – as opposed to raw Flop/s count – is currently within reach, but for a limited number of carefully chosen and optimised applications.

The next edition of the workshop, hosted by PSNC, will take place in Poznan in May 2019.