E-Infrastructures
H2020- INFRAEDI-2018-2020

INFRAEDI-01-2018: Pan-European High Performance Computing infrastructure and services (PRACE)

PRACE-6IP

PRACE Sixth Implementation Phase Project

Grant Agreement Number: INFRAEDI-823767

D5.1

Installation requirements and best practices for hosting Exascale systems

Final

Version: 0.7
Author(s): Hayk Shoukourian, BADW-LRZ; Gert Svensson, KTH; François Robin, CEA; Huub Stoffers, SURF; Ezhilmathi Krishnasamy, U Luxembourg; Sebastien Varrette, U Luxembourg
Date: 19.05.2021

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: INFRAEDI-823767**	
	Project Title: PRACE Sixth Implementation Phase Project	
	Project Web Site: http://www.prace-ri.eu/about/ip-projects	
	Deliverable ID: D5.1	
	Deliverable Nature: Report	
	Dissemination Level: PU*	**Contractual Date of Delivery:** 31 / May / 2021
		Actual Date of Delivery: 31 / May / 2021
	EC Project Officer: Leonardo Flores Añover	

* - The dissemination level is 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: Installation requirements and best practices for hosting Exascale systems**	
	ID: D5.1	
	Version: 0.7	**Status:** ***Final***
	Available at: http://www.prace-ri.eu/about/ip-projects	
	Software Tool: Microsoft Word 2016	
	File(s): D5.1-v0.7.docx	
Authorship	**Written by:**	Hayk Shoukourian, BADW-LRZ; Gert Svensson, KTH; François Robin, CEA; Huub Stoffers, SURF; Ezhilmathi Krishnasamy, U Luxembourg Sebastien Varrette, U Luxembourg
	Contributors:	
	Reviewed by:	Lukas Demovic, CCSAS Dirk Brömmel, JUELICH
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	18/January/2021	Draft	Initial draft and setup of the document structure
0.2	12/February/2021	Draft	Addition of survey results (Chapter 3). Proposal to include a subsection5.1 providing high-level information on metrics to

			monitor with regard to energy-efficiency
0.3	22/March/2021	Draft	Compilation of inputs from partners concerning Chapter 2 and Chapter 4. Addition of executive summary and introduction sections. Editing.
0.3.1	24/March/2021	Draft	Updated the floor space value for LUMI
0.3.2	09/April/2021	Draft	Includes update concerning the metrics subsection. Contains the early draft version of the conclusion section.
0.3.2.1	21/April/2021	Draft	Revisiting the conclusion section. Minor additions in the survey section. Editing of acronyms and abbreviations used.
0.4	28/April/2021	Draft	Second draft of input on Chapter 5 (metrics section excluded) received and integrated.
0.5	03/May/2021	Draft	Draft review and finalisation.
0.6	17/May/2021	Draft	Version addressing the internal review comments
0.7	19/May/2021	Draft	Addressing the comments raised during the second internal review
1.0		Final	Integration of reviewer remarks

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Preparatory Access, SHAPE, HLST

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] European Workshops on HPC Infrastructures: https://prace-ri.eu/infrastructure-support/european-workshops-on-hpc-infrastructures/

[2] https://eurohpc-ju.europa.eu

[3] https://prace-ri.eu/infrastructure-support/prace-hpc-infrastructure/

[4] "Call for Expression of Interest for the Selection of Hosting Entities for Precursors to Exascale Supercomputers", European High Performance Computing Joint Undertaking, [REF:EUROHPC-2019-CEI-PE-01], "General system requirements", p. 16.

[5] https://www.Top500.org/lists/Top500/list/2020/11

[6] https://www.hpe.com

[7] https://eurohpc-ju.europa.eu/news/lumi-new-eurohpc-world-class-supercomputer-finland

[8] https://atos.net/en/

[9] https://eurohpc-ju.europa.eu/news/leonardo-new-eurohpc-world-class-pre-exascale-supercomputer-italy

[10] https://www.bsc.es/news/bsc-news/marenostrum-5-will-host-experimental-platform-create-supercomputing-technologies-%E2%80%9Cmade-europe%E2%80%9D
https://group.vattenfall.com/press-and-media/pressreleases/2020/vattenfalls-hydro-power-runs-finnish-supercomputer

[11] "PUE: A Comprehensive Examination of the Metric". The Green Grid, White Paper #49: https://www.thegreengrid.org/en/resources/library-and-tools/20-PUE%3A-A-Comprehensive-Examination-of-the-Metric

[12] HPC Infrastructures Workshop #9, 2018, Bologna, p. 24, discloses a number of details about the site then under development. https://prace-ri.eu/wp-content/uploads/5IP-D5.4.pdf

[13] https://www.eetimes.com/u-s-launches-exaflop-supercomputer-initiative/#

[14] https://www.hpcwire.com/2018/04/11/us-plans-1-8-billion-spend-on-doe-exascale-supercomputing/

[15] https://www.olcf.ornl.gov/frontier/

[16] https://www.alcf.anl.gov/aurora

[17] https://www.llnl.gov/news/llnl-and-hpe-partner-amd-el-capitan-projected-worlds-fastest-supercomputer

[18] https://www.anandtech.com/show/15581/el-capitan-supercomputer-detailed-amd-cpus-gpus-2-exaflops

[19] https://www.olcf.ornl.gov/frontier/#4

[20] https://insidehpc.com/2020/12/building-an-exascale-class-data-center-for-ornls-incoming-frontier-supercomputer/

[21] https://www.tomshardware.com/news/el-capitan-supercomputer-cray-shasta-intel-amd-nvidia,40142.html

[22] https://www.alcf.anl.gov/aurora

[23] https://www.datacenterdynamics.com/en/news/cooling-exascale-lawrence-livermore-plans-two-exascale-supercomputers-totaling-85mw/

[24] https://www.llnl.gov/news/lab-breaks-ground-exascale-facility-upgrades

[25] https://www.tsmc.com/english/dedicatedFoundry/technology/logic/l_7nm

[26] https://www.fujitsu.com/global/documents/about/resources/publications/technicalreview/2020-03/article04.pdf

[27] https://www.nextplatform.com/2019/05/02/china-fleshes-out-exascale-design-for-tianhe-3/

[28] https://eurohpc-ju.europa.eu/discover-eurohpc

[29] Shoukourian, Hayk, and Dieter Kranzlmüller. "Forecasting power-efficiency related key performance indicators for modern data centers using LSTMs." Future Generation Computer Systems 112 (2020): 362-382.

[30] Rong, Huigui, Haomin Zhang, Sheng Xiao, Canbing Li, and Chunhua Hu. "Optimizing energy consumption for data centers." Renewable and Sustainable Energy Reviews 58 (2016): 674-691.

[31] Díaz, Andrés J., Rodrigo Cáceres, José M. Cardemil, and Luis Silva-Llanca. "Energy and exergy assessment in a perimeter cooled data center: The value of second law efficiency." Applied Thermal Engineering 124 (2017): 820-830.

[32] Frederic Souques, Jean-Marc Ducos. “Big Data & Machine Learning at TGCC - First approach with KAPSDATA” PRACE Session of the Infrastructure Workshop, October 2020. https://events.prace-ri.eu/event/1000/

[33] https://eehpcwg.llnl.gov/

[34] Shoukourian Hayk. "Learning Energy-Aware Management of Modern Data Centers". 11th Annual Energy Efficient High Performance Computing Workshop (EE HPC WG Workshop 2020) Online, November 2020. https://sites.google.com/view/eehpcwg-2020/workshop-program

[35] Scogland, T.R., Steffen, C.P., Wilde, T., Parent, F., Coghlan, S., Bates, N., Feng, W.C. and Strohmaier, E., 2014, March. A power-measurement methodology for large-scale, high-performance computing. In Proceedings of the 5th ACM/SPEC international conference on Performance engineering (pp. 149-159).

[36] https://www.ashrae.org/

[37] " ERE: A Metric for Measuring the Benefit of Reuse Energy from a Data Center ", The Green Grid, White Paper #29: https://www.thegreengrid.org/en/resources/library-and-tools/242-ERE%3A-A-Metric-for-Measuring-the-Benefit-of-Reuse-Energy-From-a-Data-Center-

[38] https://climate.nasa.gov/evidence/

[39] https://ec.europa.eu/clima/policies/strategies/2030_en
[40] https://science.sciencemag.org/content/367/6481/984/tab-figures-data
[41] https://www.vxchnge.com/blog/data-center-cooling-technology
[42] Current Cooling Systems & Methods https://www.datacenters.com/news/data-center-cooling-future-of-cooling-systems-methods-and-technologies
[43] Hot Aisle/Cold Aisle Layout https://www.energystar.gov/products/low_carbon_it_campaign/12_ways_save_energy_data_center/hot_aisle_cold_aisle_layout
[44] https://www.gigabyte.com/Solutions/Cooling/asetek-liquid-cooling
[45] Immersion Cooling Markets https://www.globenewswire.com/news-release/2020/03/24/2005538/0/en/Immersion-Cooling-Markets-in-Data-Centers-2020-to-2025-Analyzed-by-Application-Cooling-Fluid-Type-and-Geography.html
[46] https://www.lumi-supercomputer.eu/
[47] Nordic Energy Technology Perspectives https://www.nordicenergy.org/wordpress/wp-content/uploads/2013/02/Nordic-Energy-Technology-Perspectives.pdf
[48] Pre-Exa and Exascale Challenge https://eehpcwg.llnl.gov/assets/sc19_07_200_300_pre-exa_and_exascale_challanges_-cavazonni.pdf
[49] CNAF al Tecnopolo e prospettivedel calcolo INFN https://agenda.infn.it/event/16759/contributions/33914/attachments/23913/27222/Zoccoli.pdf
[50] https://www.eia.gov/international/analysis/country/ITA
[51] https://www.cineca.it/en/about-us/certifications
[52] https://www.olcf.ornl.gov/frontier/
[53] https://www.olcf.ornl.gov/2020/12/11/building-an-exascale-class-data-center/
[54] ORNL Leading computing facility https://www.olcf.ornl.gov/2018/12/03/summit-facility-design-team-snags-tennessee-energy-award/
[55] ORNL Successfully Installs 2.5-mile Power Line for the OLCF's Frontier Supercomputer https://www.hpcwire.com/off-the-wire/ornl-successfully-installs-2-5-mile-power-line-for-the-olcfs-frontier-supercomputer/
[56] https://pawsey.org.au/about-us/history/
[57] https://pawsey.org.au/magnus-last-dance/
[58] https://pawsey.org.au/groundwater-cooling-system/
[59] https://pawsey.org.au/about-us/capital-refresh/
[60] https://pawsey.org.au/our-green-credentials/
[61] https://issuu.com/meland6/docs/datacenter_2019_09_webb
[62] Data Centers In North Sweden, Making data centers sustainable and energy efficient
[63] https://www.blog.google/outreach-initiatives/sustainability/100-percent-renewable-energy-second-year-row/
[64] https://blog.google/inside-google/infrastructure/cleaner-data-centers-batteries-included/
[65] https://blog.google/outreach-initiatives/sustainability/data-centers-energy-efficient/
[66] https://blog.google/outreach-initiatives/sustainability/our-biggest-renewable-energy-purchase-ever/
[67] https://cool-shield.com/aisle-containment/
[68] https://www.dataaire.com/precision-vs-comfort-cooling/
[69] http://www.etson.eu/node/63

[70] https://pawsey.org.au/our-green-credentials/

[71] https://baxtel.com/data-center/facebook-lulea

[72] https://www.datacenterknowledge.com/archives/2013/06/12/inside-facebooks-new-lulea-data-center

[73] https://www.3m.com/3M/en_US/data-center-us/applications/immersion-cooling/

[74] https://www.gigabyte.com/Solutions/Cooling/immersion-cooling

[75] https://assets.ext.hpe.com/is/content/hpedam/a50002389enw

[76] https://www.thegreengrid.org/

[77] "THE GREEN GRID METRICS: DATA CENTER INFRASTRUCTURE EFFICIENCY (DCIE) DETAILED ANALYSIS". The Green Grid, White Paper #14

[78] "Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric". The Green Grid, White Paper #32: https://www.thegreengrid.org/en/resources/library-and-tools/241-Carbon-Usage-Effectiveness-%28CUE%29%3A-A-Green-Grid-Data-Center-Sustainability-Metric

[79] "WATER USAGE EFFECTIVENESS (WUE™): A GREEN GRID DATA CENTER SUSTAINABILITY METRIC". The Green Grid, White Paper #35: https://www.thegreengrid.org/en/resources/library-and-tools/238-Water-Usage-Effectiveness-%28WUE%29%3A-A-Green-Grid-Data-Center-Sustainability-Metric-

[80] "SPACE USAGE EFFECTIVENESS (SPUE)". The Green Grid, White Paper #61: https://www.thegreengrid.org/en/resources/library-and-tools/233-WP

[81] https://www.top500.org/lists/green500/

[82] Bode, Arndt. "Energy to Solution: A New Mission for Parallel Computing." In Euro-Par, pp. 1-2. 2013.

[83] Shoukourian, Hayk, Torsten Wilde, Axel Auweter, and Arndt Bode. "Monitoring power data: A first step towards a unified energy efficiency evaluation toolset for HPC data centers." *Environmental Modelling & Software* 56 (2014): 13-26.

[84] Wilde, Torsten, Axel Auweter, Michael K. Patterson, Hayk Shoukourian, Herbert Huber, Arndt Bode, Detlef Labrenz, and Carlo Cavazzoni. "DWPE, a new data center energy-efficiency metric bridging the gap between infrastructure and workload." In *2014 International Conference on High Performance Computing & Simulation (HPCS)*, pp. 893-901. IEEE, 2014.

List of Acronyms and Abbreviations

aisbl	Association International Sans But Lucratif (legal form of the PRACE-RI)
ALCF	Argonne Leadership Computing Facility
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BCO	Benchmark Code Owner
CFM	Cubic Feet per Minute
CMU	CPU Memory Unit
CoE	Centre of Excellence
CPU	Central Processing Unit
CRAC	Computer Room Air Conditioning
CRAH	Computer Room Air Handler
CUDA	Compute Unified Device Architecture (NVIDIA)
CVC	Calibrated Vectored Cooling
DARPA	Defense Advanced Research Projects Agency

DCiE	Data Centre infrastructure Efficiency
DCIM	Data Centre Infrastructure Management
DEISA	Distributed European Infrastructure for Supercomputing Applications EU project by leading national HPC centres
DLC	Direct Liquid Cooling
DoA	Description of Action (formerly known as DoW)
DWPE	Data centre Workload Power Efficiency
EC	European Commission
ECMWF	European Centre for Medium-Range Weather Forecasting
EESI	European Exascale Software Initiative
EFlop/s	Exa (= 10^{18}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also EF/s
EoI	Expression of Interest
ERE	Energy Reuse Effectiveness
ERF	Energy Reuse Factor
EtS	Energy-to-Solution
ESFRI	European Strategy Forum on Research Infrastructures
EuroHPC JU	The European High-Performance Computing Joint Undertaking
EWHPC	European Workshops on HPC Infrastructures
FinFET	Fin Field-Effect Transistor
Flop/s	Floating Point Operations Per Second
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
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network. The current EC-funded project as of 2015 is GN4.
GFlop/s	Giga (= 10^9) Floating-point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GPU	Graphic Processing Unit
HET	High Performance Computing in Europe Taskforce. Taskforce by representatives from the European HPC community to shape the European HPC Research Infrastructure. Produced the scientific case and valuable groundwork for the PRACE project.
HLST	High-Level Support Team
HMM	Hidden Markov Model
HPC	High-Performance Computing; Computing at a high-performance level at any given time; often used synonym with Supercomputing
HPE	Hewlett Packard Enterprise
HPL	High Performance LINPACK
HVAC	Heating, Ventilation, and Air Conditioning
HW	Hardware
ISC	International Supercomputing Conference; European equivalent to the US-based SCxx conference. Held annually in Germany.
kB	Kilo (= $2^{10} \sim 10^3$) Bytes (= 8 bits), also kByte
LINPACK	Software library for Linear Algebra
LLNL	Lawrence Livermore National Laboratory
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. DP) per second, also MF/s
MOOC	Massively open online Course
MoU	Memorandum of Understanding.
MPI	Message Passing Interface
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
NNSA	National Nuclear Security Administration
ODA	Operational Data Analytics
OLCF	Oak Ridge Leadership Computing Facility
PA	Preparatory Access (to PRACE resources)
PATC	PRACE Advanced Training Centres
PDU	Power Distribution Unit
PB	Tera (= $2^{50} \sim 10^{15}$) Bytes (= 8 bits), also Pbyte
PFlop/s	Peta (= 10^{15}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also PF/s
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PRACE 2	The current phase of the PRACE Research Infrastructure following the initial five year period.
PRIDE	Project Information and Dissemination Event
PSU	Power Supply Unit
PUE	Power Usage Effectiveness
pPUE	partial PUE
RI	Research Infrastructure
SpUE	Space Usage Effectiveness
SW	Software
TB	Technical Board (group of Work Package leaders)
TB	Tera (= $2^{40} \sim 10^{12}$) Bytes (= 8 bits), also Tbyte
Tb/s	Tera (= 10^{12}) bits per second, also Tbit/s
TB/s	Tera (= 10^{12}) Bytes (= 8 bits) per second, also TByte/s
TCO	Total Cost of Ownership. Includes recurring costs (e.g. personnel, power, cooling, maintenance) in addition to the purchase cost.
TDP	Thermal Design Power
TFlop/s	Tera (= 10^{12}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
TGG	The Green Grid
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
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources.
UPS	Uninterruptable Power Supply
WUE	Water Usage Effectiveness

List of Project Partner Acronyms

BADW-LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Germany (3rd Party to GCS)
BILKENT	Bilkent University, Turkey (3rd Party to UHEM)

BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion, Spain
CaSToRC	The Computation-based Science and Technology Research Center (CaSToRC), The Cyprus Institute, Cyprus
CCSAS	Computing Centre of the Slovak Academy of Sciences, Slovakia
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3rd Party to GENCI)
CENAERO	Centre de Recherche en Aéronautique ASBL, Belgium (3rd Party to UANTWERPEN)
CESGA	Fundacion Publica Gallega Centro Tecnológico de Supercomputación de Galicia, Spain, (3rd Party to BSC)
CINECA	CINECA Consorzio Interuniversitario, Italy
CINES	Centre Informatique National de l'Enseignement Supérieur, France (3rd Party to GENCI)
CNRS	Centre National de la Recherche Scientifique, France (3rd Party to GENCI)
CSC	CSC Scientific Computing Ltd., Finland
CSIC	Spanish Council for Scientific Research (3rd Party to BSC)
CYFRONET	Academic Computing Centre CYFRONET AGH, Poland (3rd Party to PNSC)
DTU	Technical University of Denmark (3rd Party of UCPH)
EPCC	EPCC at The University of Edinburgh, UK
EUDAT	EUDAT OY
ETH Zurich (CSCS)	Eidgenössische Technische Hochschule Zürich – CSCS, Switzerland
GCS	Gauss Centre for Supercomputing e.V., Germany
GÉANT	GÉANT Vereniging
GENCI	Grand Equipement National de Calcul Intensif, France
GRNET	National Infrastructures for Research and Technology, Greece
ICREA	Catalan Institution for Research and Advanced Studies (3rd Party to BSC)
INRIA	Institut National de Recherche en Informatique et Automatique, France (3rd Party to GENCI)
IST-ID	Instituto Superior Técnico for Research and Development, Portugal (3rd Party to UC-LCA)
IT4I	Vysoka Skola Banska - Technicka Univerzita Ostrava, Czech Republic
IUCC	Machba - Inter University Computation Centre, Israel
JUELICH	Forschungszentrum Jülich GmbH, Germany
KIFÜ (NIIFI)	Governmental Information Technology Development Agency, Hungary
KTH	Royal Institute of Technology, Sweden (3rd Party to SNIC-UU)
KULEUVEN	Katholieke Universiteit Leuven, Belgium (3rd Party to UANTWERPEN)
LiU	Linkoping University, Sweden (3rd Party to SNIC-UU)
MPCDF	Max Planck Gesellschaft zur Förderung der Wissenschaften e.V., Germany (3rd Party to GCS)
NCSA	NATIONAL CENTRE FOR SUPERCOMPUTING APPLICATIONS, Bulgaria
NTNU	The Norwegian University of Science and Technology, Norway (3rd Party to SIGMA2)
NUI-Galway	National University of Ireland Galway, Ireland
PRACE	Partnership for Advanced Computing in Europe aisbl, Belgium
PSNC	Poznan Supercomputing and Networking Center, Poland

SDU	University of Southern Denmark (3rd Party to UCPH)
SIGMA2	UNINETT Sigma2 AS, Norway
SNIC-UU	Uppsala Universitet, Sweden
STFC	Science and Technology Facilities Council, UK (3rd Party to UEDIN)
SURF	SURF is the collaborative organisation for ICT in Dutch education and research
TASK	Politechnika Gdańska (3rd Party to PNSC)
TU Wien	Technische Universität Wien, Austria
UANTWERPEN	Universiteit Antwerpen, Belgium
UC-LCA	Universidade de Coimbra, Labotatório de Computação Avançada, Portugal
UCPH	Københavns Universitet, Denmark
UEDIN	The University of Edinburgh
UHEM	Istanbul Technical University, Ayazaga Campus, Turkey
UIBK	Universität Innsbruck, Austria (3rd Party to TU Wien)
UiO	University of Oslo, Norway (3rd Party to SIGMA2)
UL	UNIVERZA V LJUBLJANI, Slovenia
ULIEGE	Université de Liège; Belgium (3rd Party to UANTWERPEN)
U Luxembourg	University of Luxembourg
UM	Universidade do Minho, Portugal, (3rd Party to UC-LCA)
UmU	Umeå University, Sweden (3rd Party to SNIC-UU)
UnivEvora	Universidade de Évora, Portugal (3rd Party to UC-LCA)
UnivPorto	Universidade do Porto, Portugal (3rd Party to UC-LCA)
UPC	Universitat Politècnica de Catalunya, Spain (3rd Party to BSC)
USTUTT-HLRS	Universitaet Stuttgart – HLRS, Germany (3rd Party to GCS)
WCSS	Politechnika Wroclawska, Poland (3rd Party to PNSC)

Executive Summary

Supercomputing centres continuously aim to improve their resources for a better service provision to the HPC community. Within this objective, High-Performance Computing (HPC) data centres position themselves at the forefront of computing technologies, with regular system hardware and software updates allowing to ensure continuous research and enable further scientific discoveries.

Accommodating these updates on a regular basis is, however, not a trivial task, as planning and commissioning a data centre with a lifespan of several years while capable of hosting diverse supercomputers of different generations is quite a challenge.

This document discusses installation requirements and best practices for hosting next generation HPC systems (i.e. pre-exascale and exascale), with a specific focus on the current European ecosystem. More specifically, this document outlines foreseeable major challenges related to a wide spectrum of data centres' building infrastructure relevant aspects: ranging from hosting facility, through power availability and usage of renewable energy sources to cooling solutions and waste heat reuse, as well as suggests best practices for addressing those issues.

This best practice guide aims to deliver information and guidance useful when planning or upgrading an HPC data centre. The target audience of this document is representatives of supercomputing sites that plan to deploy large-scale systems capable of delivering an exascale (or close to exascale) compute performance in the coming years.

1 Introduction

The design of a supporting ecosystem for the efficient, reliable, and maintainable operation of High-Performance Computing (HPC) systems across generations of such high-end machines requires a thorough consideration of various aspects ranging from floor space constraints through power delivery and cooling to environmental impacts. The continued growth in demand for compute power and online resources has led to ever-high levels of energy consumption by modern HPC data centres, which not only result in high operating costs, but also affect the environment and limit further data centre expansion. Thus, it is imperative to fully understand the complex interaction of current installation requirements as well as possible and potentially feasible solutions when planning or upgrading a current supercomputing site.

The aim of this document is to assist HPC site owners and managers in their planning to host next-generation, exascale[1]-class HPC systems in the coming years, 2023 - 2025. For that purpose, this document investigates the existing installation requirements and challenges when deploying next generation HPC systems and discusses the possible solutions and best practices. This investigation is primarily based on:

- the information gathered during European Workshops on HPC Infrastructures (EWHPC) [1]
- a survey circulated among European High-Performance Computing Joint Undertaking (EuroHPC JU) [2] pre-exascale and petascale sites as well as among additional PRACE Tier-0 and Tier-1 sites [3]
- other publicly available sources

The remainder of this document is organised as follows. Chapter 2 provides an update on current exascale projects regarding the installation requirements and solutions put in place for hosting exascale systems in Europe, as well as in the USA, Japan, and China. Chapter 3 conveys further details on installation requirements for hosting next generation HPC systems based on the survey organised by the PRACE-6IP WP5 T5.2 team and circulated among EuroHPC petascale and pre-exascale as well as PRACE Tier-0/Tier-1 sites. Based on the information given in Chapter 2 and Chapter 3, Chapter 4 derives the main installation requirements, and Chapter 5 recommends best practices for hosting next generation European HPC systems, whereas Chapter 6 outlines some of the available Key Performance Indicators (KPIs) used for monitoring the energy efficiency in modern data centres. Chapter 7 provides a further outlook and concludes this document.

2 Review of current Exascale projects

This chapter reviews various projects inside and outside of Europe concerning installation requirements and implemented decisions for hosting next generation HPC systems. The information presented in this chapter is mainly derived from the proceedings of EWHPC [1] and other publicly available sources.

[1] Systems capable of performing 10^{18} floating point operations per second.

2.1 Review of EuroHPC pre-exascale projects currently under way

One essential way in which the EuroHPC Joint Undertaking (JU) seeks to sustain, develop, and foster European HPC infrastructure and HPC operations is by procuring three top-class HPC systems in the pre-exascale range and subsequently owning and operating these systems. As defined by the JU, this top-class must have a sustained aggregate compute capacity of at least 150 PFlop/s measured by means of the High-Performance LINPACK (HPL) benchmark [4]. Thus "pre-exascale" can indeed be considered top class on a global level if the Top500 list of November 2020 is taken as a representative reference: the positions 1-3 on that list have measured HPL performances of (1) 442.01 PFlop/s, (2) 148.6 PFlop/s, and (3) 94.64 PFlop/s respectively[5].

The JU has tendered for partners willing to host a pre-exascale system and selected three centres with considerable experience in operating HPC systems and HPC data centre infrastructure: CSC in Finland, CINECA in Italy, and BSC in Spain. The procurements of the systems themselves, culminating in the delivery and acceptance of an operational computer system, each have their own timeline. At the time of writing, none of the systems is functional yet. Comparable metrics about the systems, let alone about the data centre efficiency achieved in actually accommodating the workload on these systems, are not available. The review of these systems, focusing mainly on approaches to meet key infrastructure installation requirements, is therefore based on projections disclosed by the sites involved during their presentation at the EWHPC online event of the 14th of October 2020 [1], complemented with details disclosed in earlier and later press releases.

The three EuroHPC pre-exascale systems currently underway are the following:

1. "LUMI", hosted by CSC, in Kajaani, Finland. The contract for the supercomputer system was awarded to Hewlett Packard Enterprise (HPE) [6] for a system that has an aggregate nominal peak performance of 550 PFlop/s and is expected to have an HPL performance of 375 PFlop/s [7]. System delivery is expected to start in Q2 2021.
2. "Leonardo", hosted by CINECA, in Bologna, Italy. The contract for the supercomputer system was awarded to Atos [8] for a system that is expected to have an HPL performance of 248 PFlop/s [9]. System delivery is expected to start in Q3 2021.
3. "MareNostrum 5", hosted by BSC, in Barcelona, Spain. MareNostrum 5 will consist of two main clusters. Their joint aggregate peak HPL performance is targeted to be at least 200 PFlop/s [1]. Apart from the two clusters mentioned, MareNostrum 5 will also include an experimental platform for developing and testing novel supercomputing technologies "made in Europe". System delivery is expected to start in Q4 2021.

The data centre infrastructures set up to accommodate each of these systems will, in one way or the other, have to satisfy some basic compute capability-related parameters that have been formulated by JU in their call for hosting partners as requirements to comply with:

- *Power capacity and power quality for hosting a system in the range of 10 to 15 MW total consumption for the pre-exascale supercomputers*
- *Uninterruptable Power Supply (UPS) power available to cover the critical systems, including storage and access to data of the JU system.*
- *Adequate capacity of air or liquid cooling for hosting the JU system*
- *At least 700 m2 of contiguous floor space available for hosting the EuroHPC supercomputer and auxiliary systems*

- *Raised floor able to bear at least 2,200 kg/m2 distributed load*
- *100 Gbit/s connectivity towards the rest of the GÉANT Network (link capacity)[4]*

It is no surprise and consistent with the trend in HPC over the last decade, that power and cooling requirements figure prominently in this list. Despite ongoing research and development aiming at highly potent dense processors running at low power and improvements in cooling technology, it has not been possible to grow in compute power without substantially increasing power and cooling capacity.

The explicit mentioning of at least 100 Gbit/s uplink capacity to the "world at large" in part reflects that the systems must be versatile and able to handle many different workflows and applications besides the ones that can be considered classical Message Passing Interface (MPI) based applications that can orchestrate the use of thousands of cores in a single compute job. The classical workflow is accommodated above all by a high bandwidth and low latency interconnect between the nodes of the HPC compute facility itself. However, many workflows, for example in the domain of elementary particle physics analysing data produced by the Large Hadron Collider experiments at CERN, make use of compute grid protocols that combine the need for substantial amounts of compute capacity with more frequent and more substantial network I/O from and to external compute and storage facilities.

2.1.1 LUMI, CSC, Kajaani

The LUMI system will be hosted in Kajaani, in a location that once used to house a large paper mill. Due to the industrial history of the site, there is ample power capacity in the local power grid with 230 MW, fed from the national power grid through three independent transmission lines. Over the last 38 years, there have only been two minutes of power outages.

The energy for LUMI is 100% fossil-free and renewable. CSC has a contract with Vattenfall for a yearly volume of up to 100 GWh of guaranteed origin-certified hydropower [10]. Put differently, to actually produce the 100 GWh figure on the annual power bill, the day in day out sustained average aggregate power draw would have to be slightly over 11.4 MW.

Kajaani is that far up north that free cooling, i.e. cooling with the use of any mechanical chillers, can be used all year round. The system can be operated with a Power Usage Effectiveness (PUE) [11] of 1.03. However, instead of using cooling towers, the waste heat of the system, captured in heated cooling water, will be sold to the Kajaani municipal energy company for use in their district heating system. Waste heat from LUMI thus will provide about 20% of the heat needed for the city of Kajaani. The income generated by selling waste heat reduces the operational cost of the system. Since the municipal energy company currently uses fossil fuel, the waste heat reuse will reduce emissions. In that sense, the system has a *negative* annual carbon footprint of 13,500 tons of CO_2.

The previous paper mill hall has a surface area of 15,000 m^2, which is more than enough to house all subsystems of LUMI. LUMI is a very heterogeneous system, both in terms of compute and storage resources. A CPU partition, a GPU partition, a data-analytics partition, and a cloud container partition are all coupled via the same high speed, low latency interconnect, as are storage subsystems for a shared parallel – Lustre – file system, CEPH object storage, and accelerated storage.

Figure 1: Kajaani data centre site, with an indication of the hall to house LUMI

The heterogeneity of the system reflects the ambition to cater not only to HPC workloads, but to also support different applications and workflows that may also imply steeper requirements for the frequency and volume of network I/O to and from the system. At the virtual HPC infrastructures workshop, LUMI was presented as a 'Swiss Army Knife' of computing and 'not only about Tier-0 simulation applications. The LUMI system is located in a data centre that is directly connected to the Nordic backbone. Currently, the uplink to GÉANT is 4×100 Gbit/s. But this can easily be expanded to multiple Tbit/s. Support of grid access protocols is not planned initially but can be considered later. Catering optimally to different scientific communities is a priority. External connectivity resources are recognised as a vital data centre infrastructure aspect. Lack of external bandwidth to accommodate frequent and high volume I/O will not be an issue that enters decision making on how to enable particular groups best.

2.1.2 Leonardo, CINECA, Bologna

The Leonardo system will be hosted at the new Bologna Science Park location, which is a large space with an industrial history: it once used to house a giant tobacco company. Nearby is a high-voltage substation of an electricity company (Enel) [12].

The overall size of the park is about 100,000 m^2, of which approximately 18,000 m^2 are used for the data centres. The site also houses the ECMWF computers. CINECA and the Italian National Physics Institute (INFN) are both partners in the Leonardo Consortium. They have two large rooms at their disposal, and each has a surface area of about $1{,}600\ m^2$. One of those will provide ample space to house the Leonardo system (see Figure 2).

The power draw of the system will be between 8 and 9 MW. The currently available power envelope is 10 MW, in a configuration in which one powerline is backed up by another. There is an option to expand this setup to a similar 2N redundancy setup providing a power envelope of 20 MW by 2023. Backup for power is provided by 5×2 MW Diesel rotary UPS systems.

Figure 2: Bologna Science Park, Tecnopolo di Bologna, indicating the CINECA/INFN halls next to the ECMWF site

The currently available cooling capacity is also 10 MW, of which 8 MW is provisioned for cooling "hot water" Direct Liquid Cooled (DLC) compute systems, while 2 MW is geared towards providing air cooling for storage and ancillary systems. The cooling plant utilizes dry cooling towers as well as mechanical chillers. Four locally available wells, providing chilled water, are utilized to lower the amount of chilled water to be produced by the mechanical chillers[2]. The system is expected to be operated in the data centre with a PUE below 1.1.

2.1.3 MareNostrum 5, BSC, Barcelona

To support their vision of a heterogeneous performant system, in the long run, BSC needed to build a new data centre. The new data centre building, which also houses BSC headquarters and explicitly aims to be a visitable site and exhibition centre of MareNostrum 5, is being built and made ready for production, next to the renowned chapel that has been used to house all previous incarnations of MareNostrum.

The tender for the new building was formally awarded on 26.11.2019. Initially, the new data centre was expected to be ready for production in September 2020. However, due to COVID-19, there has been some delay. The date has now shifted to April 2021. This implies that no HPC system will be installed at the BSC facilities before this date.

A new electrical substation, to provide a total capacity of 2×25 MVA, had to be built underground, as well. The expected initial power consumption for the envisaged pre-exascale

[2] Ibid.

system is 20 MW. BSC intends to keep this facility operational for at least 15 years - as was previously the case with the chapel. The location of the subsystem has room to later expand its capacity to 2×40 MVA, if needed. The first installation phase included installing a power line of 1145 m connecting BSC to the general power grid.

An independent emergency line with a capacity of 5 MVA was also installed. The decision of which equipment can switch to the emergency power is part of the implementation of the systems at the new centre. The centre will rely on the emergency line exclusively in case of failure of the main power feed and will *not* have an on-site Diesel-powered generator to support the UPS infrastructure. BSC has a contract with their power provider that guarantees that all energy supplied is from renewable sources.

Figure 3: New BSC data centre

In the request for proposals for the MareNostrum 5 system, the following key data centre infrastructure-related requirements and limitations were formulated, reflecting in part, the current limits and capabilities of the newly built centre:

- The power consumption of the system, while running the HPL benchmark, must not exceed 12 MW
- The PUE of the data centre accommodating the system must be below 1.08
- The system's cooling requirements for a cold water loop (18 °C) to provide cooling capacity for air-cooled or rear-door heat exchange cooled components must not exceed a cooling capacity of 1 MW
- The system's cooling capacity for a warm water loop (35 °C) targets a capacity for cooling up to 12 MW
- Racks using the warm water loop for cooling must be able to dissipate 95% or more of their heat through this channel

The above list reflects the power and cooling resources available to the MareNostrum 5 system. They reflect the capabilities of the site-wide BSC data centre infrastructure only in part, as BSC plans to keep MareNostrum 4 in production as well at its current capacity, but with improved efficiency, as well as another high-end system to be housed in the new building.

The diagram, shown in Figure 4, was presented at the EWHPC 2020 online event and illustrates the site-wide supporting power and cooling infrastructure schematically. There are 5 transformers shown in the bottom part that transfer the electrical energy to the main distribution panel.

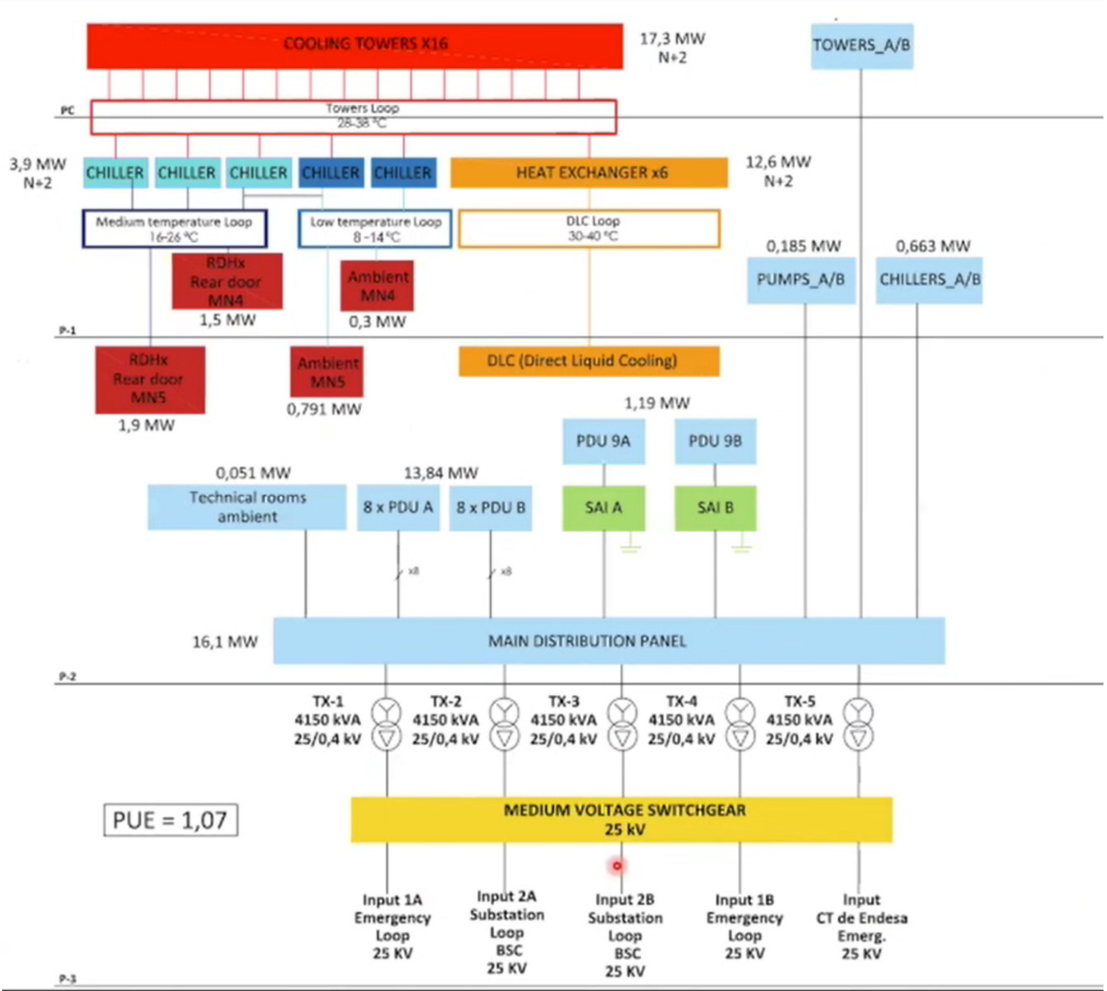

Figure 4: BSC site power and cooling infrastructure diagram

The main distribution panel, in turn, feeds the power to the Power Distribution Units (PDUs) and the UPSs.

The upper part of the diagram shows that the cooling plant contains chillers for non-DLC components on cold water loops. Sixteen cooling towers, with N+2 redundancy, on top of the new building, are operating within a temperature of 38 °C for the inlet and 28 °C for the outlet. With a flowrate of 1,500 m^3/h, the water then flows via heat exchangers to compute nodes. The cooling towers provide a total capacity for dissipating 17,300 kW of heat.

2.2 Exascale Computing Projects in the US

Around 2008, the US Department of Energy, in particular the Office of Science and the National Nuclear Security Administration, initiated funding for a range of exascale computing projects. These ongoing projects are being organised by the Department of Energy, mainly by the National Nuclear Security Administration [13]. In 2018, three research centres united and started the CORAL-2 project [14]. The main aim of this project is to build exascale machines in the US. Three exascale projects originated from the CORAL-2 project and are currently ongoing: Frontier, Aurora, and El Capitan. The following list shows which data centre is organising which of those three machines.

- Frontier [15]: Oak Ridge Leadership Computing Facility (OLCF) providing the computing facility to the government, researchers and academia. The OLCF has provided its service for the past 28 years and is currently hosting the second-fastest supercomputer (Summit) as of November 2020 (Top500). Frontier is expected to be operational around 2021.
- Aurora [16]: Argonne Leadership Computing Facility (ALCF) planning to host the Aurora exascale computing machine around 2021. The ALCF has been providing its computing resources to the research community for the past few decades.
- El Capitan [17]: Lawrence Livermore National Laboratory (LLNL) is scheduled to host the El Capitan in 2023. El Capitan's primary use will be dedicated to the National Nuclear Security Administration (NNSA). The NNSA is specialising in using computing facilities for nuclear-related simulation (science, engineering and weapons).

Table 1 provides the configuration overview of these exascale machines [18].

	Frontier	***Aurora***	***El Capitan***
CPU Architecture	AMD EPYC	Intel Xeon Scalable	AMD EPYC "Genoa" (Zen 4)
GPU Architecture	Radeon Instinct	Intel Xe	Radeon Instinct
Performance (RPEAK)	> 1.5 *EFlop/s*	1.0 EFlop/s	2.0 EFlop/s
Power Consumption	~ 30 *MW*	≤ 60 MW	< 40 MW
Footprint	> 100 Cabinets	N/A	N/A
Laboratory	Oak Ridge Leadership Computing Facility	Argonne Leadership Computing Facility	Lawrence Livermore National Laboratory
Vendor	Cray	Intel	Cray
Expected Year	2021	2021	2023

Table 1. Three ongoing exascale projects in the USA

2.2.1 Frontier

The OLCF will host Frontier, with more than 100 Cray Shasta cabinets with high density compute blades. These blades are going to feature HPC and AI optimised AMD EPYC processors and Radeon Instinct GPU accelerators. Each node consists of four GPUs connected to one CPU by the AMD Infinity Fabric links and supports coherent memory between them within the node. To support HPC and AI workloads at the exascale level, each node is configured with one Slingshot interconnect network port for every GPU with a streamlined communication between the GPUs [19]. The OLCF is currently hosting two petascale computers: Titan and Summit. Table 2 shows the hardware configurations for the present and future systems at the OLCF.

For cooling, OLCF is building a new cooling tower with a volume of 492 m3, which will deliver 18.9 m3 of high-temperature water per minute with the help of the 261 kW pumps. The water supply will be delivered to the data centre via a 152 m long, 0.6 m in diameter pipe. Frontier will be placed (in a 1,900 m2 room) where the Titan was installed. Titan was recently decommissioned, and the technicians dismantled 195,000 kg of components, which were sent for recycling. The floor is also going to be upgraded to support the new machine weight [20].

System	***Titan***	***Summit***	***Frontier***
Peak Performance	27 *PFlop/s*	200 PFlop/s	> 1.5 EFlop/s
Cabinets	200	256	> 100
Node	1 AMD Opteron CPU 1 Nvidia K20X Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 HPC and AI Optimised AMD EPYC CPU 4 Purpose Built AMD Radeon Instinct GPUs
CPU-GPU Interconnect	PCI Gen2	NVLINK Coherent memory across the node	AMD Infinity Fabric Coherent memory across the node
System Interconnect	Gemini	2x Mellanox EDR 100 GB/s InfiniBand Non-Blocking Fat-Tree	Multiple Slingshot NICs providing 100 GB/s network bandwidth. Slingshot dragonfly network, which provides adaptive routing, congestion management and quality of service.
Storage	32 PB, 1 TB/s Lustre Filesystem	250 PB, 2.5 TB/s, GPFS	2-4x performance and capacity of Summit's I/O subsystem. Frontier will have near node storage like Summit

Table 2. OLCF systems' hardware configuration

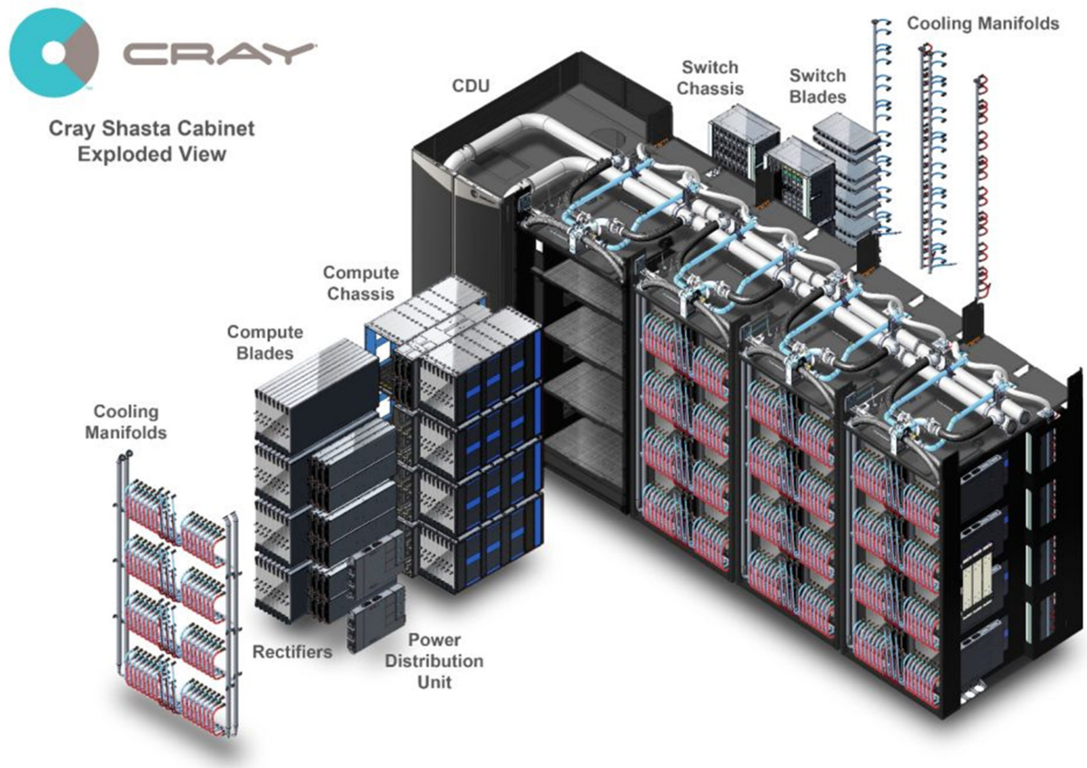

Figure 5: Cray Shasta cabinet exploded view [21]

2.2.2 Aurora

Aurora is based on Intel processors and accelerators. Each node will have 2 Intel Xeon scalable "Sapphire Rapids" processors and 6 Xe ("Ponte Vecchio") GPUs with a unified memory architecture. The interconnect between CPU and GPU is PCIe based and between the GPUs Xe Link is used. The system interconnect relies on the HPE's Cray Slingshot 11 with Dragonfly topology with adaptive routing [22]. For better I/O in the system, Aurora will feature Distributed Asynchronous Object Store (DAOS) to support a larger data workload. More than 9,000 nodes are expected to be placed in the Aurora machine. Intel's oneAPI will be available among other existing programming models. Aurora will be placed in the ALCF data centre, which is being upgraded for better power/cooling infrastructure.

2.2.3 El Capitan

El Capitan features the HPE Cray Shasta architecture, which will support a wide variety of processors and accelerators (see Figure 5). It will have a next-generation AMD EPYC processor, codenamed "Genoa", featuring the "Zen 4" processor core and next-generation AMD Radeon Instinct GPUs with Cray Slingshot HPC Ethernet interconnect. These GPUs are well optimised for HPC and AI applications. Each node will have 4 GPUs connected by 3rd Gen AMD Infinity Architecture for high bandwidth and low latency between the CPUs and GPUs. The El Capitan system will be placed in the NNSA data centre.

NNSA has launched a project in 2019 called Exascale Computing Facility Modernization (ECFM), which is scheduled to end in July 2022. This project's primary focus is to modernise (structural, architectural, mechanical, and electrical systems) building 453 in the Livermore

Computing Complex (B453) at the NNSA data centre. Additionally, it will upgrade the approximately 3.5 acres of the previously developed site adjacent to B453. Nevertheless, this project aims to provide enhanced power and cooling for the current and future computing machines at the NNSA. More specifically, the facilities' existing cooling tower capacity will be expanded to 100 MW from 35 MW. It will be achieved by installing a new low-conductivity water cooling tower and establishing the required piping and pumps. The power supply will be upgraded from 45 MW to 85 MW. Cooling at the NNSA is air-based, but it has been slowly replaced with water cooling infrastructure, reducing the air handlers from 1300 m3⁄s to under 950 m3⁄s. ASHRAE [36] W3 (water-supply temperature of 2 °C to 32 °C) is planned for water cooling. Presently the PUE for the NNSA data centre is between 1.1 and 1.2, but in the future, it is aimed to be around 1.02 [23]. The overview of the site upgrade is shown in Figure 6.

Figure 6: Building 453 in the Livermore computing complex [24]

2.3 Exascale Computing Project in Japan

The exascale computer Fugaku (named after Mount Fuji) has been operational in Japan since 2020 and is hosted by RIKEN. The Fugaku system is configured with ARM A64FX CPUs (ARMv8.2-A architecture with SVE 512-bit vector extension). Each node has 48 cores with 2 or 4 cores for OS activities. The processor is based on the 7 nm Fin Field-Effect Transistor (FinFET) [25], has a competitive logic density (integrates 8.786 billion transistors); thus, it suits very well for HPC. In total, Fugaku has 128,976 (432 racks) nodes with a total memory of 4.85 PB. System nodes are connected using the Tofu Interconnect D (data rate 28 Gbit/s x 2 lanes per link x 10 port) to facilitate high-density node configuration, link configuration & injection bandwidth and dynamic packet slicing. The PCIe Gen3 x16 supports I/O for Fugaku. Power consumption lies between 30 to 40 MW. The system achieves 1.07 EFlop/s (32-bit - single precision) and 2.15 EFlop/s (16-bit - half precision) in boost mode.

Fugaku has implemented some new strategies regarding cooling and power consumption. To minimise the power loss, the designers minimised the distance between the Power Supply Unit (PSU) and CPU Memory Unit (CMU), thus reducing the power consumption for the bus converter and overall power dissipation due to the distance. Figure 7 shows how it is implemented in Fugaku compared to the K computer. This modification overall reduces the power loss from 24% to 14%.

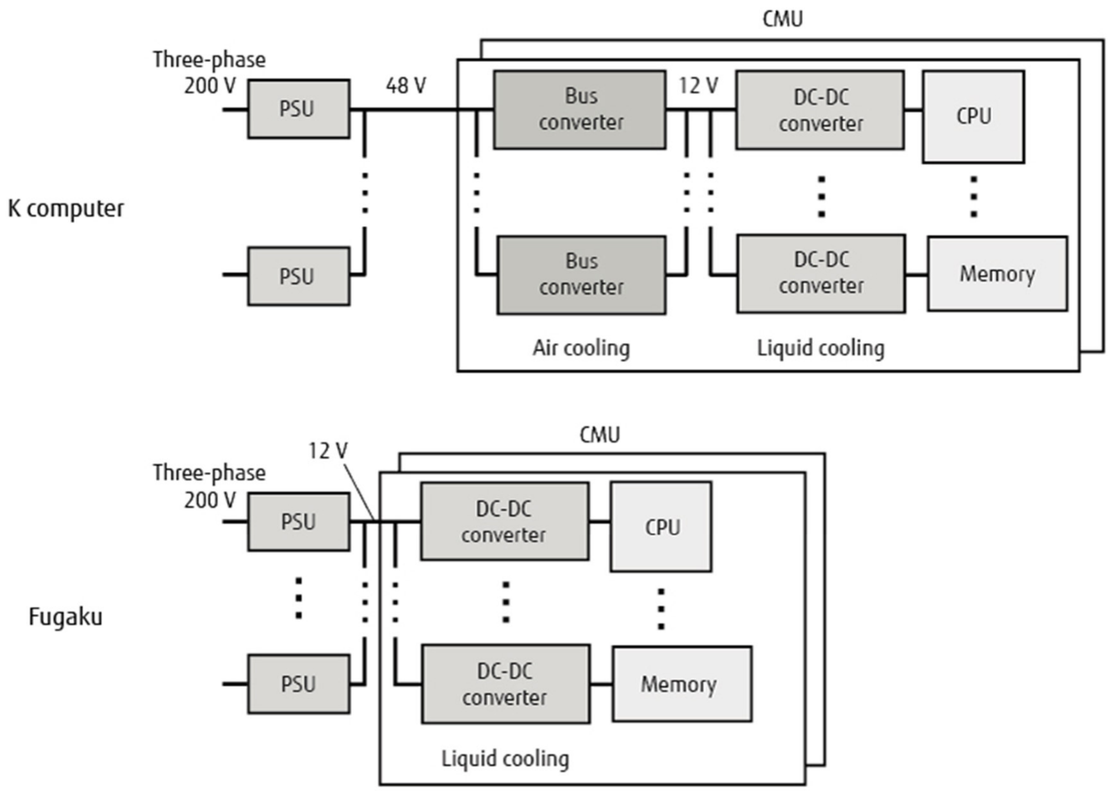

Figure 7: Power saving strategy in Fugaku [26]

The direct liquid cooling of Fugaku has introduced a fluid flow path in both directions compared to the K computer's single direction (see Figure 8), allowing for better heat rejection from the CPU.

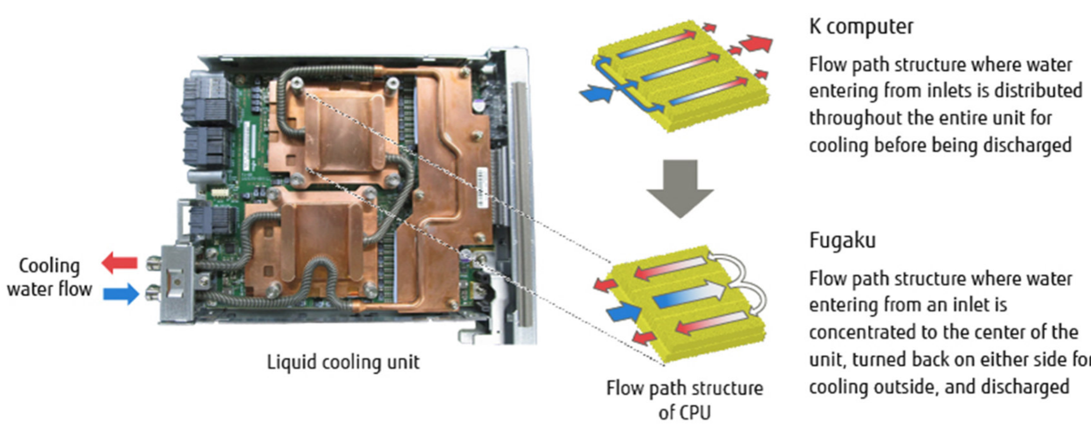

Figure 8: Different cooling liquid flow path between K and Fugaku computers

2.4 Exascale Computing Projects in China

In China, the exascale projects are coordinated by the three research units: National Research Center of Parallel Computer Engineering and Technology (NRCPC), Dawning Information

Industry, and the National University of Defense Technology (NUDT). The following list shows the ongoing exascale projects in China with the publicly available information we have [27].

- **NRCPC (Sunway prototype)**: is a pure CPU based prototype featuring Chinese ShenWei SW26010 CPUs with a total of 512 nodes using a homegrown interconnect network that provides a 200 Gbps of point-to-point bandwidth. Each node features 2 CPUs. A single CPU of SW26010 has 260 cores, which is already being used in Sunway's TaihuLight supercomputer. The Sunway exascale machine hosted in Tsingtao relies on a 28 MW cooling system from Climaveneta based on closed-coupled chilled water cooling with a customised liquid water-cooling unit. Further details are not yet disclosed.
- **SUGON (prototype)**: is a heterogeneous prototype system featuring two Hygon x86 CPUs and two DCUs (accelerators). The CPU is a licensed, cloned model from AMD's first-generation EPYC processor, whereas the DCU is built by Hygon. The interconnect of the system is homegrown with a 6D torus network. The Sugon system will be cooled by immersion cooling technology, which is proven to be very efficient for removing the processor's heat. It is mainly due to direct contact (conduction) of the liquid to the surface of the components. The corresponding exascale system is expected to be deployed in 2022 in Zhengzhou.
- **NUDT (Tianhe-3)**: is also a heterogeneous system, which will feature a Matrix-3000 DSP processor (considered as an accelerator) with an unknown future generation of CPUs. The Matrix-3000 is expected to have at least 96 cores, and the CPU is going to have 64 cores. The entire system will consist of 100 cabinets. Each cabinet will have 128 blades, and each blade is built with eight CPUs and will be paired with eight DSPs. The network is based on the 3D butterfly topology supporting the homegrown network. The data centre, hosted in Tianjin, is expected to have a PUE of 1.1 with a hybrid cooling using mixed air and liquid cooling methodologies. While the Tianhe-3 supercomputer, projected to deliver 1.29 $EFlop/s$ performance was initially announced for 2020, the global situation related to the COVID-19 pandemic postponed this release. The deployment of the system is however still expected for 2021.

3 Survey

For further identification of installation requirements, PRACE-6IP WP5 has conducted an online survey that was distributed among EuroHPC JU pre-exascale and petascale sites [2] as well as among additional PRACE Tier-0 and Tier-1 sites [3] not involved in the mentioned EuroHPC JU public-private partnership programmes. Section 3.1 provides further details concerning the participating supercomputing sites.

The survey consisted of 22, mainly multiple-choice, questions. The survey was distributed on 08.06.2020 and as of December 2020, had attracted 19 sites out of 23 approached. This survey was distributed with the help of an open-source survey tool, LimeSurvey, hosted at BADW-LRZ.

3.1 Participants

The following supercomputing sites that will host the three pre-exascale supercomputers [2]:

- CSC, Finland (hosting the LUMI system)
- CINECA, Italy (hosting the Leonardo system)
- BSC, Spain (hosting the MareNostrum 5 system)

as well as the following EuroHPC petascale sites:

- IT4Innovations National Supercomputing Center, Czech Republic (hosting the EURO_IT4I system)
- Luxprovide, Luxembourg (hosting the MeluXina system)
- Sofiatech Park, Bulgaria (hosting the PetaSC system)
- Institute of Information Science in Maribor, Slovenia (hosting the Vega system)

have participated in the survey. In addition, the following PRACE Tier-0 and Tier-1 sites have participated in the survey [3]:

- CEA, France
- JSC, Germany
- BADW-LRZ, Germany
- CINECA (data centre in Casalecchio)[3], Italy
- IDRIS (GENCI), France
- GRNET, Greece
- KIFU, Hungary
- UIO (Sigma2), Norway
- PSNC, Poland
- CCSAS, Slovakia
- PDC, Sweden
- SURFsara, The Netherlands
- EPCC, UK

3.2 Results

3.2.1 Major challenges when deploying next generation HPC system

One of the questions asked in the survey was to rank the foreseeable challenges for data centres when deploying a next generation (i.e. pre-exascale, exascale) HPC system. The presented bar chart in Figure 9 indicates the recorded average importance across all the responses received: the lower a number, the more relevant an aspect is.

[3] Different from the data centre hosting the pre-exascale system, changing the considered total number of surveyed sites to 20.

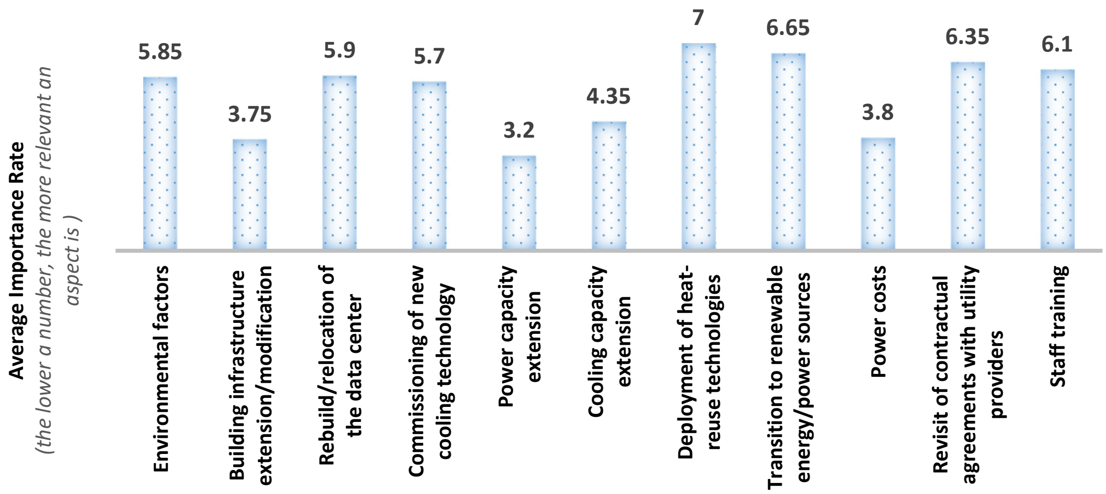

Figure 9. Average importance of identified challenges for hosting next generation HPC system

As can be seen, the three major challenges identified are power capacity extension, followed by building infrastructure extension/modification and power costs – once again indicating the continuous necessity of energy/power-efficiency improvements across all the pillars of a modern data centre.

3.2.2 Facility aspects

The extension/modification of the building infrastructure was indicated as one of the major challenges for hosting a next generation HPC system. Figure 10 provides an overview on the currently available infrastructure floor space of the surveyed European HPC centres. In addition, around 35% of the surveyed supercomputing sites have indicated their intention to extend their infrastructure floor space for hosting a next generation exascale system.

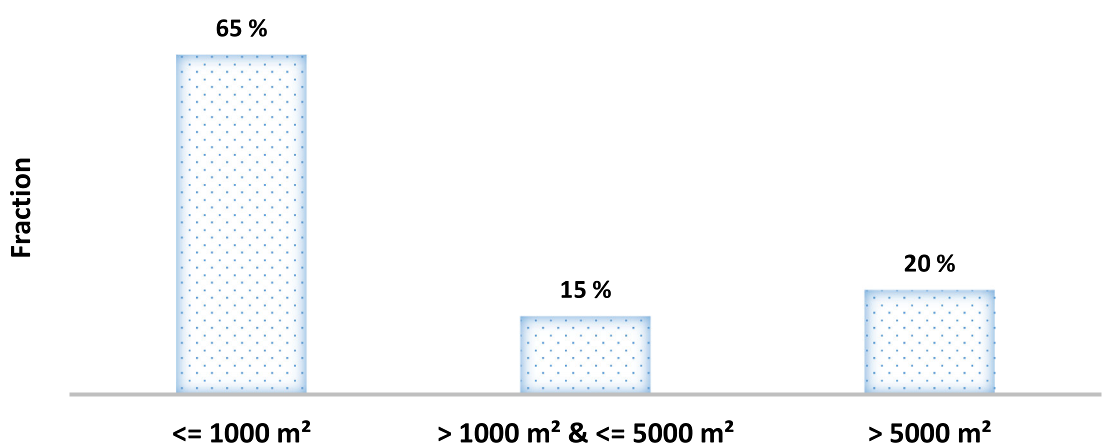

Figure 10. Currently available infrastructure floor space of the surveyed sites

Figure 11 outlines further the occupied area of the flagship HPC systems at the surveyed supercomputing sites.

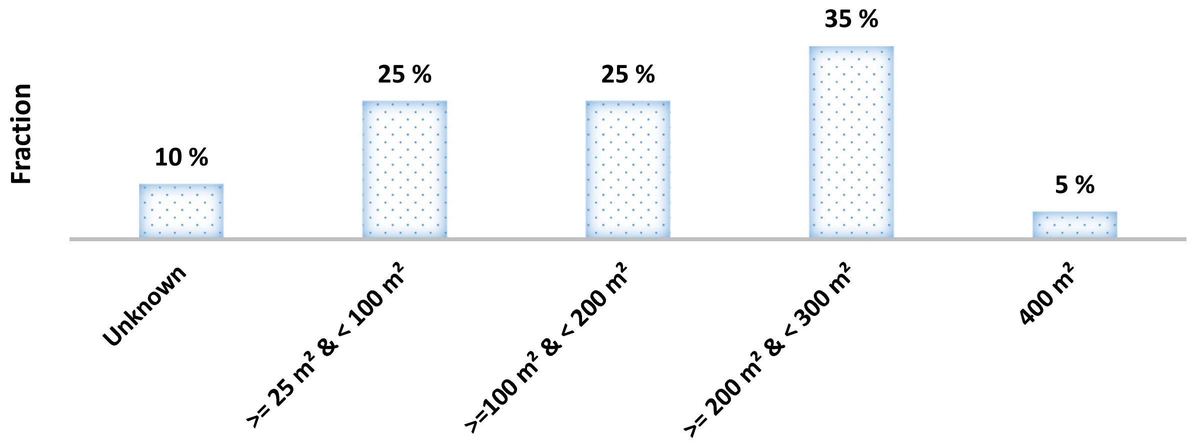

Figure 11. Currently occupied area by the flagship HPC systems at the surveyed sites

Sites were further asked if they expected their next generation HPC systems to occupy more floor space, as compared to the occupied space by the current flagship systems. Figure 12 illustrates their expectations.

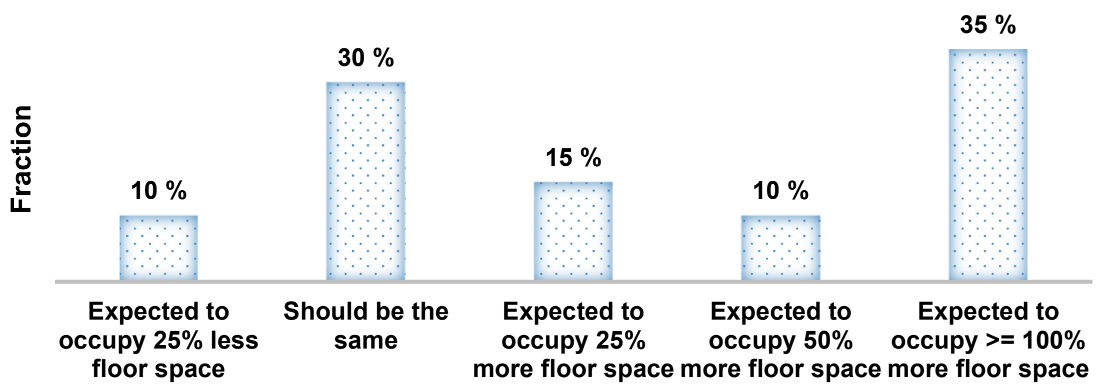

Figure 12. Expectations regarding the foreseen floor space occupation compared to the current usage by the flagship systems

During the recent European Workshops on HPC Infrastructures, the purpose of the raised floors was actively questioned, given the costs and projected weight increase of the compute racks. The current weight of a compute rack, as indicated by the surveyed sites, varies from 500 kg to 2,200 kg, and around 75% of surveyed sites expect the compute racks of next-generation systems to be heavier due to foreseen increase in server densities. This might imply that data centres will in future request system vendors to strengthen the raised false floors and carry associated costs in case the delivered racks are heavier than a given threshold.

For this reason, the surveyed sites have been further asked if they would like to move away from the usage of the raised floors given the possibility of a data centre redesign. Interestingly enough, 80% of the surveyed sites indicated that they will not be in favour of the elimination of the raised false floors. The sites justified that by the current space limitations, difficulties that would arise for efficient air-cooling while emphasising the ease of under-rack plumbing in case of leakages and possibilities of installation separation (e.g. water and power) given a raised floor solution.

3.2.3 Power availability and usage of renewable energy sources

Power capacity extension along with power costs were the other two major challenges identified by surveyed European HPC sites for hosting a next generation supercomputer.

Figure 13 illustrates the current power capacities of surveyed data centres. As can be seen, around 10% of surveyed data centres have less than 1 MW power capacity. On the other hand, approximately 60% of surveyed sites indicated to have more than 1 MW and less than 10 MW power capacity, and 30% to have surveyed sites reported to currently have more than 10 MW power capacity.

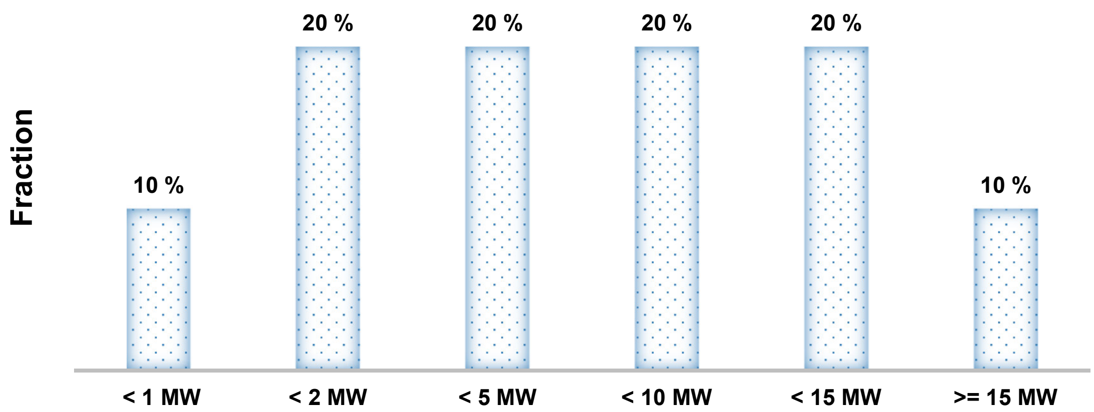

Figure 13. Current power capacity of surveyed data centres

The survey further revealed that 70% of the surveyed sites intend to increase their power capacities even further. Figure 14 provides a detailed view. As can be seen, 70% of surveyed sites intend to increase their current power capacities, with 20% of the sites intending to increase the capacity by 50% and 25% of the sites beyond 100%.

Given the rising power costs and the various incentives introduced by different countries for the reduction of CO_2 emissions in a cost-effective manner (examples include Germany with its "Energy Wende" program, the Netherlands, etc.) [29] the question of interest was whether or not the surveyed sites already use renewable energy for powering their flagship HPC systems. As can be seen from Figure 15, 40% of surveyed data centres currently power their flagship systems using renewable energy sources, while 30% partially rely on renewables for powering their high-end IT systems.

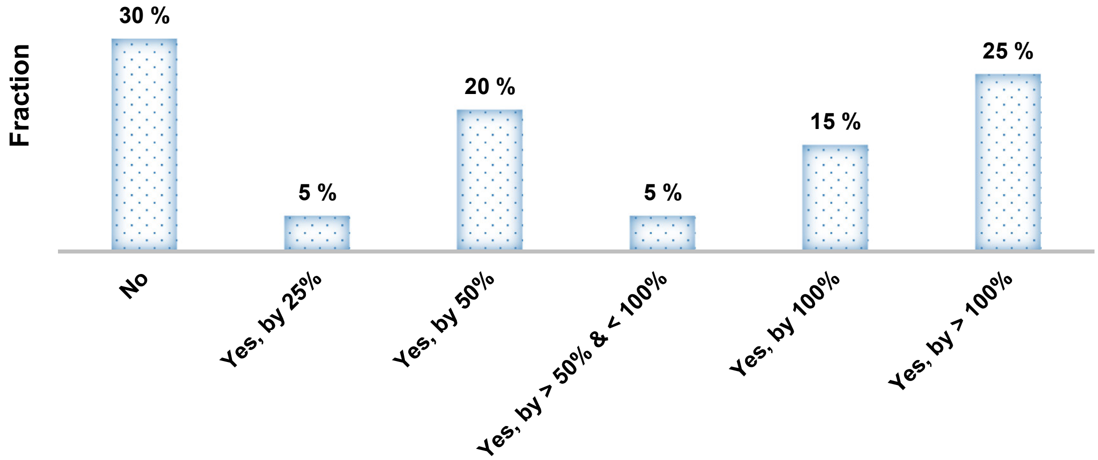

Figure 14. Planned extensions of power capacity

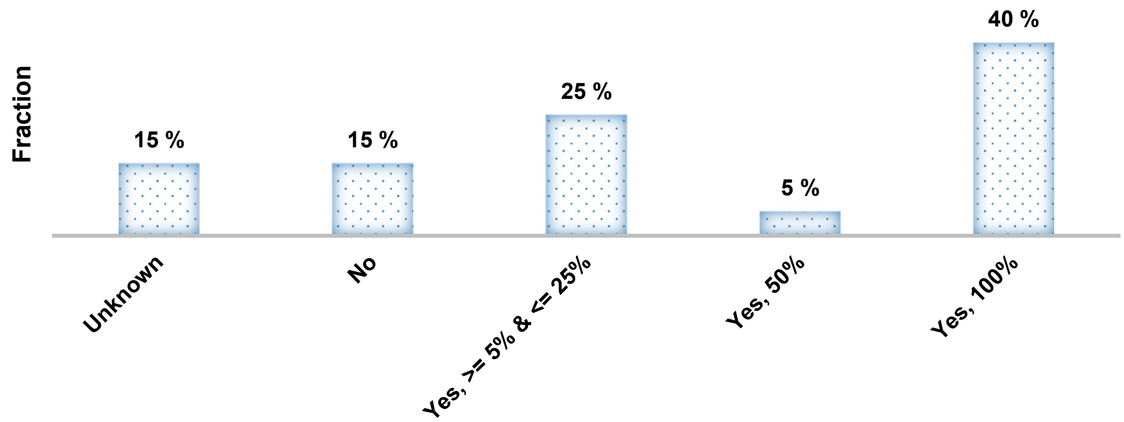

Figure 15. Fraction of renewable energy usage for powering the flagship system

Figure 16 provides an overview of the foreseen intentions regarding the usage of renewable energy sources. As can be seen, there are mixed intentions concerning the future. 25% have indicated that they are not planning to use renewable for their next system, and 15% indicated that they are going to look into the possibilities of renewable energy sources for powering their flagship system to some extent.

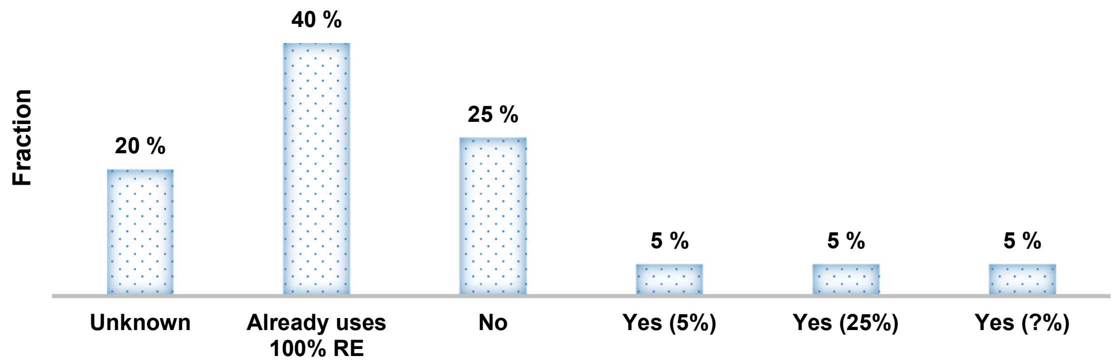

Figure 16. Foreseen intentions regarding the usage of renewable energy sources for powering next flagship system.

3.2.4 Cooling solutions

Given that the energy consumption of certain cooling infrastructure solutions might account for approximately 40% of the data centre's power consumption [30], [31] it was essential to understand the currently used cooling technologies as well as to get an overview on foreseen usage of various cooling solutions for operating next-generation HPC systems. Figure 17 outlines the currently used cooling technologies by the surveyed data centres for the heat removal from their flagship HPC systems.

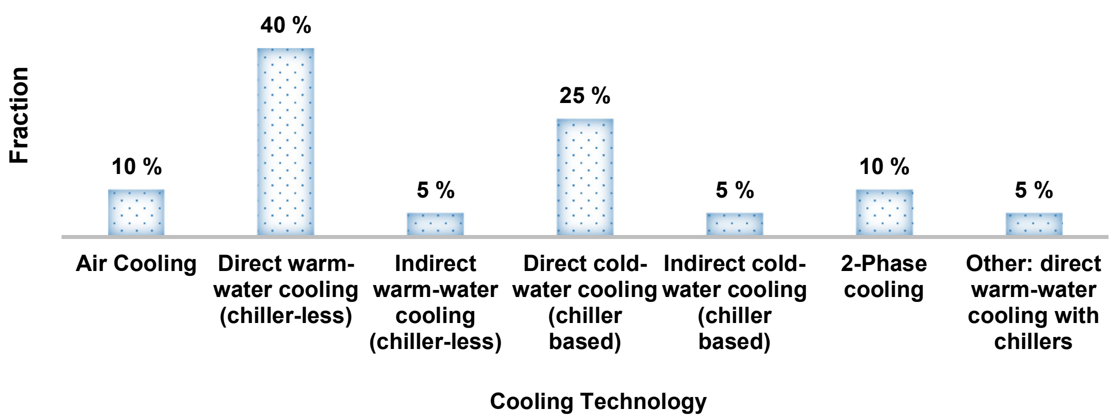

Figure 17. Currently used cooling technologies

As of now, the most popular cooling technologies seem to remain direct warm-water cooling (chiller-less) and direct cold-water cooling (chiller based) followed by 2-Phase cooling and air cooling. Around 75%, however, indicated the need to revisit the existing cooling solutions when hosting the next-generation supercomputing system. The reasons are manifold, ranging from the migration to new data centre facility through the extension of the cooling capacity while ensuring the operation of older generation HPC systems to better utilisation of free cooling capabilities.

3.2.5 Foreseen major investment areas

For understanding the current major focus areas of modern large-scale European HPC centres in preparation for hosting a next generation HPC system, surveyed sites were also asked to indicate the foreseen investment areas. Figure 18 outlines the results. As for the case with Figure 9, here also the bars indicate the recorded average importance across all the responses received with a lower rate indicating higher relevance.

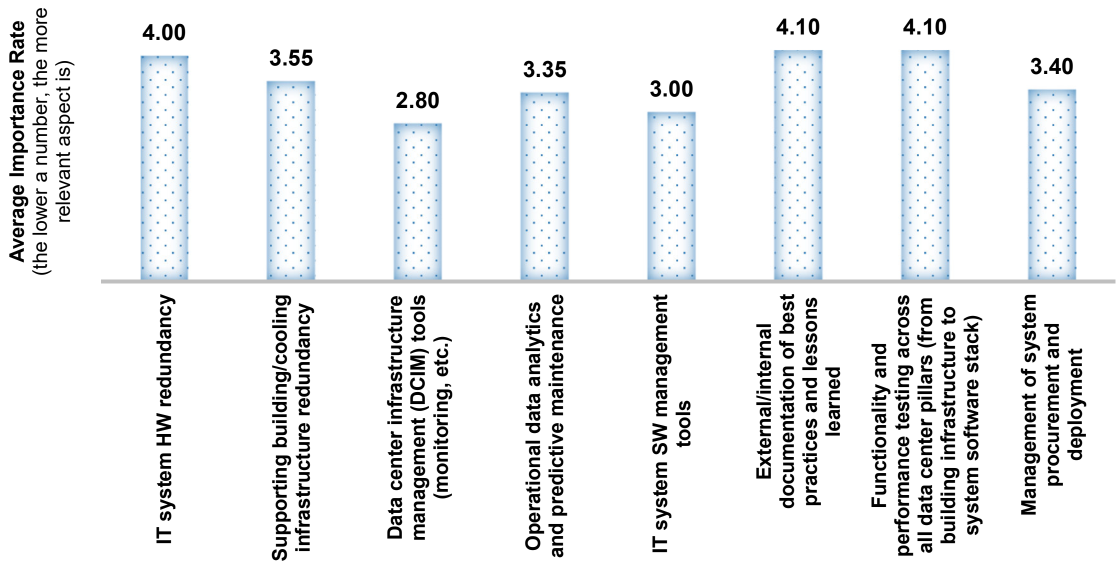

Figure 18. Areas in which the surveyed sites are going to invest for preparation of (during) next generation HPC system deployment

As can be inferred from Figure 18, the top three investment areas indicated by surveyed European HPC sites are:

- data centre infrastructure management (DCIM) tools (monitoring, etc.)
- IT system SW management tools
- operational data analytics (ODA) and predictive maintenance

It is worth noting that some of the PRACE sites, namely CEA [32] and LRZ [29], have already developed ODA prototype frameworks that look at the optimisation of the underlying building/cooling infrastructure. The latter is also currently being further assessed for applicability at one of the partner Energy Efficient High Performance Computing Working Group (EE HPC WG) [33] sites [34].

4 Main installation requirements for hosting next generation European HPC systems

4.1 Categories of installation requirements

Understanding the main installation requirements for hosting future HPC systems well in advance is a major concern:

- For HPC sites planning major long-term evolutions of their hosting infrastructure and for being able to procure systems without putting unreasonable installation constraints on the systems to be procured.
- For entities like EuroHPC willing to conclude hosting agreements with HPC sites and for selecting sites able to provide convenient hosting for systems to be procured.
- For (European) system vendors and integrators as well as possibly for technology providers for better understanding and architecting solutions not only satisfying contemporary requirements of HPC user community but also addressing the constraints and peculiarities of the hosting entities.

These requirements can be organised[4] into the following domains:

- Power supply
- Cooling
- Floor space (computer room)
- Network connectivity
- Monitoring
- Environmental impact
- Security
    - Physical access security
    - Fire mitigation
    - IT access security
- Service
    - On-call service teams
    - Measurement of the users' satisfaction

The purpose of this chapter, based on the two previous ones, is to summarise the main installation requirements expected for hosting next generation HPC systems with a special focus on European exascale systems[5] around 2024. It covers only the first six items of the list above as security measures and service are somehow independent of the size of the systems – current requirements may be kept for future systems.

4.2 Trends in terms of installation requirements

In order to understand the trends in terms of installation requirements for future systems, it is of interest to summarise the most important relevant data presented in the second chapter of this document. The table below summarises this data for EU pre-exascale systems, US plans for exascale systems and Japanese plans for an exascale system. It highlights the power consumption of future large systems as well as the ambitious targets in terms of PUE in order both to reduce the electricity bill but also to reduce the environmental impact of the systems.

[4] The organisation of these requirements draws from the organisation of the hosting requirements described in the EuroHPC "Call for expression of interest for the selection of Hosting Entities for Precursors to Exascale Supercomputers" (ref: EUROHPC-2019-CEI-PE-01) issued in January 2019.
[5] Exascale means here a system with an aggregate HPL performance of at least 1 exaflops.

System	Site	Planned Operation Date	Peak perf. (PFlop/s)	HPL perf. (PFlop/s)	Power consumption (MW)	PUE	Floor space (m^2)
LUMI	CSC	2021	550	375	8.5 (HPL)	1.03	200
Leonardo	CINECA	2021		248	9 (HPL)	< 1.1	360
MN5	BSC	2021/2022		> 200	< 12	< 1.08	< 847
Frontier	ORNL	2021/2022	> 1500		30 – 40		
Aurora	ANL	2022/2023	> 1000		60		
El Capitan	NNSA	2022/2023	2000			1.02	
Fugaku	RIKEN	2020/2021	537	442	30 (HPL)	1.1	

Table 3. Information publicly available on pre-exascale and exascale systems (empty cells correspond to a lack of publicly available information)

This table, in addition to the information gathered during the annual European Workshop on HPC centre infrastructure, makes it possible to identify the key elements to take into account when it comes to trying to anticipate the needs for hosting new generation HPC systems and especially exascale systems. These keys elements, and their main consequences, can be summarised in the following way:

- The increase of power consumption.
- The considerable fluctuations of power consumption.
- The environmental impact needs to be addressed with more attention as the power consumption increases, and consciousness of its importance[6] is growing.

These factors have direct consequences in terms of target PUE, heat-reuse, and low-carbon electricity to reduce the electricity bill and minimise the environmental impact. It can also impact the relationship with utilities providing power to the site and the need to put in place a power capping mechanism[7] in order to stay within the available power envelope, which is also useful for avoiding preparing an oversized power distribution system.

In addition, a limited increase in the physical size of systems and possibly a significant increase in the weight of racks can be expected. However, it is unlikely that the weight per rack will lead to a floor load higher than the value set in the installation requirements for the precursors to exascale supercomputer as this would strongly limit the possibility to sell such systems to the current large HPC centres.

It should be noted that these challenges are understood and taken into account by the sites involved in the survey presented in the previous chapter. For example, most of the sites are planning a substantial increase of their power capacity and are willing to increase the usage of low-carbon electricity.

[6] See for example: Shaping Europe Digital Future, European Commission, Feb. 2020.

[7] A power-capping mechanism able to limit the power consumption of the supercomputer without large impact on application performances should be a mandatory requirement when procuring a supercomputer.

4.3 Suggested installation requirements to take into consideration for future HPC systems

The purpose of this section is to suggest what installation requirements should be taken into consideration for future HPC systems. In turn, when implemented at the level of an HPC centre, some of these requirements may also be considered as requirements towards system vendors or integrators.

The content of this section derives from the two previous sections and should be considered as an estimate to be updated over time when more data from systems or suppliers become available. It is also based on the experience of the team involved in the preparation of this deliverable and more generally on the experience of the experts engaged in WP5 as well as on information from previous European Workshop on HPC Infrastructures and on data publicly available such as Green500. For illustration, figures are shown for exascale supercomputers with an architecture mix similar to the precursors to exascale supercomputers. For systems of smaller size, most figures can be scaled down linearly.

	Requirements set in the EuroHPC JU "Precursors to Exascale Supercomputers" call (for operation around 2021)	**Suggested requirements for Exascale supercomputers (for operation around 2024)**
Power supply	a) Power capacity and power quality for hosting a system in the range of 10 to 15 MW total consumption for the pre-exascale supercomputer. b) UPS power available to cover the critical systems, including storage and access to data of the JU system. c) Ability to perform at least a Level 1 measurement quality [35] for a Top500 submission.	a) Power capacity and power quality for hosting a system in the range of 20 to 25 MW total HPL consumption[8] for the exascale supercomputer. b) UPS power available to cover the critical systems, including storage and access to data (5% to 10% of the total power capacity) with a duration defined according to the quality of the power feed (from historical data) and to the availability of diesel generators. c) Ability to perform at least a Level 1 measurement quality [35] for a Top500 submission. d) Ability to handle large power usage fluctuation without compromising systems availability and accessibility[9] e) Low power losses (from electrical distribution and conversion) f) Floor space needed for the power supply system
Cooling	a) Enough capacity of air or liquid cooling for hosting the system of the EuroHPC Joint Undertaking.	a) Enough capacity of air or liquid cooling for hosting the system b) Chiller-less liquid cooling[10] c) Efficient air-cooling system

[8] Based on the current and expected performance/watt ratios shown in the table of the previous section and on the expectation of a moderate improvement over time in accordance with the latest trends from the Green500 list.

[9] This needs to be further defined later in terms of duration and amplitude of power swings.

[10] ASHRAE [36] water reference W4 and W5 classes. This needs to be further defined later: one could for example ask for a chiller-less system able to provide water at a temperature lower than 35 °C 95% of the time and able to do so the remaining 5% of the time with the assistance of chillers.

		d) Floor space needed for the cooling system
	Requirements set in the EuroHPC JU "Precursors to Exascale Supercomputers" call (for operation around 2021)	**Suggested requirements for Exascale supercomputers (for operation around 2024)**
Floor space	a) At least 700 m^2 of contiguous floor space available for the hosting of the EuroHPC supercomputer and its auxiliary systems b) Raised floor able to bear at least 2200 kg/m^2 distributed load	a) 700 m^2-1000 m^2 of floor space available for the supercomputer and its auxiliary systems b) Raised floor or concrete slab (in case no raised floor) able to bear at least 2200 kg/m^2 distributed load
Network connectivity	a) At least 100 Gbit/s connectivity towards the rest of the GÉANT Network (link capacity)	a) At least 400 Gbit/s connectivity towards the rest of the GÉANT Network (link capacity)
Monitoring		a) Capacity to implement a holistic monitoring approach, i.e. monitoring ranging from the facility over the deployed heat reuse systems and high-end IT equipment to large-scale applications b) Potential, thanks to this approach, for optimising the operation
Environmental impact		a) At least 80% of low-carbon electricity b) Heat reuse for heating and/or adsorption chillers providing an ERE larger than 0.5 [37] c) PUE less or equal to 1.1

Table 4. Summary of the suggested installation requirements for hosting next generation HPC systems

5 Best practices for hosting European pre-exascale and exascale systems

5.1 Overview of global data centre energy consumption and sources of renewable energy

Greenhouse gas emissions are the main reason for global warming, mainly by human-made activities, such as industries and deforestation [38].

There are many international and local organisations around the world working on solving the global warming issue. For example, EU countries (European Green deal) aim to reduce greenhouse emission and increase renewable energy resources by 2030. The following list shows a few of the highlights drawn by the European Commission [39].

- At least 40% cuts in greenhouse gas emissions (from 1990 levels)
- At least 32% share of renewable energy
- At least 32.5% improvement in energy efficiency

Figure 19 and Figure 20 show an overview of the worldwide power consumption in the data centres.

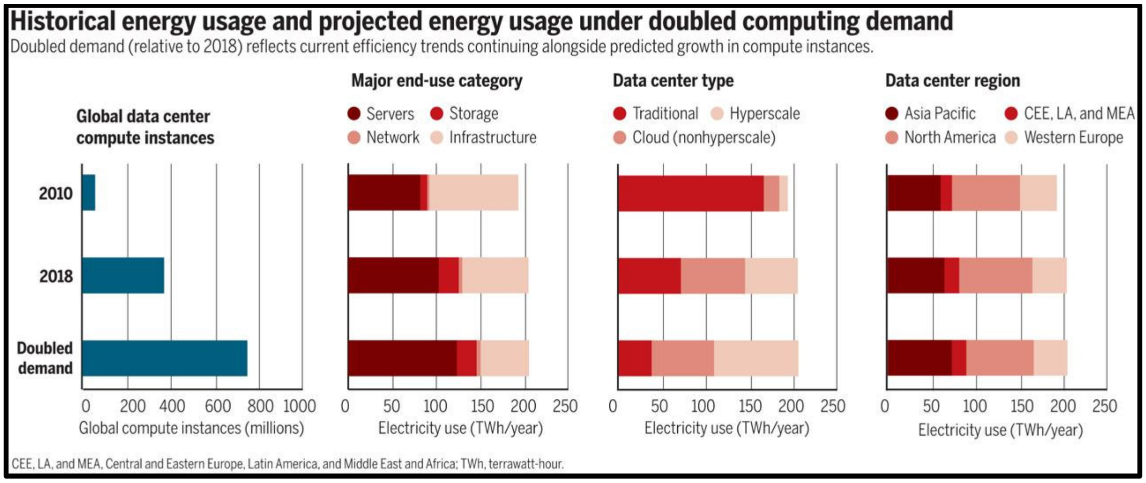

Figure 19. Historical energy usage and projected energy usage under doubled computing demand [40]

As the demand for computing power and storage will increase further over the years to come, the main question asked by the policy makers would be whether or not the provided computing is green? In addition, green computing might get subsidiaries from a local or international organisation or the government in the future.

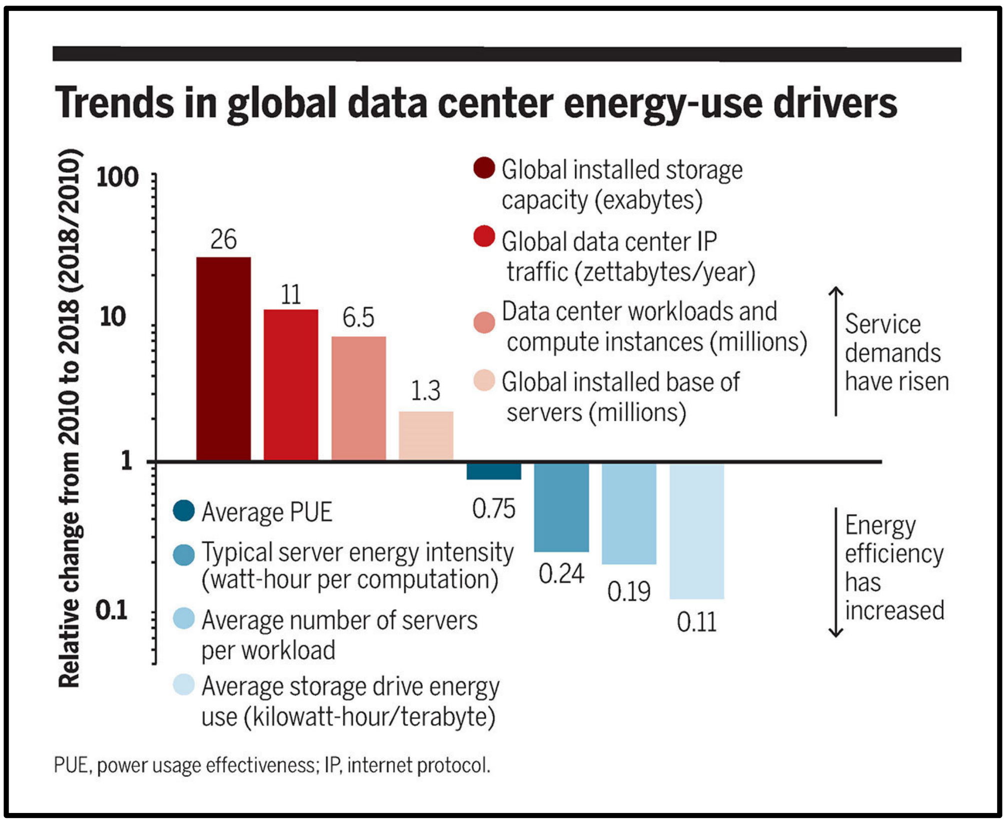

Figure 20. Energy usage in the data centres globally [40]

One of the ways for achieving green computing is to use renewable energy sources (e.g. solar, wind, etc.), which might be considered as an additional (main) power source for a supercomputing facility.

5.2 Overview of the cooling technologies for the data centres

The computer components at the data centre have some working temperature range, which should be maintained to keep the computer parts working correctly. Taking away the excess heat from the computer components is required for them to stay healthy and operate without any malfunction in the system. With the technological advancements in science and engineering, various cooling technologies are presently available to ensure efficient heat rejection in modern data centres.

First, we differentiate between air cooling and liquid cooling. Air cooling is the type of cooling when air is used to transport the heat inside the computing centre from the actual computers to some cooling device. If a liquid is used to transport the heat from the computers, that is called liquid cooling or water cooling if the liquid is water. For computer systems designed to be liquid-cooled, there is an opportunity for considerable energy savings compared to air-cooled systems. Since liquids have more heat capacity than air, smaller volumes can achieve the same level of cooling and can be transported with minimal energy use. Liquids can also easily be lead in narrow pipes to places where cooling is needed – this is due to the smaller volume of liquid required to transport the same amount of energy. In addition, if heat can be removed through a fluid phase change, heat removal capacity is further increased. The disadvantage with liquid cooling is the higher investment cost for liquid-cooled computer systems.

Air cooling

Air cooling exists in many forms that are more or less efficient. Here we start with the less efficient:

Computer Room Air Conditioner (CRAC): It is very much similar to an air conditioner, where the compressor in the refrigeration unit is cooling the air, and the heat is transported outdoors with the refrigerant, where the refrigerant is cooled in a condenser. This method is not very efficient because the compressor and the required fans draw considerable amounts of energy. It can also be challenging to spread the cooled air evenly in the computer hall. Cool air and hot air can easily mix.

Computer Room Air Handler (CRAH) with a Chilled Water System: Here, chilled water will be supplied to the server room. This chilled water will be flowing to the coils (array of pipes), and next to the coils, there will be fans that will make hot air contact with the chilled water flow coil. By doing this, there is convection taking place between the chilled water coil and the hot air. Finally, warm water will be sent back to some cooling device to get cooled by chillers or cooling towers and sent back to the server room. If the temperature is sufficiently high, a cooling tower can be used. Figure 22 shows a simplified schematic view of the chilled water system in the data centre and the difference between CRAC and CRAH. This type of cooling is widely used in small-scale or medium-scale data centres [42]. The chillers or cooling towers will cool down hot water.

Cold Aisle/Hot Aisle: Here, cold air is fed into one side of the server cabinets, and hot air is coming out from the other side of the server. The exhausting system (i.e. ventilation) draws hot air from the hot aisle. It is then sent back to the Computer Room Air Conditioning (CRAC) or Computer Room Air Handler (CRAH), where it is being cooled before being sent back to the server room [43]. Figure 21 shows the working model of the cold aisle/hot aisle methodology. This is one of the traditional ways of cooling the server room, and it is somewhat more efficient than using CRAC/CRAH units without any steering of the air-flow. The cold/hot aisle can be further improved by encapsulating the air, so it goes directly to where it is needed without mixing too much with other air in the computer hall.

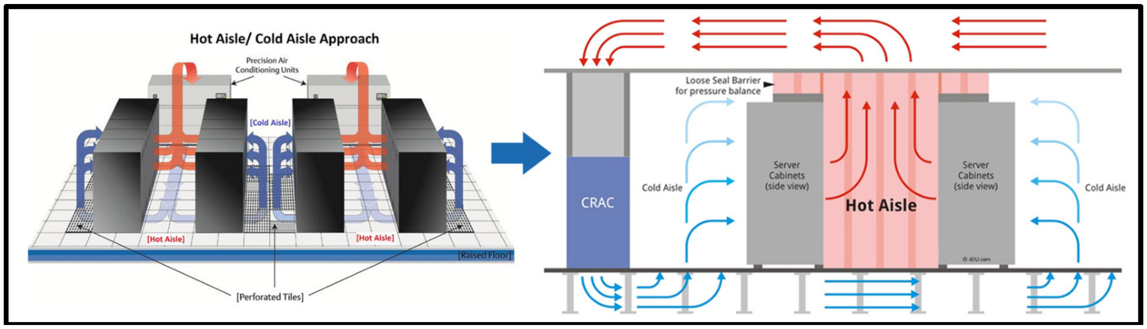

Figure 21. Cold aisle/heat schematic view [67]

Calibrated Vectored Cooling (CVC): CVC is developed by IBM and is suitable for higher density servers [41]. The working principle of the CVC is to transport the air to the electronics in exact amounts where needed. This methodology can be implemented in a blade server system in a small enclosure/chassis.

Free Cooling: It is referred to as using the outdoor air for the cooling without compressors. The term is used both for CRAH systems cooling towers (dry or wet) without compressor-based chillers and for systems where the outdoor air is just brought into the computer hall with fans. Used air can just be let out in the open air. This kind of methodology is applicable in a colder climate, where the outdoor air temperature is low. The latter works well if the air is clean but may require filtration or an air-to-air heat exchanger.

Liquid Cooling

Liquid cooling refers to the case when liquid is transported to the computer enclosure in some way to cool the computer. Inside the enclosure, air or liquid can be used to cool the components that require cooling. Using liquid cooling instead of air cooling generally has a lower energy overhead.

Hybrid liquid and air cooling: This refers to a mix of liquid cooling and air cooling where the fluid is transported to a cooler that is integrated with the rack, for example, as a cooled door on a rack that is cooled by air internally. This approach has an efficiency between air- and liquid-cooling, and the investment cost is also in the range between air- and liquid cooling. One advantage is that standard servers can easily be used. Due to the kind of servers, extra fans may be required and can, in that case, be an integrated part of the door design.

Direct-to-chip Liquid Cooling (DLC): This is a very efficient method of liquid cooling because of the direct contact with the cold plate of the computer component. A liquid coolant will pass through the cooled components' hot plate or cooler. Often CPUs and memories are the cooled components. Later, hot water is sent back (often via heat-exchangers) to an outdoor cooling device or cooling tower, where it will get cooled by the air. In most climates, chillers can be avoided since the temperature of the cooling liquid can be high (however, somewhat lower than the allowed surface temperature of the component). Components that are not directly liquid-cooled will still heat the air in the room or in the rack and may still need some air-cooling.

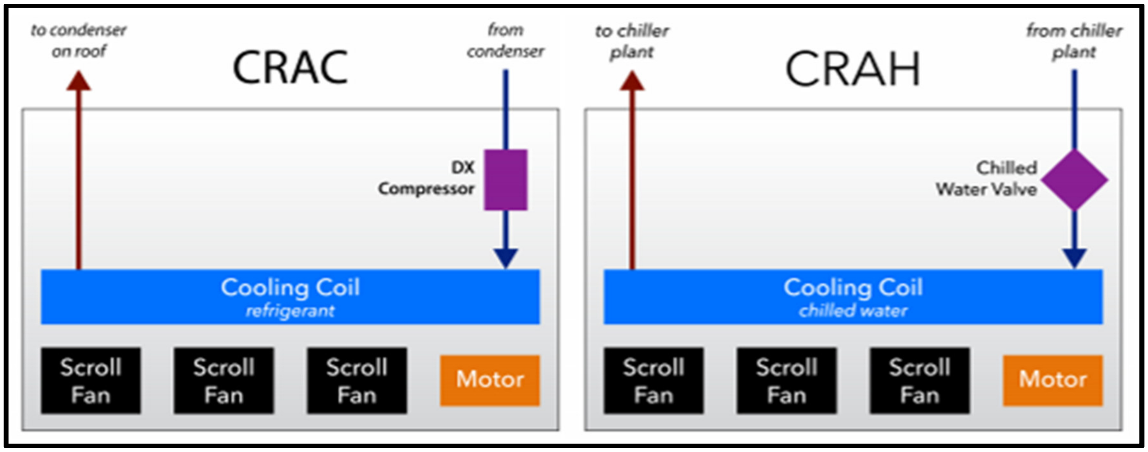

Figure 22. CRAC vs. CRAH schematic view [68]

Immersion system: This is referred to as keeping the computer circuit boards or even complete servers entirely enclosed into dielectric liquid cooling the system [45]. Here, most of the heat is transported away with the cooling fluid, and the computer room will be only heated indirectly.

5.3 Overview of few data centres that use renewable energy for power and cooling

This section discusses the usage of above-mentioned cooling solutions at some large-scale HPC data centres. And it also shows how these large scale HPC data centres use some of the renewable energy categories discussed earlier in this chapter.

LUMI, CSC, Kajani, Finland

As mentioned in Section 2.1.1, the CSC - IT centre for science will operate the LUMI pre-exascale supercomputer. LUMI will be using renewable energy for both heating and cooling. More generally, several private and commercial organisations operate data centres in Nordic countries (Sweden, Finland, Norway and Denmark), which offer their geographic location as a unique capability in the EU landscape. More importantly, power, heating and cooling supply can be made efficiently and environmentally friendly. Typically, Nordic countries have an eco-friendly approach for the data centres, which are as follows:

- Renewable energy: all the countries in the Nordic region produce energy (ca. 40% of the total production) from hydropower, wind turbines, biomass and waste heat [47]
- Reuse of the waste heat for district heating and greenhouse farms
- The cold climate of the Nordic countries allows minimising the energy used for the cooling while reaching among the lowest PUE rates in Europe

Leonardo, CINECA, Bologna, Italy

CINECA in Bologna, Italy (see Section 2.1.2) is hosting one of the EuroHPC JC pre-exascale machines. This machine will be installed in the Bologna science park and it is expected to reach PUE < 1.1. This park is already hosting the data centre of the European Centre for Medium-Range Weather Forecasting (ECMWF). In 2023, the ENEA Bologna Research centre will also be placed here; see Figure 23.

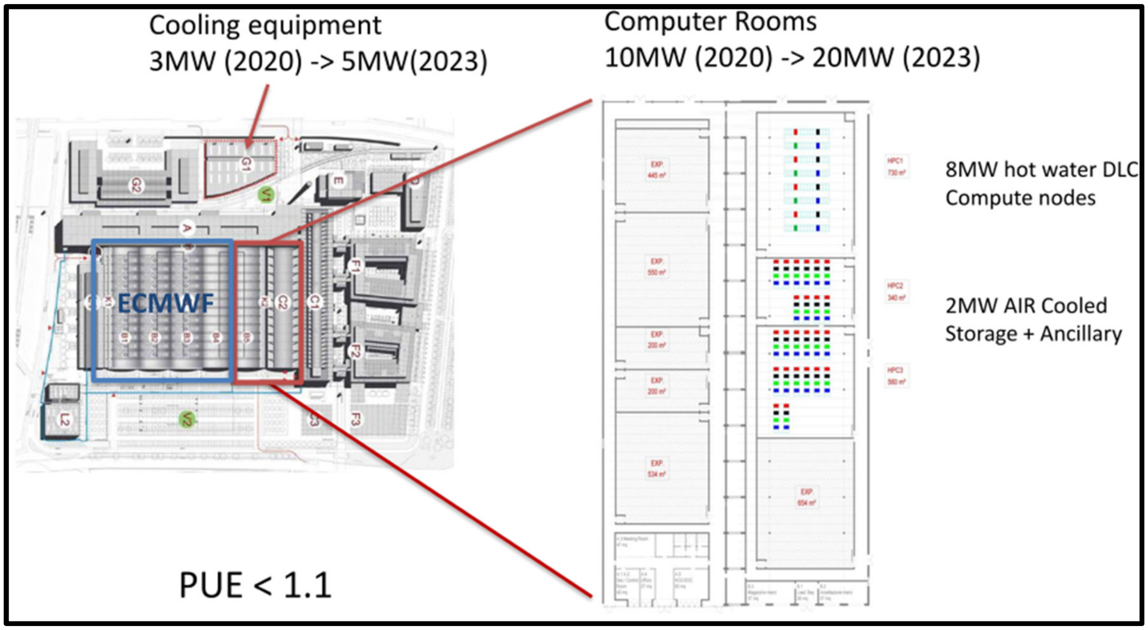

Figure 23. Future expansion of Bologna Science Park [69]

A few of the buildings at the Bologna science park have photovoltaic cells on their roofs and are producing 500 MWh of electricity per year [49]. Although the centre is not entirely based on renewable energy, it is partially powered by solar energy. But we do not know how the rest of the power (either from renewable energy or fossil fuel) is supplied to the science park. Based on available statistical data, half of the country's electricity comes from natural gas [50]. The science park occupies more than 2,500 m^2. It has already a couple of ISO certificates: ISO/IEC 9001:2015 (Quality Management) and ISO/IEC 27001:2013 (Information Security Management) [51], ensuring that the centre follows all the necessary international protocols for hosting a big data centre.

MeluXina, LuxProvide, Luxembourg

MeluXina is one of the EuroHPC petascale systems. Hosted in Luxembourg within LuxConnect's data centre DC2 in Bissen, the facility is cooled by green energy, more specifically by surplus heat from Kiowatt, a cogeneration plant fuelled by waste wood and biomass. In particular, 65% of the heat generated by the cogeneration plant is used to cool the data centre located on the other side of the road. Two absorption cold production machines with a refrigerating capacity of 2,500 kW each are powered by superheated water from the cogeneration plant. The cold water is then delivered via a district heating network between the Kiowatt heating plant and the data centre energy facility.

Frontier, ORNL, USA

The ORNL will host the Frontier exascale machine, expected to be operational from 2021 [52]. The ORNL has a long history of hosting supercomputers. Furthermore, it is trying its best to reduce the PUE over the years. Before the Summit system, ORNL employed traditional air cooling with a chilled water mechanism to cool down the computers. Nevertheless, it has been changed since Summit's installation - ORNL is now using the warm-water cooling system. This is more efficient than the traditional air cooling methodology. To achieve a PUE below 1.1, ORNL has implemented new technologies at the data centre, which are as follows [53], [54]:

- Epoxy-coated pumps to increase pumping efficiency
- 480 V racks to prevent excessive electrical transmission losses
- Variable frequency drives (VFDs) on motors, which adjust cooling capacity to load
- Variable refrigerant flow system that filters and dehumidifies the room air while removing any residual heat released into the room from IT equipment
- LED lighting and timer system

Due to those implementations, the centre was awarded the "Tennessee Chamber of Commerce and Industry's 2018 Energy Excellence Award" for the data centre's energy-efficient solution [54]. The above technologies also reduced the annual cost by $430,000. Finally, ORNL plans a more efficient cooling system for Frontier, reducing the operational cost and it is expected to be 30 to 40 percent cooling efficient compared with Summit [55].

Pawsey Supercomputing Centre, Perth, Australia

The Pawsey data centre has been operational since 2009 in Perth, Australia [56]. It is planning to host the new petascale machine, which can deliver up to 50 PFlop/s [57]. The machine vendor is HPE Cray and will use the AMD EPYC CPUs and AMD Instinct GPUs with a Cray ClusterStor E1000 storage system. This centre has quite an exciting way of supplying power and cooling to the centre discussed in the following paragraph [58].

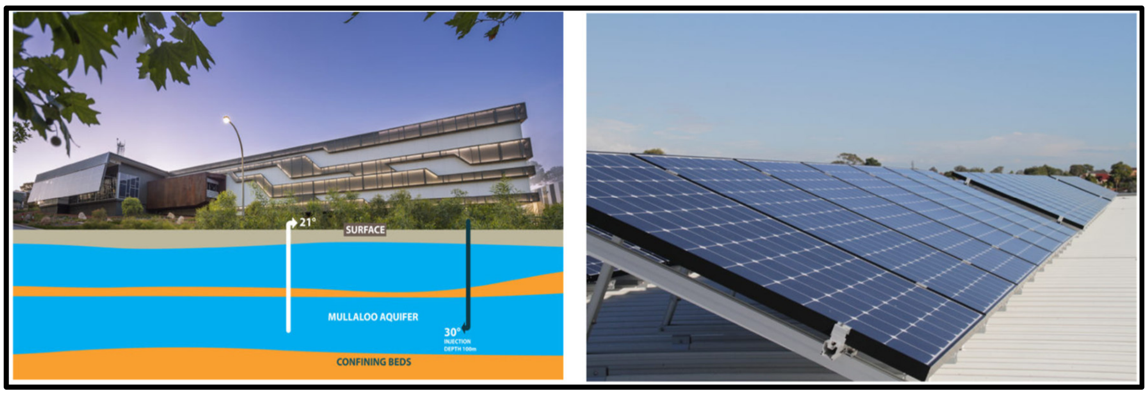

Figure 24. The Pawsey ground water cooling and solar panels [70]

Two kinds of renewable resources are being used in the Pawsey data centre: groundwater cooling and solar energy, see Figure 24. The following list shows the Pawsey eco-friendly and renewable approach:

- Proper insulation since temperature in Perth goes above 30 °C a few months a year. The insulation is efficient enough not to let external heat enter the data centre.
- A photovoltaic system is incorporated into the building's shaded facade and the building's roofs. In total, this installation produces up to 120 kW of electricity. This is, of course, carbon-free energy [59].
- It uses the geothermal cooling methodology; heated water from the data centre will be sent to an aquifer in the ground. This methodology saves up to 7 million litres of water every year compared to conventional cooling towers [60].
- The data centre has an intelligent system that will monitor the entire building for power supply, cooling and lighting. This setup reduces overall power costs.

- The entire data centre is designed to accommodate any future expansion with pre-exascale or exascale machines.

Facebook Data centre in Luleå, Sweden

One of the Facebook data centres is located in Luleå, Sweden, with 6,400 m^2. According to Facebook, it is 100% powered by renewable energy produced from the Luleå river as hydro-power. It consumes 38% less power and 60% less water than the traditional data centres. This is mainly due to the use of free air cooling available due to the geographic location. This all makes this Facebook data centre green [61], [62]. Figure 25 shows the inside of the data centre and the cold air-inlet.

Figure 25. Facebook data centre in Luleå, outdoor air inlet and computer cabins [71], [72]

Google Data centres

Google has data centres operating around the world, within the four continents, North America, South America, Europe and Asia.

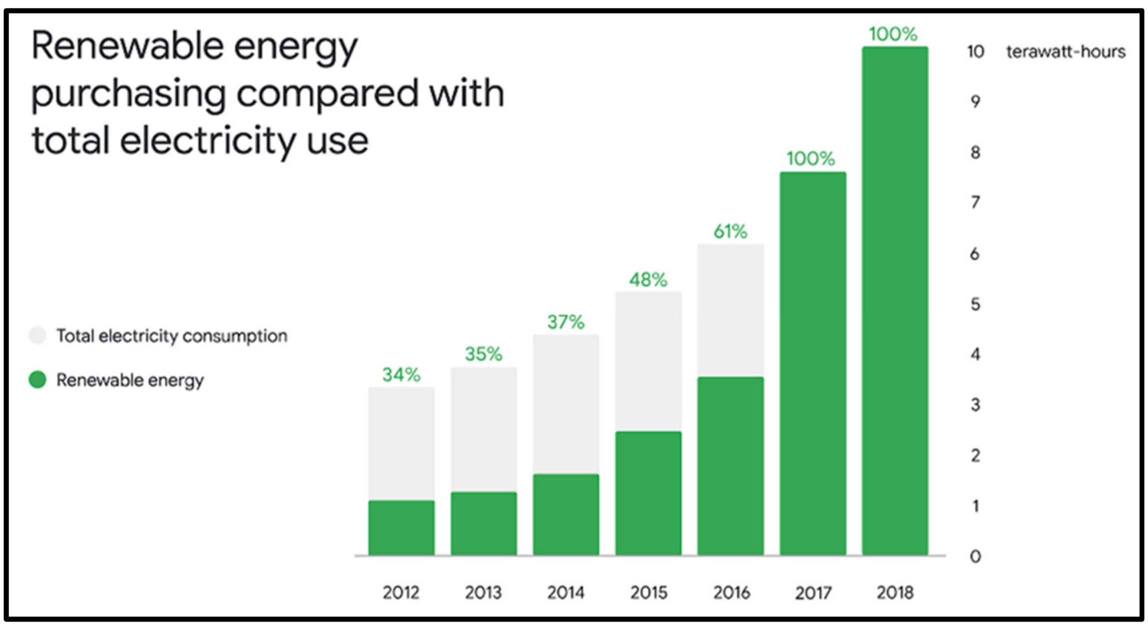

Figure 26. Google data centre power purchase [63]

Among those, Google data centres have been investing in innovative technology to make their data centres environmentally friendly. The following highlights a few of the ongoing technological innovations and green strategies:

- In Belgium, the Google data centre is going to install a battery-based backup power system data centre to avoid power shortages; this will eliminate the on-site fuel-based power production, often in the form of diesel-generators [64].
- Since 2007, Google has been carbon neutral, and during 2017, 2018, and 2019 Google only has been using the only renewable energy for all their data centres [66].
- All of Google's data centres have been using a machine-learning algorithm to optimise the data centres' cooling since 2014 [65]. Moreover, they also use an intelligent system that controls the data centres' temperature and lighting to reduce energy consumption. This strategy has achieved a PUE of around 1.10, which is less than the industrial average of PUE 1.67 [65].
- In 2019, Google made another electricity purchase (the data centre expansion led to the need of additional purchase), which is about 1,600 MW; with this, they will have in total 5,500 MW of electricity from renewable energy [66]. Figure 26 shows the yearly overview of Google's electricity purchase.

5.4 Recommendations for efficient data centres

The following section discusses a few criteria that need to be considered for hosting efficient data centres.

Power Supply:

Most new data centres and old data centres adopt power supplied from renewable energy - for example, LUMI, Google, and Facebook sites. Some renewables are more prone to interruptions like solar and wind. During a sudden power outage, battery technology or biofuel (gas/liquid) for electricity generators can be used, and this will allow data centres to rely on power from renewable energy continuously. Using the power supply from renewable energy gives not only zero-emission but also local/international government subsidies. Most of the time, depending on the location, renewable energy types can be chosen. For example, sunny places could consider solar energy, windy sites could consider wind turbines, and areas with a continuous flow of a river (near water dams) can consider hydropower.

Cooling:

Currently, two cooling methods are quite efficient and popular among the data centres. They are:

- Immersion cooling (single-phase and two-phase)
- Direct-to-Chip Liquid cooling

Immersion cooling:

Single-phase: in this type of cooling, the entire system is submerged into the liquid (dielectric coolant), and the coolant does not change its phase since the coolant has a higher boiling point, which means it does not evaporate due to higher heat. Due to its dielectric property, the fluid does not act as a conductive medium, making it safe to immerse the electronics directly into the liquid. Once the heat is removed from the module (system), the coolant gets hot, and it will be pumped into the cooling facility, where it gets cooled. The schematic overview of this process can be seen in Figure 27. This kind of process is very efficient, especially when the module (system) has high density. The main drawback is access to the boards when maintenance is required. Some problems with liquid that dissolves plastics in the boards and cables have also occurred.

Two-phase: here, the process is very similar to single-phase immersion cooling, but the coolant remains in the enclosed box of a container along with the system, and the coolant evaporates once it is heated. Later, this evaporated vapour is condensed by the cooling coils which are inside of the enclosed container. Figure 28 shows the design model of the two-phase immersion cooling technology. Drawbacks of this kind of system can be that liquid can be lost if the enclosure is not completely air-tight. This requirement also makes maintenance more complex. Residues on the hot surfaces have also occurred.

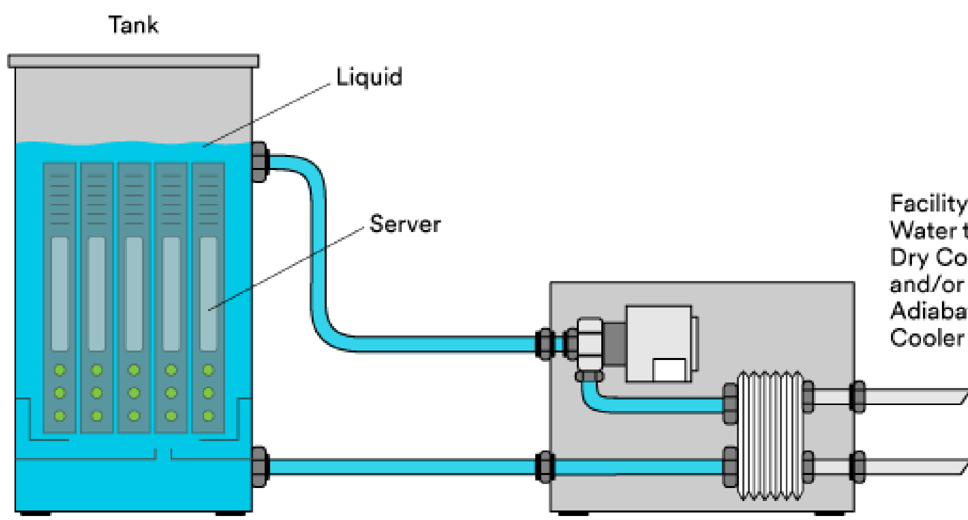

Figure 27. Single-phase immersion cooling [73]

Direct-to-Chip Liquid cooling:

Here, pipes lead the coolant to each component that needs cooling. The coolant does not have direct contact with the system hardware; instead, they contact through the surface of a specially designed cooler or cooled plate. This technology is getting quite popular among computing hardware manufacturers. This process makes less noise for data centres and can deliver a cooling capacity of up to 100 kW per rack for a standard-sized rack.

Benefits of using liquid cooling (immersion and direct-to-chip):

- Due to its higher efficiency of the cooling, the higher density system can make the most benefit from it. Whereas, if we use air-cooling technology for the high-density system, the cooling will not be uniform and efficient. Currently and in the near future, the system will have a higher density. To be able to cool such high-density modules (systems), liquid cooling is an optimal solution.

- Liquid cooling is more efficient compared to other cooling technologies, which were mentioned earlier in this document. The cost is reduced by requiring lower energy overhead for the cooling, higher density racks (less space) and other accessories requirements. Overall the power for cooling can be reduced from 20-30 percent compared to other cooling technologies. However, the investment costs are usually higher. This makes liquid cooling most attractive for high-density and high-power computers.
- Due to less noise, these types of data centres are easier to be placed anywhere. However, this is most of the concern for smaller compute centres (like edge computing).

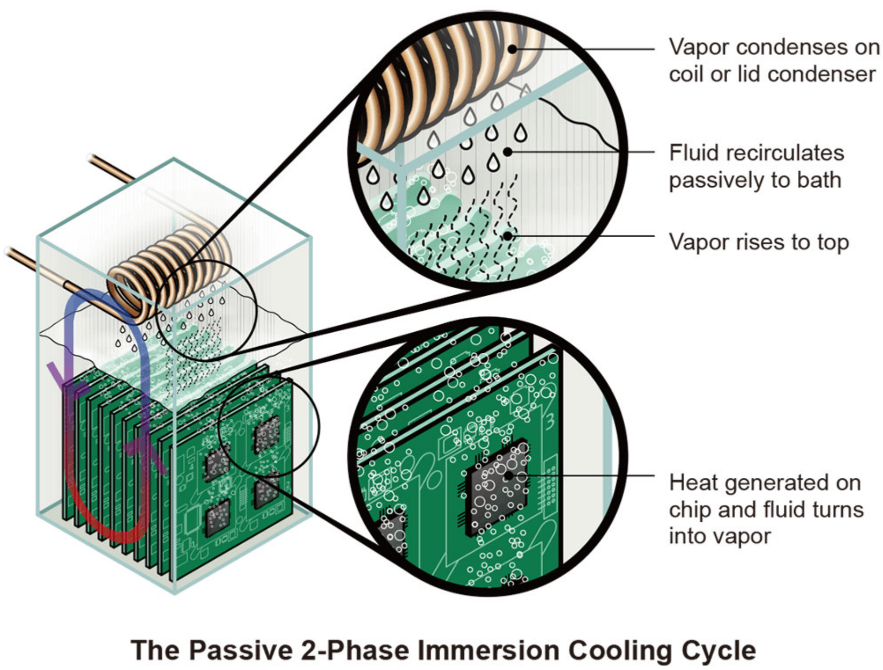

Figure 28. Two phase immersion cooling [74]

Floor Planning:

Traditionally the data centres floor and Heating, Ventilation, and Air Conditioning (HVAC) should be well planned to accommodate extensive systems. But, currently and in the future, the entire data centre will be cooled by liquid cooling. This minimises the effort for the HVAC design in detail for the traditional cooling concepts in the data centres. For example, the HPE Cray EX supercomputer offers closed-loop liquid cooling [75]. This will prevent the heat from dissipating to the surrounding. In this case, the entire system can be placed in a minimum HVAC facility place, reducing the effort to plan the air circulation in the computer hall. However, instead planning of the cooling pipes is required.

6 Green metrics to monitor

As mentioned earlier, the energy consumption of HPC systems continues to grow hand in hand with their performance improvements, making the efforts towards power reduction of modern HPC data centres imperative. This section provides a high-level overview of some of the available metrics used for monitoring the energy efficiency in modern data centres and outlines some of their advantages and disadvantages[11].

- **Power Usage Effectiveness (PUE)[12].** PUE is a widespread metric. It is defined as the ratio of total facility energy to the deployed IT equipment energy [11], i.e.

$$PUE = \frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy}$$

  However, it implies specific issues, ranging **from measurement** and reporting aspects:
  - Data centre owners and operators should adhere to certain guidelines in measurements and reporting (typically provided by The Green Grid (TGG) [76])

  to interpretation:
  - PUE typically worsens with decreasing IT power consumption, i.e. if your data centre has a newer generation of more energy-efficient servers, then your power consumption for the same workload you used to run will obviously be less, which in its turn will imply a higher PUE.

  The other issue with PUE is that it is related to a whole centre that may host systems of different types/purpose. However, since the next-generation exascale machines are likely to be the biggest in term of energy use, the other systems may possibly be neglected. Nevertheless, partial PUE (pPUE) can be used further for calculating the effectiveness of power usage for a specific segment of a data centre or a mixed-use facility. Please refer to [11] for more information on PUE and pPUE.

- **Sustainability metrics.** The environmental impact should be constantly monitored and reported during the construction of the system and the supporting ecosystem as well as during the system's lifetime. The list below outlines three sustainability-related metrics proposed by TGG [78].
  - **Carbon Usage Effectiveness (CUE).** CUE is a metric used for assessing the impact of operational carbon usage of data centres. It is defined as a ratio of total carbon emissions caused by the total data centre energy to the deployed IT equipment energy. Please refer to [78] for more information regarding the metric.
  - **Water Usage Effectiveness (WUE).** This metric allows accessing the water usage associated with the data centre operations. It is defined as the quota of the annual water usage (water used for cooling, humidity regulation, etc.) over the energy usage of the deployed IT equipment. Please refer to [79] for more information regarding the metric.
  - **Space Usage Effectiveness (SpUE).** SpUE assists in optimising the facility space usage while considering energy efficiency. SpUE uses the *(i)* power consumption of the deployed IT equipment, *(ii)* the number of racks inside a data centre; and *(iii)* the data centre space as its calculation basis, and in contrast to earlier mentioned three metrics (i.e. PUE, CUE, and WUE), SpUE is not represented by solely one number and does not have theoretical ideal values. Please refer to [80] for more information regarding the metric.

[11] This subsection does not intend to provide an overview on the complete list of the available green metrics, but to outline the most frequently used ones.
[12] Data Centre infrastructure Efficiency (DCiE) [77] is another metric defined as the reciprocal value of PUE and is expressed as a percentage.

- **Metrics accounting for energy reuse.** As mentioned in Chapter 4, it is equally important to consider the possibilities of energy reuse for the reduction of the overall operating costs and environmental impacts. Below are two additional metrics suggested by TGG.
  - **Energy Reuse Effectiveness (ERE)**. ERE [37] is a metric to account for the amount of energy reuse by data centres. Similarly to PUE, it is defined as a fraction of total facility energy over the deployed IT equipment energy with a difference that the numerator of that fraction also accounts for the amount of reused energy, i.e.

$$ERE = \frac{Total\ Facility\ Energy\ - Reuse\ Energy}{IT\ Equipment\ Energy}$$

    It is worth noting that while PUE ranges from 1.0 to infinity, the possible ERE range is 0 to infinity, with ERE of 0 indicating a 100% reuse of used energy.

  - **Energy Reuse Factor (ERF)**. ERF is yet another metric allowing to account for energy reuse, defined as:

$$ERF = \frac{Reuse\ Energy}{Total\ Energy}$$

    In contrast to ERE, ERF ranges from 0 to 1.0, with 0 indicating the absence of any energy reuse and 1 showing a 100% energy reuse. From the definitions presented above, it is easy to derive the interrelation between ERE, ERF, and PUE: $ERE = (1 - ERF) \times PUE$.

  Please refer to [37] for more information regarding ERE and ERF metrics.

- **Metrics accounting for cooling**
  - **Coefficient of Performance (COP) of the cooling infrastructure.** It is known that the energy consumption of cooling infrastructure accounts for a substantial fraction with respect to the total data centre energy consumption [29], [31]. In case a data centre uses multiple cooling technologies, the PUE will combine the cooling efficiency (COP) from all cooling technologies into one number. This would make it impossible to calculate the cooling overheads for individual IT systems. The COP of the cooling infrastructure is an easy to calculate metric and is (typically) defined by a fraction of generated cooling supply (i.e. how much cooling is needed) and the power consumed for generating that supply (i.e. how much power is consumed for getting the required amount of cooling) [29]. It is worth noting that this metric can be used for chiller-based as well as for chiller-less cooling solutions. For the latter, it can help with assessing the number of required active cooling towers, flow rate(s)/speed of the pump(s) given the outside temperature and the current system load, etc.
  - **Delta-T**. Metric indicating the difference between the inlet and outlet temperature of the IT equipment (e.g. per cabinet). Useful for assessment of cooling efficiency and for subsequent minimisation of the associated costs.

- **Metrics accounting for workload**
  - **Flop/s per W (also FLOPS/W)**. Useful if the focus is more on the HPC system side. Some synthetic data regarding this metric is available via the Green500 list

[81]. The drawback of this metric is that it does not indicate the efficiency of a particular machine when deployed in a specific data centre.

- **Energy-to-Solution (EtS)**. This metric indicates the aggregated energy consumption of a given application [82], [83]. This value indicates the aggregated energy consumption of compute nodes used to run a given workload, including also partial sub-system components (e.g. system networking, system cooling, and infrastructure), which were utilized during the run.
- **Data centre Workload Power Efficiency (DWPE)** is a metric aimed at showing the energy-efficiency of a given workload when run on a particular HPC system deployed at a specific data centre [84].

In addition, further data centre-level KPIs regarding the availability, reliability, and redundancy of the infrastructure equipment should also be considered. These include:

- **Infrastructure power capacity forecast**. Analysing the power consumption trends with corresponding alerting mechanisms in case of a potential violation of the thresholds
- **Infrastructure equipment availability**. Availability percentage of critical infrastructure equipment that ensures the reliable and uninterrupted (critical) operation of the target HPC system
- **Infrastructure equipment failure.** Number of infrastructure-related components failures and corresponding requests for change/repairing work (per month)
- **Infrastructure incidents affecting HPC system availability**. Number of infrastructure incidents affecting the availability of the system (per month)
- **Infrastructure incidents affecting HPC system equipment**. Number of infrastructure incidents affecting the equipment of the underlying HPC system resulting in a mandatory change/repairing of certain system components (per month)

7 Conclusion

Currently, the HPC community is actively working towards the planning and the deployment of next-generation supercomputers capable of delivering exascale (or close to exascale) computational performance. This involves not only the design and development of the target HPC system but also the aspects of hosting infrastructure, i.e. facility, cooling, power delivery, etc. The latter can be equally challenging, as the planning and commissioning of a supporting ecosystem are time-consuming and costly and should support the hosting of diverse high-end IT systems over several generations.

This document, built upon the work conducted in PRACE-IP projects in terms of best practices for energy-efficient design and operations of large-scale HPC data centres, provided a review of current exascale projects and initiatives in Europe as well as overseas, followed by an analysis of survey results circulated among European High-Performance Computing Joint Undertaking (EuroHPC JU) pre-exascale and petascale sites as well as among additional PRACE Tier-0 and Tier-1 sites. The main aim of this survey was the identification of installation requirements across European HPC sites for hosting next generation HPC systems. Most importantly, this survey revealed that:

- *power capacity extension*, followed by *building infrastructure extension/modification* and *power costs* are the three top challenges that the European data centres will face when hosting next generation HPC systems

- 40% of the surveyed data centres currently power their flagship systems using renewable energy sources, while 30% only partially rely on renewables for powering their high-end IT systems
- *DCIM* and *IT system SW management tools*, together with the *frameworks for operational data analytics and predictive maintenance* are the foreseen major investment areas by the European HPC data centres

Based on the mentioned information presented in Chapter 2 and Chapter 3, this document then derived the main installation requirements and discussed the best practices taken (or planned to be taken) into account to host next-generation supercomputing systems.

The following bullet list outlines the most notable of these requirements and suggestions for the supporting ecosystem of a data centre when hosting next generation exascale systems.

- The supporting building infrastructure should ensure a power delivery for hosting a system in the range of 20 to 25 MW maximum consumption.
- The supporting building infrastructure should be capable of handling large power fluctuations without compromising the availability and accessibility of the target IT systems.
- The supporting building infrastructure should be capable of accommodating state-of-the-art cooling solutions.
- The raised floor or the concrete slab (in case there is no raised floor) should be able to bear at least 2,200 kg/m2 distributed load.
- Data centres should be capable of taking advantage of the available renewable energy sources and ensure the reuse of generated waste heat. Data centre operators should ensure the usage and continuous monitoring of various energy-efficiency and sustainability-related Key Performance Indicators (KPIs) during the planning and commissioning. In particular, the following should be monitored and ensured:
    - Power Usage Effectiveness (PUE) should be less or equal to 1.1.
    - Energy Reuse Effectiveness (ERE) should be larger than 0.5.