Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

HPC Infrastructures Workshop:
2020 Online Event

Hayk Shoukourian [a*], Volker Weinberg [a], Radosław Januszewski [b], Huub Stoffers [c], Andreas Johansson [d], Susanna Salminen [e], Ezhilmathi Krishnasamy [f], Norbert Meyer [b], Dirk Pleiter [g], François Robin [h], Frederic Souques [h], Jean-Marc Ducos [h], Jean-Philippe Nomine [h]

[a] Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities (BADW-LRZ), Germany
[b] Poznan Supercomputing and Networking Center (PSNC), Poland
[c] SURF, Netherlands
[d] National Supercomputer Centre (NSC) at Linköping University, Sweden,
[e] CSC IT Center for Science, Finland
[f] University of Luxembourg, Luxembourg
[g] Jülich Supercomputing Centre (JSC), Germany and PDC Center for High Performance Computing, Sweden,
[h] Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), France

Abstract

The annually held series of European Workshops on HPC Infrastructures (EWHPC) aim to bring together worldwide specialists in HPC centre design and operation to discuss the latest trends in infrastructure and supporting technologies for supercomputing centres.

Due to the COVID-19 pandemic, the 11th EWHPC was cancelled and a shorter online event was offered instead. It attracted more than 100 participants from 40 different HPC sites and provided a general overview on EuroHPC Joint Undertaking (JU) as well as updates on three EuroHPC JU pre-exascale consortia and their systems: Leonardo (hosted at CINECA), LUMI (hosted at CSC), and MareNostrum 5 (hosted at BSC).

The EWHPC online event was held on 14th of October and virtually collocated with an online PRACE Topical Session on "Exascale for European Datacentres" held on 15 October 2020, then followed by the traditional PRACE session on HPC infrastructures with technical presentations from PRACE sites on 16 October 2020.

This document summarises the presentations and discussions held at the online event and the PRACE session on HPC infrastructures.

[*] Corresponding author. Email address: shoukourian@lrz.de

DOI: 10.5281/zenodo.5838044

31/05/2021

Table of Contents

Abstract 1

Table of Contents 2

1. Introduction 3

2. Online Session: EuroHPC Pre-Exascale, chaired by Herbert Huber 4

2.1 EuroHPC state-of-play-HPC 4

2.2 EuroHPC Leonardo Consortium 5

2.3 EuroHPC LUMI Consortium 7

2.4 EuroHPC MareNostrum 5 8

3 PRACE Session, chaired by François Robin 17

3.1 HLRS Update 17

3.2 HPC infrastructure at Luxembourg University 19

3.3 Distributed national HPC/cloud and data infrastructure 21

3.4 Update on the PPI4HPC project 26

3.5 Big Data & Machine Learning at TGCC - First approach with KAPSDATA 28

4. 11th European Workshop on HPC Infrastructures 29

5. Conclusions 30

Acknowledgements 32

List of Acronyms and Abbreviations 33

DOI: 10.5281/zenodo.5838044

1. Introduction

The 11th European Workshops on HPC Infrastructures (EWHPC), initially planned to be held at the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities on 25-27 May 2020, was cancelled due to the COVID-19 pandemic. Instead, a shorter online version of the event was offered on 14th of October.

The program committee for this online event consisted of following site representatives:

- François Robin and Jean-Philippe Nominé (CEA)
- Ladina Gilly and Tiziano Belotti (CSCS)
- Herbert Huber and Michael Ott (LRZ)
- Javier Bartolomé (BSC)
- Norbert Meyer (PSNC)
- Gert Svensson (KTH)

In addition, Evangelos Floros (EC) and Oscar Diez (EC) were invited to attend the meetings of the program committee.

Similar to previous events of the EWHPC series, the 2020 online event was by invitation only. It brought together more than 100 participants from 40 different supercomputing sites and provided a unique platform for sharing the latest infrastructure trends, issues, and lessons-learned among HPC sites facing similar challenges and requirements.

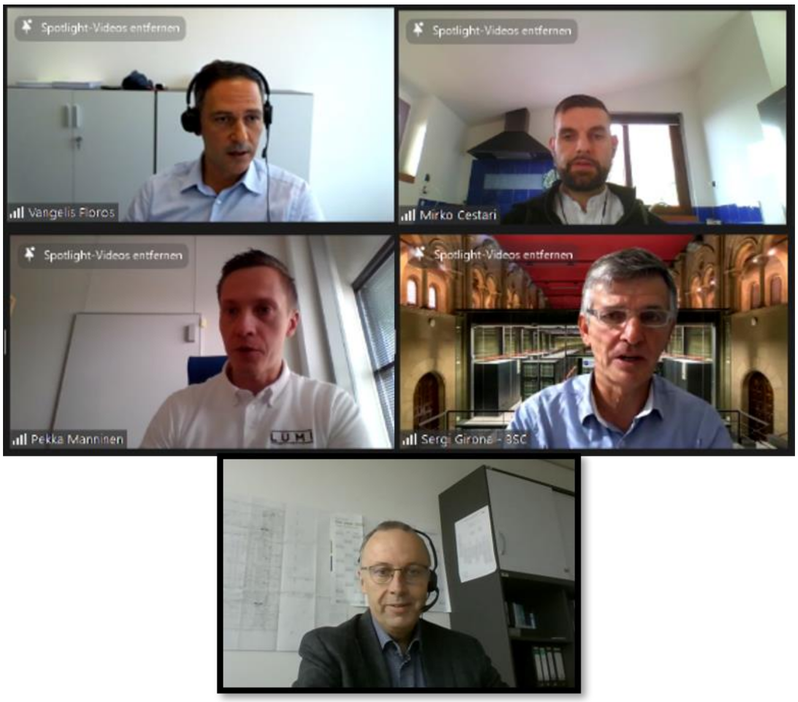

Spotlight-Videos entfernen
Vangelis Floros
Spotlight-Videos entfernen
Mirko Cestari
Spotlight-Videos entfernen
Pekka Manninen
Spotlight-Videos entfernen
Sergi Girona - BSC

Figure 1. Speakers of EWHPC 2020 Online Event (from top left to bottom): Evangelos Floros (EC); Mirko Cestari (CINECA), Pekka Manninen (CSC), Sergi Girona (BSC), and session chair Herbert Huber (LRZ)

The agenda of EWHPC 2020 online event contained five talks, starting with an overview presentation on the EuroHPC Joint Undertaking (JU) followed by updates[1] on the three EuroHPC JU pre-exascale systems, namely:

[1] As of known by 14th of October 2020.

Leonardo hosted by CINECA; LUMI hosted by CSC; and MareNostrum 5 hosted by BSC. The talk concerning next year's 11th EWHPC concluded the online session.

Due to the COVID-19 pandemic, the 11th EWHPC will be an online event taking place from May 18th to 19th, 2021 followed by the PRACE session on 20th of May, 2021. For up-to-date information, please refer to the EWHPC official website[2].

The virtual EWHPC 2020 event featured a PRACE Topical Session on "Exascale for European Datacentres", held on 15 October 2020, followed by the traditional PRACE session on HPC infrastructures with technical presentations from PRACE sites on 16 October 2020. The PRACE session on HPC infrastructures gathered attendees from PRACE Tier-0 and Tier-1 sites and contained five talks given by PRACE partners from HLRS, University of Luxembourg, PSNC, JSC, and CEA. This session gave further opportunity for exchanges of best practices among experts involved.

This document provides a summary of the facts, insights, and best practices discussed during the EWHPC 2020 online event and the following PRACE session on HPC infrastructures for a wider dissemination among interested parties and experts involved in HPC facility management. The proceedings of the previous series of EWHPC can be found via this link.[3]

2. Online Session: EuroHPC Pre-Exascale, chaired by Herbert Huber

2.1 EuroHPC state-of-play-HPC

Presenter: *Evangelos Floros, EC*

EuroHPC is a legal and funding agency with 32 participating states and two private members: ETP4HPC and BDVA. Having a budget of 1.5 billion Euro from the EU and participating states its mission is to establish a world-class supercomputing and data infrastructure. This mission is implemented via three main activities/pillars: HPC ecosystem; Infrastructure and Operations; and Research and Innovation (R&I).

The planned infrastructure will consist of three large (called pre-exascale) and five smaller (called petascale) supercomputers. The three pre-exascale supercomputers will be hosted by CSC in Kajaani, Finland (supercomputer LUMI which will have a theoretical peak performance of 550 PFLOPS), CINECA in Bologna, Italy (supercomputer Leonardo which will have a theoretical peak performance of 248 PFLOPS) and BSC in Barcelona, Spain (supercomputer MareNostrum 5 which will have a theoretical peak performance of 200 PFLOPS).

The smaller petascale systems will be hosted by:

1. LuxProvide in Bissen, Luxemburg: MeluXina supercomputer
2. IZUM, Maribod, Slovenia: Vega supercomputer
3. IT4I, Ostrava, Czechia: Euro-IT4I
4. Univ. Minho, Riba de Ave, Portugal, Deucalion supercomputer
5. Sofia Tech Park, Sofia Bulgary, PetaSC

Pre-exascale procurement of the systems was managed by EuroHPC as different lots of the same procurement. In general, the procurements of petascale systems were managed by each respective hosting country. The Bulgarian and Portuguese supercomputers were the exception and managed by EuroHPC as a separate procurement.

For the LuxProvide and IZUM systems tender procedures are closed and the contracts are already signed. The machines will be operational by the April 2021 and March 2021 respectively[4].

For 30.4 million Euro, MeluXina will have a maximum performance of 12.26 PFLOPS provided by x86 CPU and GPU partitions. With a budget of 17 million Euro, Vega will offer a maximum performance of 6.8 PFLOPS with x86 and GPU partitions. Both systems will be delivered by Atos.

[2] https://www.euhpcinfrastructureworkshop.org/

[3] https://prace-ri.eu/infrastructure-support/european-workshops-on-hpc-infrastructures/

[4] Based on the information provided during the EWHPC 2020 online event.

For the other three systems the tender process is not yet finished, the most advanced is IT4I with a signed but not officially announced contract. For 14.8 million Euro the system will offer 9.1 PFLOPS of aggregated maximum performance for x86 and GPU partitions and should be operational in May 2021.

The last two machines completed the evaluation phase with award notification pending. For budgets of 20.1 million Euro (Deucalion) and 11.6 million Euro (PetaSC) it is expected that these machines will correspondingly deliver 7.2 and 4.4 PFLOPS computing power. While PetaSC will be purely traditional x86 system, the Deucalion will consist of x86 and ARM partitions.

The following bullet list provides a summary of the procurement status and the expected delivery timelines for the three EuroHPC pre-exascale systems:

- **LUMI**: the procurement process is finalised and award decision has been submitted to vendors. The delivery should start Q2 2021.
- **Leonardo**: the procurement process is finalised and award decision has been submitted to vendors. The delivery should start Q3 2021.
- **MareNostrum 5**: the tender was still under evaluation and start of the deployment is planned for Q4 2021.

Detailed information about pre-exascale system configuration and tender procedures will be covered in in the following chapters.

Another pillar of the EuroHPC project is the R&I pillar. The goal of this pillar is to “elevate the participating countries to a common high level in the fields of HPC, HPDA and artificial intelligence (AI)” which is being implemented by the H2020-JTI-EuroHPC-2019-2 call - Innovating and Widening the HPC use and skills base. This call has a budget of 54 million Euro and aims at creating European HPC competence centres that will help to strengthen the skills base required for an efficient use of modern HPC and AI supercomputers for solving critical problems. Currently two projects are launched within this call: CASTIEL with the goal to combine the National Competence Centres (NCC) formed in EuroCC into a pan-European network and FF4EuroHPC established to promote innovation using high-performance computing in small and medium-sized enterprises across Europe.

The pillar “HPC ecosystem” aims at supporting the creation of European HPC technologies and applications. It is implemented by the H2020-JTI-EuroHPC-2019-1 call with a 55 million Euro budget.

This call has three main topics:

- Support the European technology supply industry in developing next generation power-efficient and highly resilient HPC and data technologies
- Help maintain Europe’s world leadership in HPC applications by stimulating the innovation potential of businesses and industry users to develop applications in different industry sectors (such as manufacturing, farming, health, mobility, natural hazards, energy, climate, space, finance and cybersecurity).
- Help European software vendors to improve their offer of industrial software and codes for industrial users to make full use of new, very high-performing supercomputers.

The activities that are part of the pillar are not confined to the call mentioned before. In 2020 several other calls were established. The calls covered topics such as advanced pilots of European supercomputer, quantum computer simulator and development of European low-power CPU.

2.2 EuroHPC Leonardo Consortium

Presenter: *Mirko Cestari, CINECA*

CINECA (Bologna, Italy), is one of the three hosting entities selected by the European High-Performance Computing Joint Undertaking (EuroHPC JU) for hosting and operating a pre-exascale system. A pre-exascale system is a system in the larger of the two classes of systems that are funded by the EuroHPC JU:

- Pre-exascale systems (3 systems)
- Petascale systems (5 systems)

A pre-exascale systems must have a nominal compute capacity of at least 150 PFLOPS. While the five petascale systems are to be owned by the hosting entity and subsidised by the EuroHPC JU, the EuroHPC JU is and will remain the formal owner of the pre-exascale systems. The Euro HPC JU covers 50% of the acquisition budget as well as 50% of the budget for operations.

The pre-exascale system to be hosted at CINECA is expected to start operations in Q3 of 2021, and will be named 'Leonardo'. At the time of the presentation, the procurement process was still ongoing therefore not many details about the system can be disclosed[5]. However, some marked and interesting aspects of the *vision* for the system can be discussed by focusing on the procurement and the processes leading up to it. The procurement process for the system formally started in January 2020, with the release of the technical specifications for entering the competitive dialog and ran until mid-October 2020.

The procurement process was a joint effort of the parties involved in the Leonardo Consortium: The EuroHPC JU, CINECA, and the Italian National Institute for Nuclear Physics (INFN). Along with Italian Ministry of Universities and Research, the International School of Advanced Studies (SISSA), this consortium includes participants from the following member states: Slovenia, Slovakia, Austria, Hungary, and Greece[6]. EuroHPC JU conducted the procurement process and covered many of the formal aspects of this process. CINECA, with ten staff members involved, covered the technical aspects of the procurement and focused more on the architecture of the machine. INFN, with two staff members involved, provided consultancy and support pertaining more to the integration of the system in the designated hosting facility.

Hosting site candidacy

CINECA embarked on a pre-exascale hosting site candidacy in 2019. In the period of the past twelve years, CINECA has been able to renew their Tier-0 petascale systems about every three to four years. CINECA began their candidacy as a hosting site for EuroHPC with a roadmap towards the goal of having a national full exascale system by 2025. CINECA's vision for a 2021 pre-exascale system was that of a heterogeneous single phase system, operational for 3 to 4 years, consisting of three distinct modules:

- A 'booster' subsystem in which the main HPC power is to be concentrated, of about 3500 compute nodes
- A general purpose subsystem, to accommodate many different legacy workflows of about 500 nodes
- A subsystem for data-centric data intensive workflows of about 1000 nodes

The system's total storage capacity should be at least 150 PB, with an aggregate bandwidth of 1 TB/s. Targeted HPL performance was 150-180 PFLOPS, with a nominal peak performance of 210-250 PFLOPS. The targeted bandwidth for the central interconnect was at least 200 GB/s. The power envelope for the system was 8 – 9 MW, with a PUE of at most 1.1.

The system must be able to cater to the needs of quite a diverse set of fields, such as High Energy Physics (HEP), Oil&Gas, and Pharma. It must also be able to contribute to European leadership and gaining sovereignty on strategic technologies for European economic wealth, such as artificial intelligence, cybersecurity, and internet of things.

Besides the general outline of a system, CINECA also had a clear idea on where to host a system with such demanding infrastructure needs, viz. at the new Bologna Science Park location. The location has two adequately sized rooms available either of which can be used to host the system. The cooling capacity is currently about 10 MW: 8 MW of direct liquid cooling (DLC) can be provided for dense HPC nodes. 2 MW of air-cooling capacity for storage and other auxiliary racks. It is feasible to expand the cooling capacity to 20 MW in 2023.

Procurement strategy

The acquisition budget for the system was 120 million Euro for the system, software stack and licenses, and support and maintenance for a period of 5 years. The Leonardo Consortium supported CINECA's roadmap and allowed further development towards heterogeneous compute systems.

A main goal of the competitive dialog procurement procedure and the technical specifications released in the request for proposals (RfP) was to promote competition among vendors. The consortium perceived that the way to achieve this was by relaxing requirements pertaining to the node design. Most of the mandatory requirements focused on aspects of facility integration. The few mandatory requirements for node design were mainly for the data-centric and general-purpose nodes, and specified a minimal number of cores, NVMe capacity, and RAM capacity. Vendors were allowed to propose the same design for data-centric and general-purpose nodes. There is a risk in reducing mandatory requirements, but the ability to discuss design in a competitive dialog with the

[5] This meeting was held one day before the disclosure of the party that won the tender and consequently nothing specific about the new system itself could be disclosed. At the time of writing this report, we know that the contract was awarded to Atos.

[6] https://www.cineca.it/en/hot-topics/Leonardo-announce

technology providers was perceived, and indeed found to work, as a sufficient guarantee to avoid relatively esoteric architectures.

There were three dialogue rounds of 4 hours with each candidate, in mid-February, beginning of April and mid of May. The benchmark specification was released during the dialog phase. The final RFP was provided on June 4.

Evaluation of offers

The offers received were evaluated on three sets of criteria:

- System performance, based on HPL and HPCG (40%)
- Total cost of ownership (TCO) analysis, based on the benchmark suite (30%)
- Benchmark analysis assessment, taking into account the completeness of results and the quality of projections and methodology (30%).

The TCO analysis aims to define "value for money". The "value" is based on an estimate of the average number of workloads that can be run on the system during its entire lifetime. The estimate is based on the vendor's benchmark results. The "money" is an approximation of the TCO. The TCO approximation made use of the common framework for comprehensive cost analysis that was developed by BSC, CINECA, CEA, and JSC while they took part in the PPI4HPC project[7].

The benchmark, on which the "value" estimate is based, took about four months to prepare using 4 applications. The advantage of using only 4 applications, is that the applications are all very well tested on different architectures, and make a sound and reliable comparison of a system's feasibility. In addition, a low number of applications are easier to manage for both the procurer and the candidates. The candidates received iteratively refined versions of the benchmark. Receiving a first provisional version of the benchmark very early allowed the vendors to immediately start on the projection methodology. A changelog to track repository updates and a benchmark information document with all rules pertaining to compilation and runs was provided from the start.

The downside – especially in the context of the ambition that this machine is to cater to much more that the classical MPI-based HPC workloads - is that the four applications may not be representative enough for the very heterogeneous workloads that the system is supposed to accommodate during its lifetime.

2.3 EuroHPC LUMI Consortium

Presenter: *Pekka Manninen, CSC*

The Large Unified Modern Infrastructure (LUMI) consortium is hosting one of the EuroHPC Pre-exascale systems. Ten countries – Finland, Belgium, Czech Republic, Denmark, Estonia, Iceland, Norway, Poland, Sweden, and Switzerland – are participating in the consortium. In Finnish "lumi" is also a word for snow.

LUMI will be hosted at the CSC datacentre in Kajaani, Finland. Before the site started to be converted into a data centre in 2012 the facility was used as a paper mill and parts of the site are already in use for hosting other systems. Current systems have been hosted in a converted warehouse building, but now a 15 000 m^2 building that was previously used for a paper machine will be remodelled. Due to the requirements of the previous paper mill, there is ample power capacity in the local power grid with 200 MW of hydro power available fed from the national power grid through three independent transmission lines. In the previous 38 years there has only been two minutes of power outages. The industrial facility has also fulfilled the demands of ISO 27001 for a number of years.

Free cooling can be used all year round in the northern climate and using warm water direct liquid cooling a PUE of 1.03 can be achieved. However, instead of using roof chillers the waste heat will be reused by selling the heated cooling water to the Kajaani municipal energy company. To increase the temperature from 40–45°C and make it suitable for district heating a heat pump is used. Waste heat from LUMI will provide 20% of the heat needed for the city of Kajaani. The income from this will reduce the operational cost of the system and put the effective energy price at € 35 per MWh. Currently the energy company uses fossil fuel, and this waste heat reuse can reduce emissions by 13 500 tons of CO_2 per year.

Building preparations for the installation will be finished during Q4 2020. System installation is planned for Q1 2021 and general system availability in Q2 2021. LUMI is planned to be operational until Q4 2026.

Network capacity through the Finnish NREN to the European GÉANT network is currently 4 x 100 Gbit/s. This can be expanded to multiple Tbit/s if needed since the Nordic backbone network runs through the data centre.

[7] https://www.ppi4hpc.eu

This presentation was made before the public announcement on 21 October and could not go into detail about the system and only presented the general architecture. LUMI is a GPU accelerated system and partitioned into multiple parts with different characteristics. All partitions share the same high-speed interconnect to allow using nodes from multiple partitions in the same job. The partitions are:

- LUMI-C CPU x86 “Tier-1”
- LUMI-G GPU “Tier-0”
- LUMI-K Container cloud services
- LUMI-O Object Storage using Ceph
- LUMI-D Data analysis with large shared memory nodes
- LUMI-F Performance optimised flash storage
- LUMI-P Lustre parallel storage

A partition based on future technologies (LUMI-Q) is planned but has been excluded from the procurement stage and phase one installation.

Moving applications to a new large system usually involves adaptations and users can start preparing for LUMI by considering which projects are suitable for running on it. Larger problems can be solved with the increased capacity, both by scaling up existing research and new grand challenges. Applications that are not already GPU enabled needs porting to take advantage of LUMI-G. This work can be started before LUMI is operational and will be useful for multiple new pre-exascale systems.

The focus of LUMI will be placed on the user experience and being best at that is more important than being at the top of the TOP500. Many software packages will be pre-installed and multiple user interfaces made available. Traditional command-line access will be complemented with approaches such as Jupyter Notebooks and Rstudio. Large reference data sets, such as genome databases and AI training data, will be installed system wide for use by jobs. LUMI will be a general-purpose system and not only focused on very large-scale Tier-0 simulation jobs that utilise the majority of the system. Use cases that need to analyse output data can take advantage of LUMI-D for post processing and do not need to move data to another system. Services for data management will also be provided.

All LUMI member countries will individually contribute one FTE for user support, with staff hosted in all countries and not only at CSC. User support is not limited to answering questions, the same group will provide training and manage the software installations. Additional resources for higher level support will be added via EuroHPC Competence Centres.

Half of the installed system will be allocated to EuroHPC with the remainder split among the consortium members for their own national allocation. A peer review process similar to the PRACE Tier-0 access will be used for allocations from the EuroHPC slice, while the national allocation policies will vary.

2.4 EuroHPC MareNostrum 5

Presenter: *Sergi Girona, BSC*

The Barcelona Supercomputing Center (BSC) is a public research centre located in Barcelona, Spain, providing supercomputing services as well as performing R&D in various disciplines. BSC is a consortium of the Spanish government (60%), Catalan Government (30%), and the Technical University of Catalonia (10%). Currently there are more than 700 employees at BSC facilities conducting research and supporting the HPC services.

Figure 2 represents the history of the MareNostrum system, different installation phases which were deployed inside of a disused chapel since 2004.

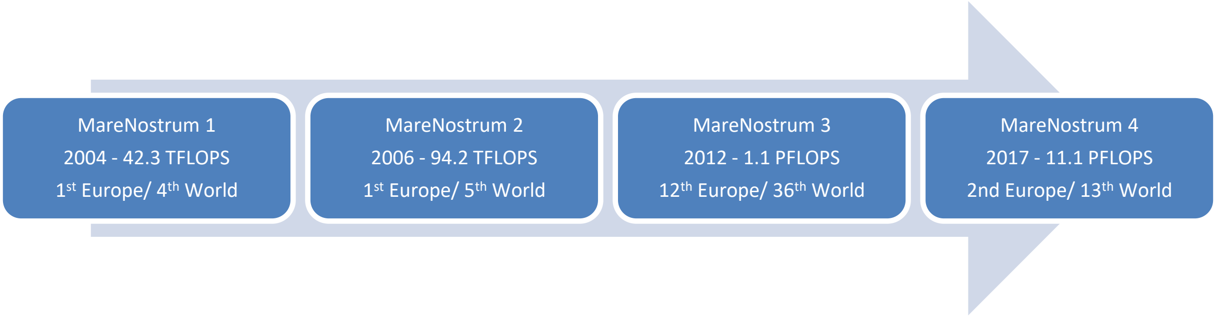


Figure 2. History of MareNostrum

In 2004, the data centre had a power capacity of 3 MVA, which was upgraded to 5 MVA in 2012. The power consumption of the system was increasing throughout the subsequent installation phases of MareNostrum reaching a limit of cooling capacity, despite the continuous efforts aimed at increasing the cooling capacity. **Table 1** outlines the details of the chapel's evolving infrastructure.

The mentioned cooling capacity limit required a change of the supporting building infrastructure. Currently BSC is in the process of moving to new headquarters, construction of which is going to be finalised by the end of 2020. **Figure** 3 illustrates BSC's new building built next to the chapel (see **Figure 4**) which will host the MareNostrum 5 pre-exascale EuroHPC system.

Figure 3. BSC new headquarters

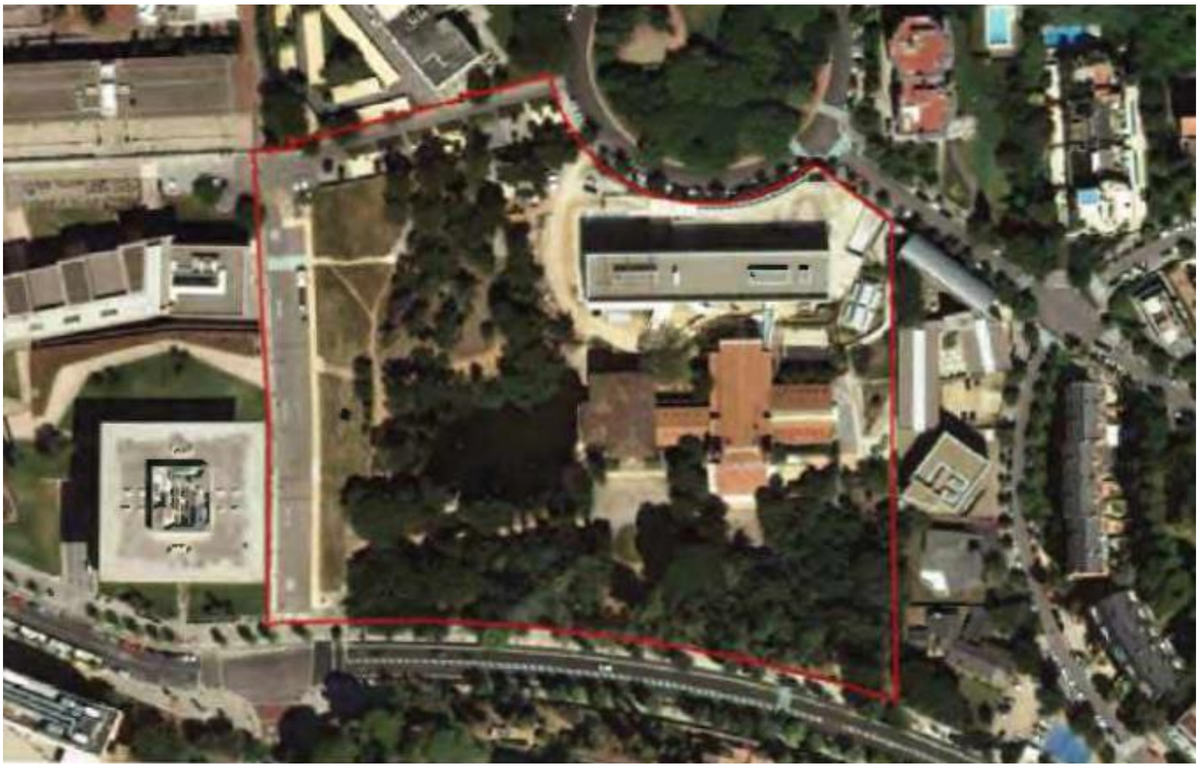


Figure 4. Location of BSC new headquarters

System	Year	Total Power	Power Consumption (kW)	Total Cooling Capacity (kW)	Cooling
MN1	2004	3x1 MVA (2+1)	650	Outdoors: 940 Indoor: 755	Air cooled Chillers: 4 x 235: STULZ MODELO CSO 2352 Crahs: 10 x 75,5: STULZ ASD-740
MN2	2006		750	Outdoors: 1175 Indoor: 896,4	Air cooled Chillers: 5 x 235: STULZ MODELO CSO 2352 Crahs: 8 x 75,5 + 2 x 146,2: STULZ ASD-(740-1500)
MN3	2012	2x2 MVA+1 MVA (partial redundancy)	1080	Outdoors: 2202,6 Indoor: 1400	Air cooled, RDHX Chillers: • 5 x 235: STULZ MODELO CSO 2352 • 2 x 513,8: CLIMAVENETA NECS/CA 2015 HxB: 2 x 1400 Crahs: 6 x 75,5 + 2 x 146,2: STULZ ASD-(740-1500)
MN4	2017		1300		Air cooled, RDHX

Table 1. MareNostrum - Chapel Infrastructure

MareNostrum 5. A European pre-exascale supercomputer

MareNostrum 5 is a EuroHPC Joint Undertaking (JU) consortium joined by Spain (hosting entity), Portugal, Croatia, and Turkey with a total investment of 217 million Euro: 150 million Euro for capital investment, with the rest being allocated to operational costs and additional projects[8].

The peak performance goal for MareNostrum 5 system is set to 200 PFLOPS. **Figure 5** illustrates the concept of MareNostrum 5 system.

[8] The acquisition and operation of the system will be funded jointly by EuroHPC JU, through EU's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as participating states Spain, Portugal, Croatia, and Turkey.

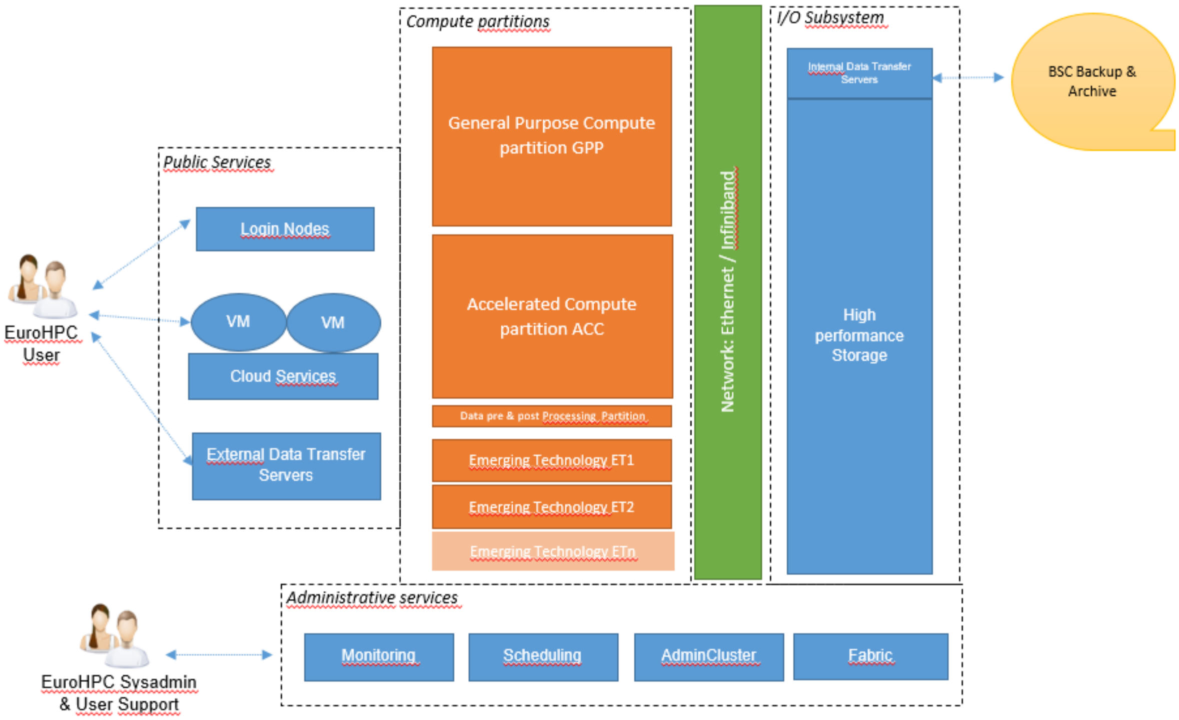


Figure 5. MareNostrum 5 system overview

The system will have a general-purpose (GP) compute partition, as based on the feedbacks received from the user community there is still a need to support scientists with general purpose processors. This on the other hand makes it rather difficult to achieve the objectives set by EuroHPC JU for having a system with a high peak performance. For that reason, an accelerated compute partition was added, thus making system heterogeneous. Certain maximum and minimum performance limits were set for each of the partitions. For example, for accelerated partitions it was decided to have a maximum sustained performance of 110 PFLOPS when running High Performance LINPACK (HPL). In addition, the system will feature emerging compute technologies (e.g. new general purpose, new accelerated technologies) together with high performance storage and very high bandwidth interconnect. The end-goal objective is to ensure the system's compatibility with existing applications and efficient assistance to stemming requirements from various domains ranging from earth and life sciences over engineering to AI and AI driven applications.

For achieving this, BSC needed to upgrade its power supply feeder and substation with a total capacity of 110 kV 2x25 MVA, expandable, if necessary, to 2x40 MVA. The expected initial power consumption is 20 MW. The first installation phase included the installation of a power line of 1145 m connecting BSC to general power grid. This underground connection is expected to be ready by system installation time. Currently it is estimated to have the system running in full production towards the end of 2021. The objective of the substation is to transform 110 kV to 25 kV. There will also be an emergency line of 5 MVA installed. One of the requirements set in the tender was to ensure that all energy provided, by contract, will be green energy. The list below outlines further tender requirements set for MareNostrum 5 infrastructure.

- The power consumption should not exceed 12 MW when running HPL
- The PUE should be below 1,08
- Certain requirements per rack were also indicated, such as power dissipation, weight and dimensions, cabling, visibility, etc.
  - One of the important requirements was to ensure that each rack dissipates minimum 95% of heat generated
- Cold water loop 18 °C up to 1 MW
- Warm-water loop 35 °C up to 12 MW
- Have an exhibition centre of MareNostrum5, i.e. to have a facility that can be visited

Under these conditions a public tender was conducted. The tender was awarded on 01.08.2019 and formalised on 26.11.2019. The awardee is Climava SL[9] with an awarded prize of € 12.557.990 (excluding VAT). The expected date for the delivery of the site was initially set for September 2020. Due to COVID-19 impact, it was later shifted to April 2021, meaning that no HPC system will be installed at BSC facilities before this date.

Figure 6 and **Table 2** indicate the available space for hosting MareNostrum 5 system.

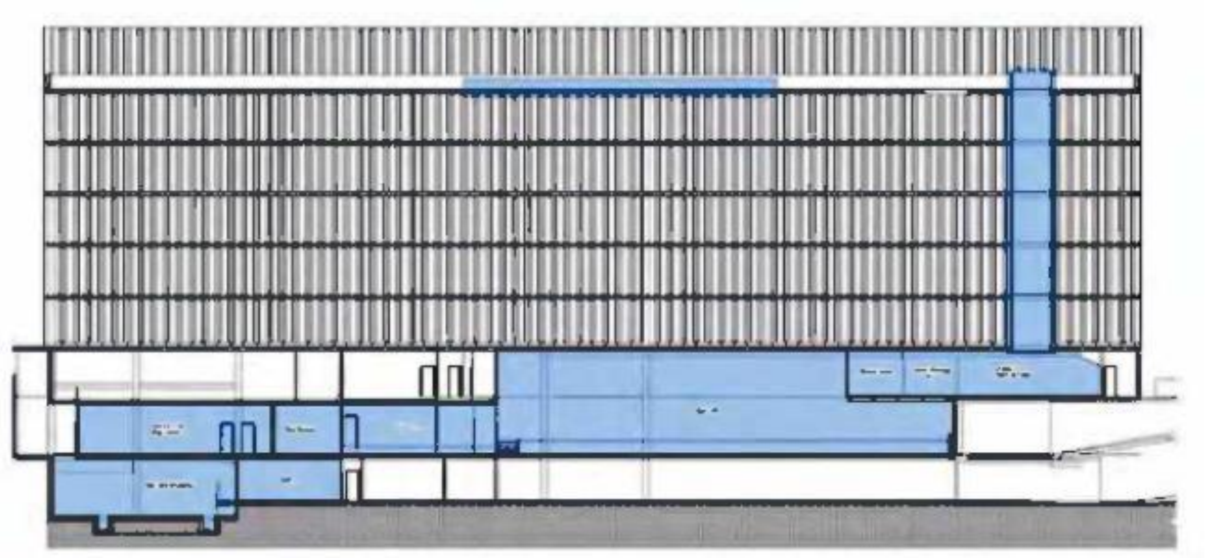


Figure 6. Space available for MareNostrum 5

As BSC intends to keep this facility operational for at least 15 years (as was previously the case with the chapel), the electrical loads were designed accordingly. **Table 3** outlines the further details, indicating the expected power consumption of IT systems as well as supporting infrastructure (e.g. chillers, pumps, etc.). As can be seen, it is planned to also keep MareNostrum 4 in production (i.e. thus also maintaining chapel in production with a better efficiency) with its full capacity together with another Torre Girona high-end IT system.

Floor		**m^2**	**Total**
P-3	Transformers	426	475
	Fire extinction	49	
P-2	Compute room	847	1374
	Access to compute room	46	
	Batteries room	73	
	Low voltage room	408	
P-1	Room for chillers and pumps	466	711
	Riser/ “PATIO”	9	
	Visitors area	236	
Roof		320	320
Total		Rounded	2880

Table 2. Overview of floor space occupation

[9] https://www.climava.com/

Electrical Loads		
Area	**Load Type**	**kW**
MareNostrum 5	IT Load	13840
	Critical IT Load	1160
	Total IT	**15000**
	Chillers	663
	Cooling Towers	182
	Pumps	185
	M&E rooms ACV	51
	Auxiliary UPS	30
	Total M&E	**1111**
	Total MareNostrum 5	**16111**
MareNostrum 4	IT Load	1300
	Critical IT Load	200
	Total IT	**1500**
	M&E Load	100
	Total M&E	**100**
	Total MareNostrum 4	**1600**
Torre Girona	IT Load	480
	Critical IT Load	120
	Total IT	**600**
	M&E Load	200
	Total M&E	**200**
	Total Torre Girona	**800**
Total		18511
PUE[10]	MareNostrum 5 + MareNostrum 4	1,07

Table 3. Electrical loads

Figure 7 illustrates the diagram of the supporting infrastructure. As can be seen, there are 5 transformers shown in the bottom part of **Figure 7**, that transfer the electrical energy to the main distribution panel.

[10] Overall PUE for MareNostrum 4 + MareNostrum 5 is estimated by considering that they both share the Cooling Tower and Chiller Plant. During this estimation, rated equipment capacity was assumed, underestimating partial or seasonal loads.

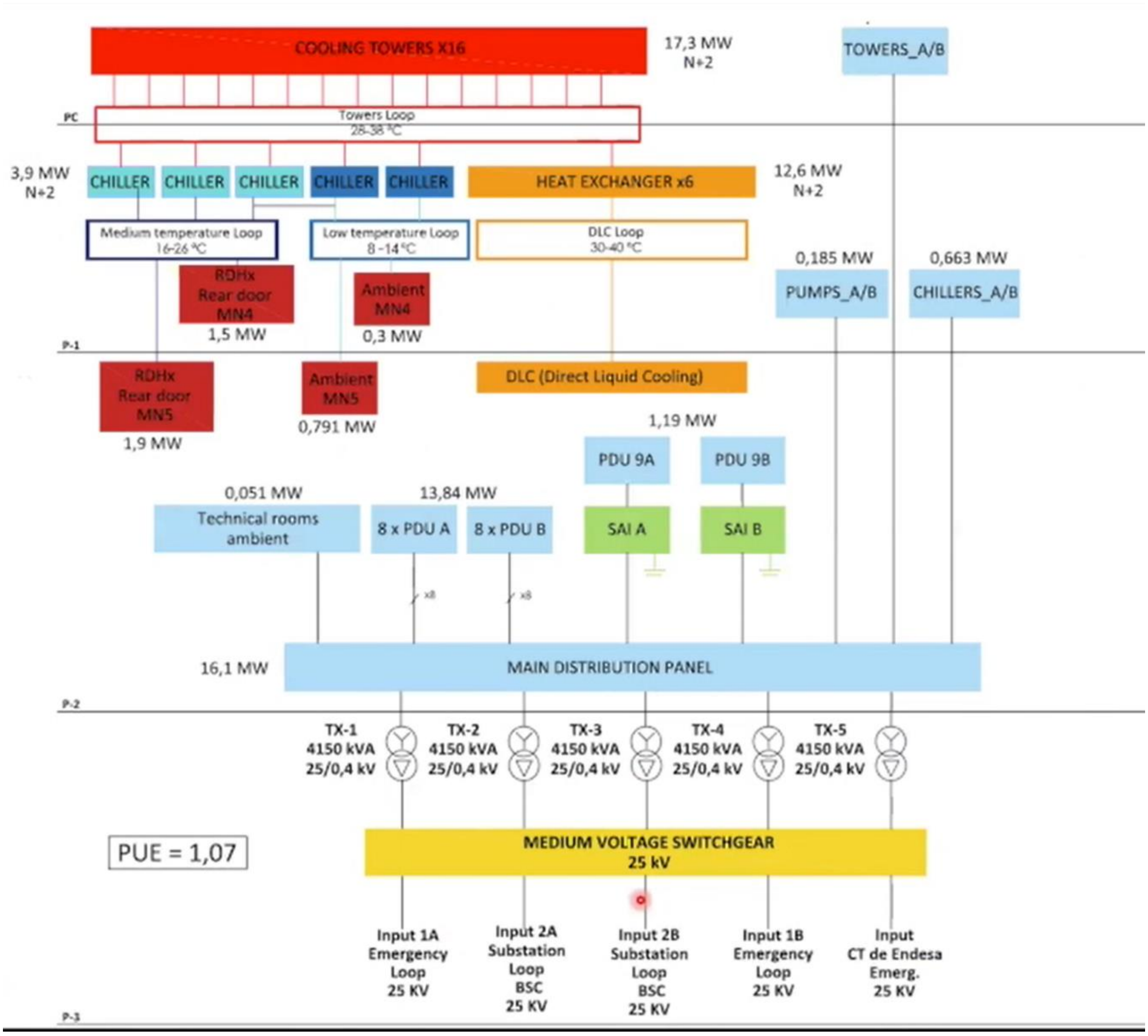
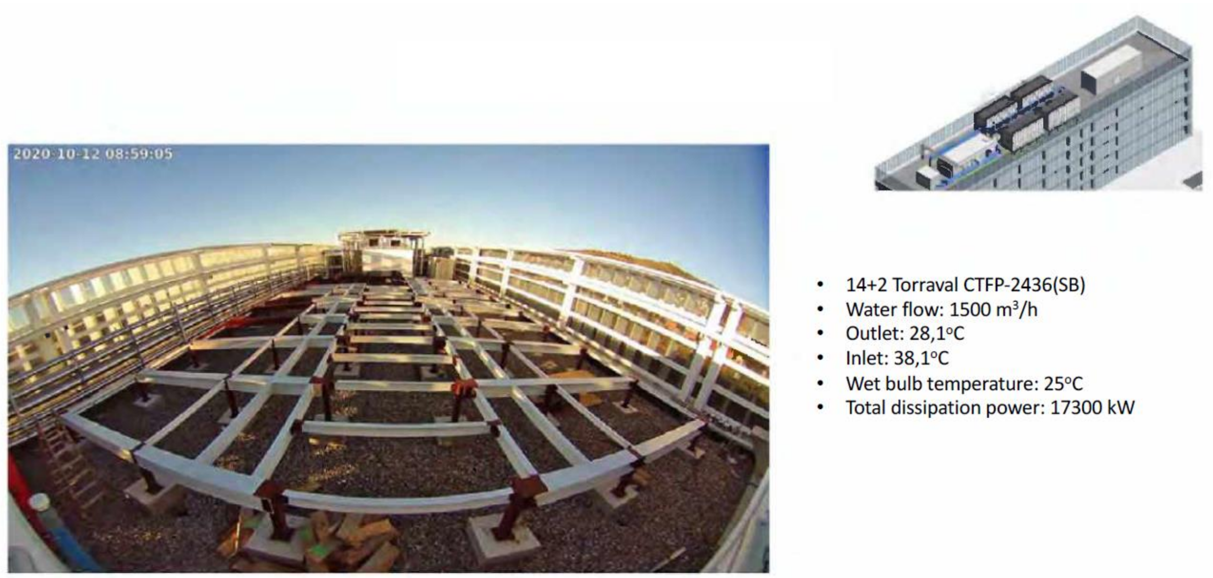


Figure 7. Infrastructure Diagram

The main distribution panel in turn feeds the power to Power Distribution Units (PDUs) and Uninterruptable Power Supplies (UPS)[11]. From the upper part of **Figure 7**, it can be seen that the cooling towers (14+2) are operating within a temperature of 38 °C for inlet and 28 °C for outlet. As illustrated in the diagram, the water then flows via heat exchangers to compute servers. **Figure 8** shows the current roof view and indicates further specifics of the cooling towers, while **Figure 9** and **Figure 10** show the compute room and a snippet from a virtual compute tour. The latter one indicates how visitors can see as the flagship system as well as specifics of the supporting building facility.

[11] Labeled SAI (Spanish word for UPS) on Figure 7.



- 14+2 Torraval CTFP-2436(SB)
- Water flow: 1500 m^3/h
- Outlet: 28,1^{o}C
- Inlet: 38,1^{o}C
- Wet bulb temperature: 25^{o}C
- Total dissipation power: 17300 kW

Figure 8. Roof of the facility

Figure 9. Compute Room

Figure 10. Snippet from a virtual tour

3 PRACE Session, chaired by François Robin

3.1 HLRS Update

Presenter: *Bastian Koller, HLRS*

HLRS I

In the beginning of 2005 HLRS had one computer room and some offices in Nobelstrase 19.

At the time, the 1 MW total energy supply was considered enough for current and future supercomputers. The energy needs for a supercomputer did increase a couple of years after becoming operational and is still a constant source of debate.

The system has an 860 kW UPS supported by batteries and with a hold-up time of 15 minutes. The cooling capacity at that time was 1.2 MW and was done using dry coolers with approximately 50% free cooling[12] and around 50% district cooling.

HLRS II

Since 2019 there are 5 buildings and the new "green building" has 3.6 MW energy supply. With a UPS installation of 3.6 MW that can hold all infrastructure for 15 seconds based on rotating masses technologies composed of a tri-phasic with 1500 A per phase.

Cooling for 4 MW has 4 wet-cooling towers with 85% free cooling around the year and around 15% is used for district cooling. There are 8 transformers of 1.2 MVA each. Six transformers are actively used for HPC system production, one is on standby mode, and one is used for building infrastructure. Inlet temperature was 10-14 °C on the old systems and 25-30 °C on the current one.

The current system is still under balancing state so the temperatures can vary. During the balancing the building infrastructure tries to adjust its cooling capacity to the changing needs of the computing equipment. Especially the high and low temperatures of the computing equipment and the rate of change need to be tested during the balancing time for the cooling infrastructure to be able to adjust its own functions.

The computer system is not the limiting factor in the inlet temperature but the building infrastructure. This means the inlet temperatures accepted by the computer system are wider in range than what the building infrastructure is providing.

[12] Yearly average.

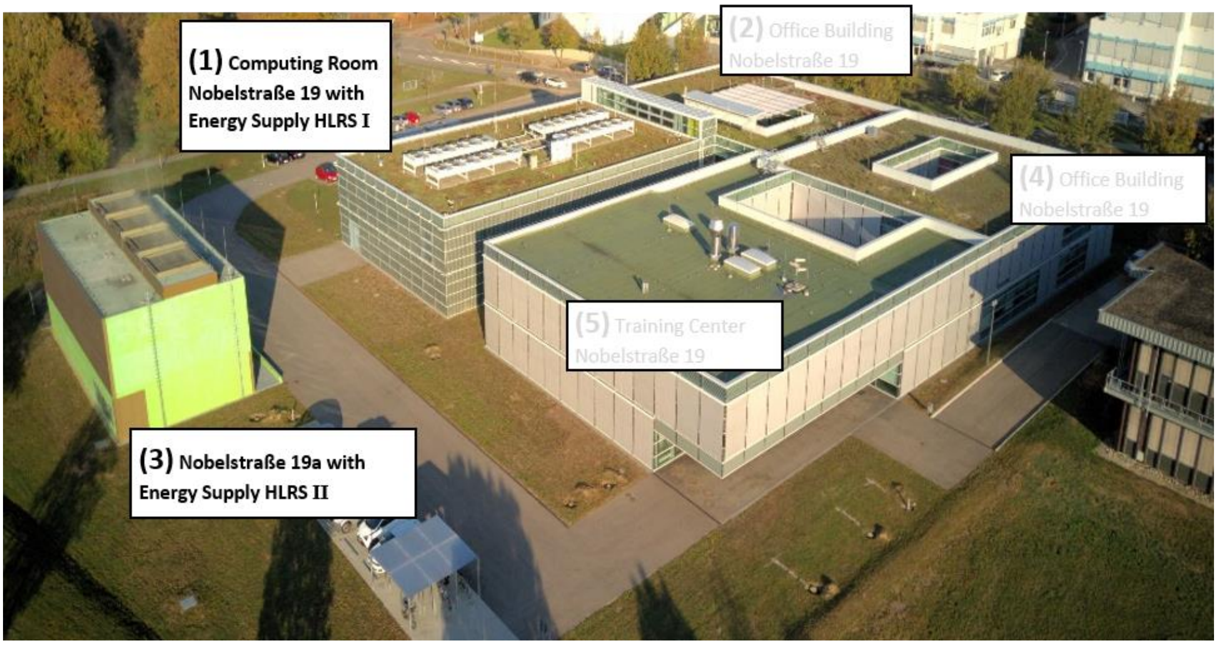

(1) Computing Room Nobelstraße 19 with Energy Supply HLRS I
(2) Office Building Nobelstraße 19
(4) Office Building Nobelstraße 19
(5) Training Center Nobelstraße 19
(3) Nobelstraße 19a with Energy Supply HLRS II

Figure 11. Picture of the site, Nobelstrase 19: 1. HLRS I was the first building, 2. office building, 3. HLRS II, 4. office building, 5. training centre.

The new system named HAWK with a maximum 4 MW power consumption is specifically designed for engineering requirements of computational fluid dynamics (CFD) and capability computing. It has been optimised for sustained performance of HLRS key applications.

Infrastructure constraints were important and had to be taken into account for system design. It seems the supercomputer centres making a step towards pre-exascale are divided into groups of those who already have the facilities and infrastructure and those who do not.

The Hawk system has 5632 AMD Rome 2.25 GHz CPUs with around 720000 cores. Hawk has 1.44 PB total memory, peak performance of around 27 PFLOPS. Work space filesystem capacity is 25 PB with bandwidth of 250 GB/s and interconnect done by HDR200 InfiniBand. Interconnect topology is enhanced hypecube (9D).

Rack configuration: 44 compute racks each having a maximum power consumption of 90 kW and a total of 2112 power supplies in the system. During normal operating conditions, the system uses around 3.5 MW and during HPL it consumes around 4.1 MW. This HPL run is already on the maximum limit of the total power supply and needs some very specific conditions for a proper run.

There are 16 racks per row and 128 nodes per rack. Each row of racks uses direct liquid cooling with the cooling units at both ends of the rack rows. A cooling unit of 1 MW has temperature difference between inlet and outlet water of 10 Kelvin.

Compute node configuration: AMD Rome CPU with 64 Zen2 cores per socket and 2 sockets per node. A standard node has 128 GB DDR4 memory of 3200 MHz 8 GB RDIMM. Each chip consists of an I/O die built at 14 nm and 8 CPU chiplets at 7 nm with 8 Zen2 cores per chiplet. 200 nodes are equipped with 256 GB memory. The memory bandwidth is around 380 GB/s. Nodes also have a PCIe4 HDR InfiniBand adapter card.

The interconnect topology of a partial 9D Hypercube is constructed of 1 rack of 128 nodes build in a 3D cube. Mellanox Infiniband HDR 200 GBit/s 40-port premium switch is used as building block for the interconnect. The MPI latency is around 1.3 microseconds to the nearest neighbour. The hypercube was seen to be the best working option for the applications.

The high-performance storage system of 25 PB usable capacity with 14 TB hard drives is made by DDN. The product is called Exascaler which provides a Lustre filesystem. DDN Infinite Memory Engine (IME) solution works as buffer cache and has 250 GB/s bandwidth of 800 TB NVMe SSD's. There are 48 InfiniBand links into the fabric. Storage system is easily expandable for performance and capacity.

Sustainability

During the last few years HLRS has taken environmental sustainability into account by obtaining certificates.

HLRS has the ISO 14001 environmental management certificate, an internationally recognised standard for the implementation of a comprehensive sustainability management system. ISO 14001 is a central component of Eco-Management and Audit Scheme (EMAS).

At HLRS these efforts impact all levels of the organisation, including:

- considering environmental impact in all purchasing decisions
- improving energy efficiency across the centre's operations
- supporting advanced research that will lead to widespread sustainability gains in domains such as transportation, energy and climate.
- minimising waste and avoiding use of environmentally damaging materials
- reusing resources, including heat generated during operation of its HPC systems
- promoting sustainability improvements among its peers
- maintaining a work environment that promotes the health and well-being of employees and their families.

ISO 50001 energy management is an international norm for energy management systems that supports companies and organisations of all sizes and sectors in the development of a systematic energy management system. The main goal of an energy management system is the improvement of energy efficiency, including the reduction of energy usage as well as the adherence to legal requirements.

At high-performance computing centres like HLRS that require large amounts of power for the operation and cooling of their computers, this means committing to implement measures that will minimise carbon emissions and consumption of non-renewable resources. At HLRS this has involved setting targets for energy usage, tracking actual energy consumption and making infrastructure improvements to optimise energy efficiency.

The last certificate has been the Blue Angel certificate and HLRS is probably the first HPC centre getting this label. The Blue Angel is the best-known eco label in Germany and since 1978 has served as a neutral and credible guide for selecting environmentally friendly products and services. In contrast to other labels in the IT sector such as EnergyStar, the Blue Angel comprehensively tests the entire environmental performance. It also considers the capacity utilisation of servers, the usage of green electricity and the avoidance of climate-damaging gases in cooling systems.

HLRS has also EMAS certification for voluntary environmental management developed by the European Community. EMAS is the most demanding system for environmental management worldwide; HLRS was the first high-performance computing centre to qualify for EMAS certification.

By participating in EMAS, companies and organisations from numerous sectors can improve their ability to make progress in environmental protection. Organisations that achieve EMAS certification commit themselves to maintain stringent environmental standards, to continuously improve their environmental performance, to participate in regular reviews by an environmental auditor and to publish an environmental statement.

Centre sustainability

The HLRS centre has reached the limit of power consumption which is enough for the current system. But with the ongoing potential increase in power consumption and internal projects there are things to consider for the future. The two options for the future are either improving the existing building or build a new one. At this point, a new building seems likely, but it involves the same problems as before, which is where to place it. During the next Infrastructure Workshop there will be more information concerning future HLRS plans.

Demand for power starts from 25 MW to 30-40 MW for the future. The demands need to be considered against the price, views of the government and the support of the public. Electricity is restricted to all university buildings in total to around 8 MW. For the future, the coexistence with the university has to be decided so the computing systems are not restricted by the rest of the university power needs. For sustainability of the power it is important to prepare for the future which can be more than what is seen now.

For now, 8 MW is sufficient but it is not enough for the future. HLRS needs to figure out what is sufficient for power consumption 5-6 years from now. Pursuing a 25 MW capable infrastructure in the future may bring political hurdles, so 12-15 MW may be more easily obtainable.

Since the site is near residential areas, the noise from the cooling equipment has to be taken into consideration.

3.2 HPC infrastructure at Luxembourg University

Presenter: *Ezhilmathi Krishnasamy, University of Luxembourg*

(Summary provided by the presenter)

The HPC department at the Luxembourg University has the four domains, namely: 1) infrastructure; 2) services; 3) expertise; and 4) education & training. The infrastructure is comprised of state-of-the-art HPC systems with 2.7 PFLOPS compute capacity (Iris and Aion). Both HPC machines are housed in a highly capable data centre (Centre De Calcul CDC), which provides air cooling facilities for the Iris cluster and liquid cooling for the the Aion cluster. The UL HPC provides compute and data services to researchers from the Luxembourg University, connected researchers, and private sector within Luxembourg. Any researcher who is connected with Luxembourg University is eligible to use the UL HPC compute and data service. UL HPC services are mostly related to supporting the researchers' questions and facilitating their research. Third, the UL HPC has a domain expert in parallel programming, computer science and data science. Finally, UL HPC provides education and training in data science and parallel programming. The UL HPC conducts annual HPC schooling, covering a top to bottom approach related to HPC (parallel programming, data science, engineering simulation and machine learning); and gives master courses in parallel programming and grid computing. This makes the UL HPC the best computing centre in Luxembourg. Since then it has been growing and now operates Iris and Aion clusters with a compute capacity of 2794.23 TFLOPS and 10713.4 TB storage capacity. The UL HPC's compute capacity is shared by two clusters: the Iris system shares TFLOPS (39%), and the Aion shares 1693 TFLOPS (61%).

The UL HPC started in 2007 under the supervision of Prof. Pascal Bouvry and Dr. Sebastien Varrette.

The Iris cluster has an Intel CPUs (Skylake and Broadwell) with Nvidia V100 GPUs. In total Iris has 196 nodes, and 24 nodes have 4 Nvidia V100 GPUs. The Aion cluster has AMD EPYC CPUs with total of 318 nodes and 40704 cores.

UL HPC’s compute and storage capacity places the UL HPC in the Tier 2 support category of the HPC centre in Luxembourg. In particular, the Iris cluster has a total of 5824 compute cores with 52224 GB RAM and fast InfiniBand (IB) EDR network. On the other hand, the Aion cluster has, 4 BullSquana XH2000 adjacent racks with 318 compute nodes of AMD EPYC CPUs. With a total of 40704 compute cores and a total of 81408 GB RAM. The network of the Aion has a fast InfiniBand (IB) HDR100 Fabric network. And finally, UL HPC has a small cluster of g5k with Intel Xeon CPUs of 38 nodes, 368 cores and 4.48 TFLOPS capacity. **Table 4** shows the evolution of computing capacity in recent years.

Year	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
CPU nodes	66 TFLOPS	66 TFLOPS	306 TFLOPS	306 TFLOPS	306 TFLOPS	1933 TFLOPS
GPU nodes	83 TFLOPS	83 TFLOPS	83 TFLOPS	881 TFLOPS	881 TFLOPS	789 TFLOPS
Large Memory nodes	13 TFLOPS	13 TFLOPS	13 TFLOPS	46 TFLOPS	46 TFLOPS	33 TFLOPS
Total	162 TFLOPS	162 TFLOPS	402 TFLOPS	1233 TFLOPS	1233 TFLOPS	2764 TFLOPS

Table 4. UL HPC’s computing capacity over the past five years and year 2020 shows the computing capacity of both Iris and Aion

The storage capacity has four file systems, namely: GPFS/SpecturmScale (HOME, projects); Lustre (SCRATCH), OneFS (Projects, Backup) - shared with UL IT Department and other (for Backup). **Table 5** shows the shared capacities of the file systems at the UL HPC.

Year	2015	2016	2017	2018	2019	2020
GPFS/SpectrumScale (Home, Projects)	1140 TB	1140 TB	2449 TB	2449 TB	3320 TB	4411 TB
Lustre	600 TB	600 TB	600 TB	1902 TB	1902 TB	1302 TB
OneFS (Projects, Backup) shared with UL IT Department	1594 TB	1594 TB	1594 TB	1594 TB	3360 TB	3360 TB
Other	2030 TB	2030 TB	2630 TB	2630 TB	2630 TB	1608 TB

Table 5. UL HPC's storage capacity over the past five years

So far, we have seen the UL HPC's compute capacity and storage capacity; now, we will briefly focus on the software available at the UL HPC. Computational software at UL HPC is installed through the EasyBuild framework, which has 319 packages, 239 software packages and 19 categories. In addition, the updated software sets are released twice per year for the users.

The operating system of the UL HPC is Linux CentOS/Redhat with the following setup:

1. User single sign-on – Redhat idM/IPA
2. Remote connection & data transfer – SSH/SFTP
3. User portal – Open OnDemand
4. Server/Compute Node Deployment – BlueBanquise, Bright Cluster Manager, Ansible, Puppet and Kadeploy.
5. Virtualisation and Container Framework – KVM and Singularity
6. Platform Monitoring (User level) – Ganglia, SlurmWeb and OpenOndemand
7. ISV software – ANSYS, ABAQUS, GROMAS, LAMMPs, compilers and performance tools.

The UL HPC is also the part of the Grid 5000[13], a large nationwide infrastructure (large scale parallel and distributed computing research) in France. It has in total 8 sites with 7 in France and 1 in Luxembourg. The Grid 5000 comprises 38 clusters with 15812 cores, connected to RENATER[14] network with a with 10 GB/s link. The software stack is deployed by the kadeploy, kavlan and kwapi.

3.3 Distributed national HPC/cloud and data infrastructure

Presenter: *Norbert Meyer, PSNC*

(Summary provided by the presenter)

The presentation described the vision of Polish HPC and Data infrastructure, which is being developed under three national projects located at the Polish Roadmap of Research Infrastructure of the Ministry of Science and Education (Polish part of ESFRI roadmap)[15,16].

Modern science and economy use advanced technologies such as HPC to process huge amounts of data (Big Data) and elements of artificial intelligence (AI). Computing services of this scale can only be provided with the use of comprehensive management platforms, in the infrastructure of technologically advanced communication and data

[13] https://www.grid5000.fr/w/Grid5000:Home

[14] https://www.renater.fr/

[15] Polish ROadmap of Research Infrastructure (PMDIB), http://www.bip.nauka.gov.pl/g2/oryginal/2020_01/3175b4b7b9daae8d0f255b3670d54361.pdf

[16] ESFRI (European Strategy Forum on Research Infrastructures), https://www.esfri.eu/esfri-roadmap

systems. PRACE-LAB[17] and PRACE-LAB2[18] are projects included in the Polish Roadmap of Research Infrastructure, aiming to fulfil HPC and cloud requirements of the scientific community and industry in Poland.

The two projects mentioned above are the Polish equivalent of the European PRACE (Partnership for Advanced Computing in Europe). The hardware and service platform that is being developed in the projects will be a part of the European research infrastructure: PRACE (HPC), EuroHPC, EUDAT CDI (data), EOSC-hub and EOSC-Synergy (cloud computing and data clouds).

The main goal of the PRACE-Lab projects is to provide advanced HPC computing services using supercomputers and data infrastructures for the scientific community in Poland and Europe, and for research in industry, especially Industry 4.0 applications. It is assumed that thanks to the implementation of the project, the Polish economy will improve its position in European and world markets because of the support and strengthening of the use of innovative solutions. Particular emphasis was placed on the possibility of cooperation between small and medium-sized enterprises (SMEs) and the science sector. The innovative approach of SMEs and the search for new solutions and their own position on the market are largely the driving force of the economy, which was also noted in the economic development plan of the Polish government.

The projects introduced three implementation stages:

1. Development of system architecture and preparation of computing services and data management on HPC e-Infrastructure (2019-2020)
2. Integration of the implemented services in the form of a demonstrator along with conducting system tests (2021)
3. Testing and operation of the PRACE-LAB platform (2022-2023).

The HPC and Data ecosystems are geographically distributed and built in a cooperation with a scientific consortium consisting of 5 HPC centres and 4 universities supported by IT vendors and SME companies. At present, the first stage of building a distributed HPC and data infrastructure in Poland is presented at Figure 12.

[17] http://www.prace-lab.pl/

[18] http://www.prace-lab2.pl/

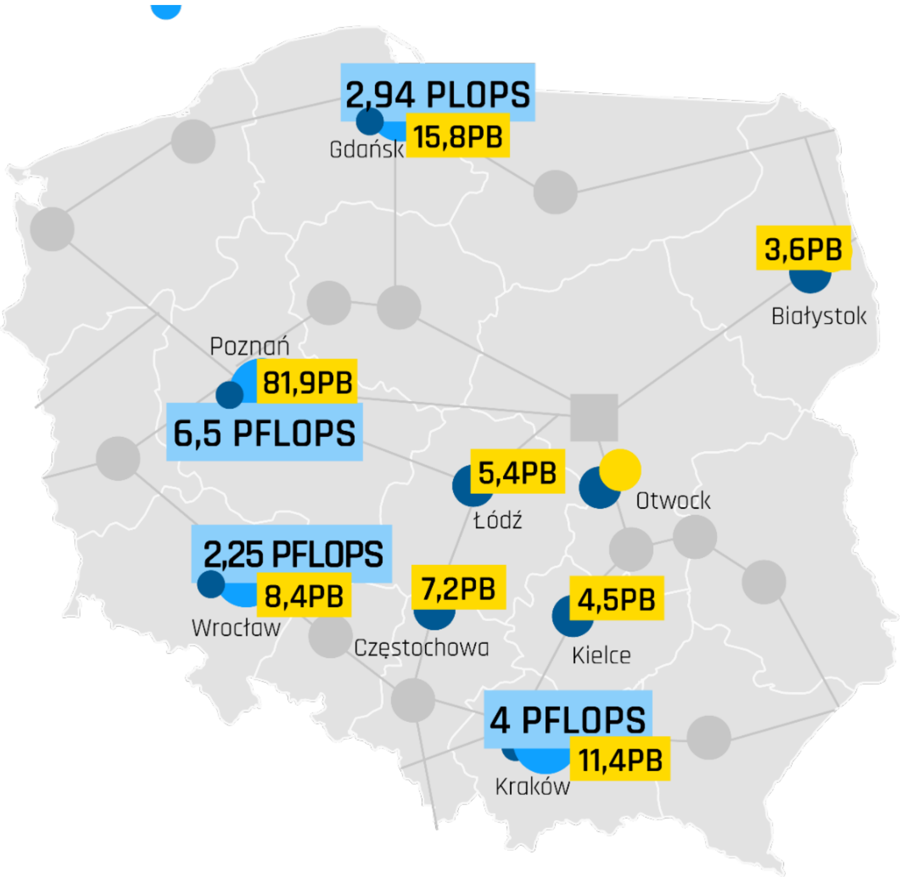


Figure 12. Distributed e-infrastructure locations and parameters of PRACE-LAB – the first stage completed by 16 PFLOPS of conventional power and 139 PB of data infrastructure

The resulting platform will be used by industry (40% of the installation) and academic sector (60%). The HPC and cloud environments (both using the same Infrastructure) will be used by governmental institutions. The scope of providing services extends well beyond the country's borders. Access to the non-economic part of the system is granted to research units from universities, institutes of the Polish Academy of Sciences and National Research Institutes.

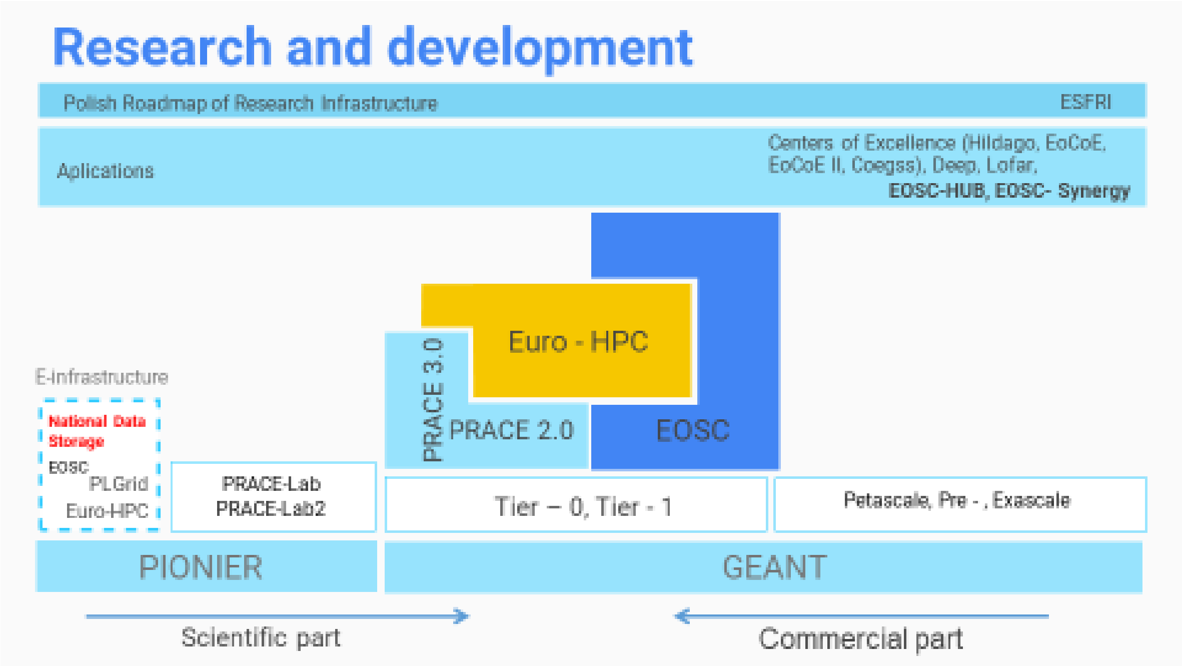
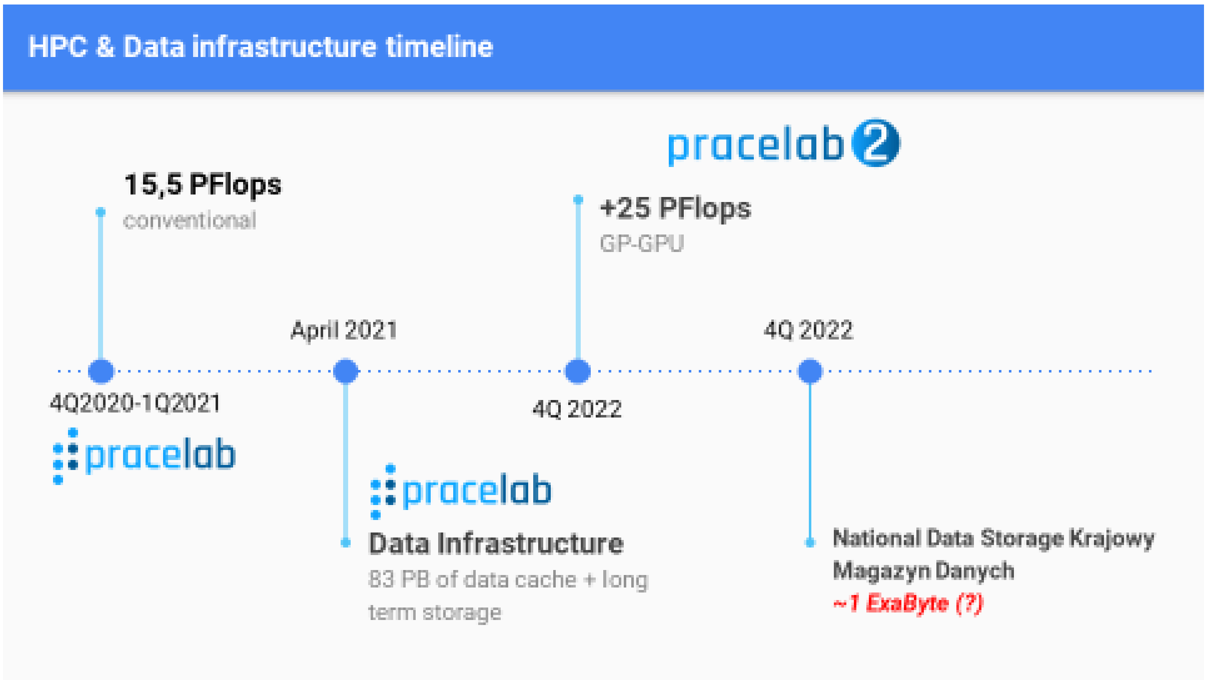


Figure 13.The position of PRACE-LAB and PRACE-LAB2 at the European infrastructure projects

The timeline of infrastructure development in Poland aims to reach at least 40 PFLOPS of HPC and 1 EB of data storage (Figure 14) the end of 2022.



HPC & Data infrastructure timeline

15,5 PFlops
conventional

4Q2020-1Q2021

April 2021

pracelab

Data Infrastructure
83 PB of data cache + long term storage

pracelab 2

+25 PFlops
GP-GPU

4Q 2022

4Q 2022

National Data Storage Krajowy Magazyn Danych
~1 ExaByte (?)

Figure 14. Development of Polish HPC e-infrastructure data in PRACE-LAB, PRACE-LAB2 and NDS (National Data Storage)

In order to meet the user requirements, it is necessary to meet many points regarding the security and reliability of the services provided. The expectations of the scientific community and industry create a new quality that we have not realised so far supporting only scientific users and its demands.

These new requirements affect not only the security and quality of direct services provided, but also the reliability of data centre infrastructure and the communication layer between data centres, including:

1) Due to the provision of computing/cloud services for the economy (40% of infrastructure), it was decided to require certification of HPC and data centres following ISO9001 and ISO27001 procedures in the field of:

    a. Provision of data processing and storage services;

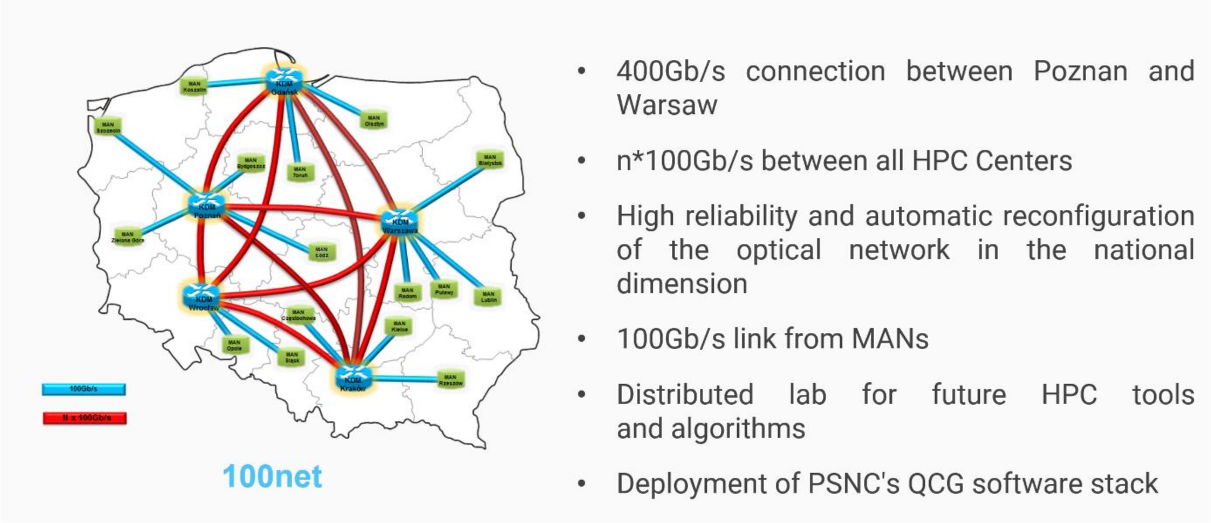
    b. Provision of cloud services;

    c. Internet platforms;

    d. Provision of colocation services;

    e. ICT consulting and security.

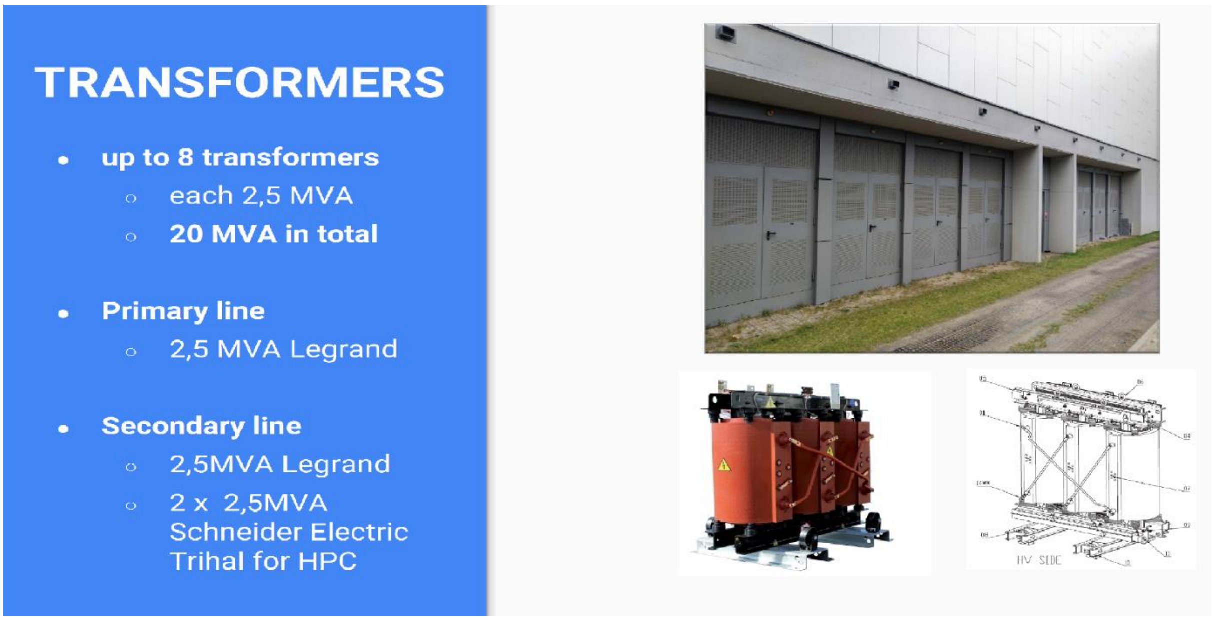
2) Dedicated fiber optic connections (full mesh) between HPC centres and distributed data infrastructure points of access were provided - dedicated 400 Gbps connections



- 400Gb/s connection between Poznan and Warsaw
- n*100Gb/s between all HPC Centers
- High reliability and automatic reconfiguration of the optical network in the national dimension
- 100Gb/s link from MANs
- Distributed lab for future HPC tools and algorithms
- Deployment of PSNC's QCG software stack

Figure 15. Dedicated "100net" network connecting Polish HPC Centres

3) In each HPC centre, a protected and unprotected redundant power supply was prepared for elements of critical computing services and data storage services

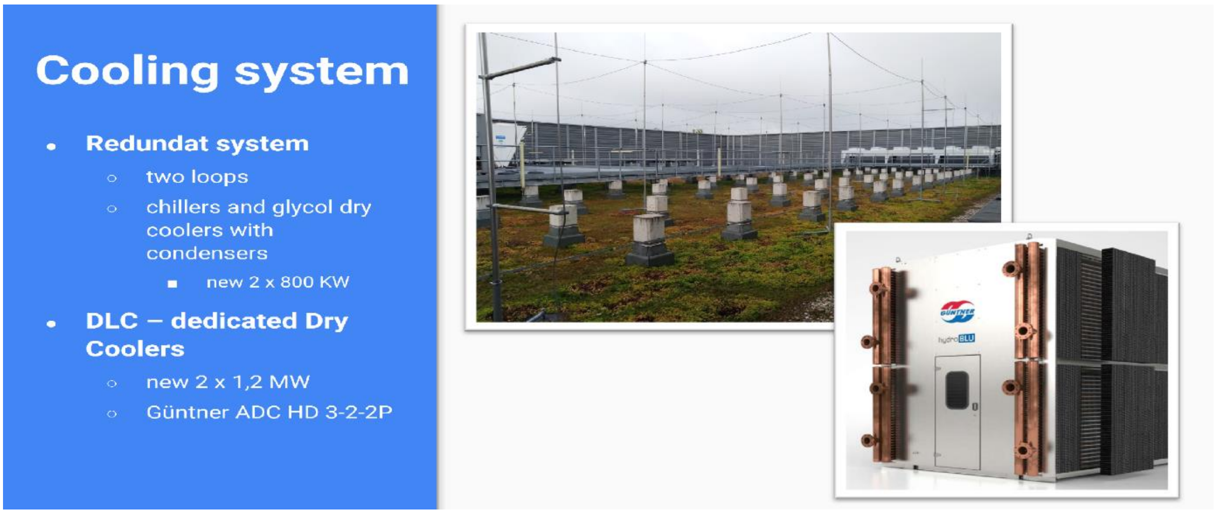


TRANSFORMERS

- **up to 8 transformers**
  - each 2,5 MVA
  - **20 MVA in total**
- **Primary line**
  - 2,5 MVA Legrand
- **Secondary line**
  - 2,5MVA Legrand
  - 2 x 2,5MVA Schneider Electric Trihal for HPC

Figure 16. Overview of Transformers

4) Each site has a high-efficiency cooling system for HPC - DLC clusters, capable of supporting the cooling of the entire infrastructure with redundant cooling towers

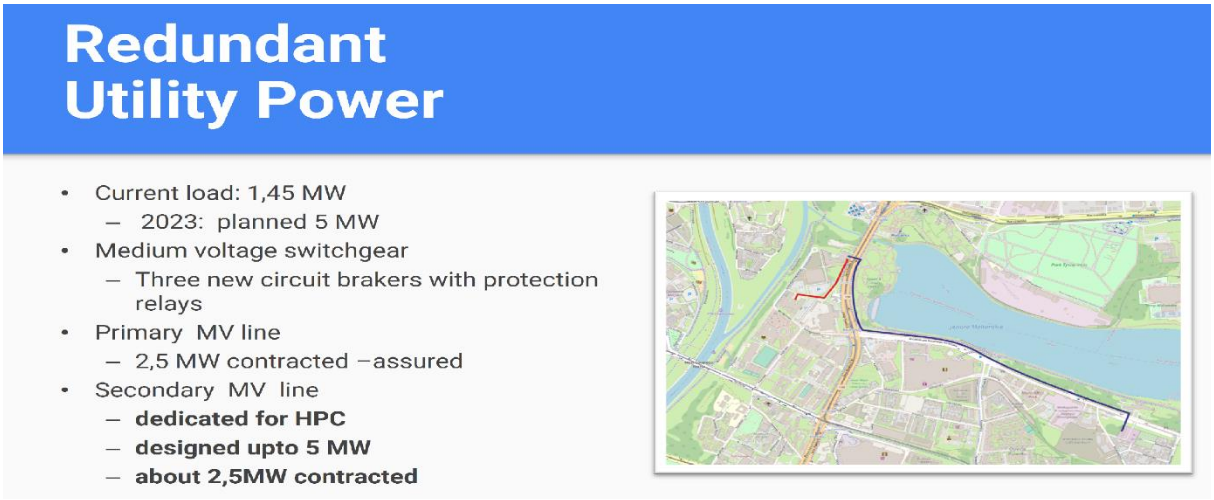


Cooling system

- Redundat system
  - two loops
  - chillers and glycol dry coolers with condensers
    - new 2 x 800 KW
- DLC – dedicated Dry Coolers
  - new 2 x 1,2 MW
  - Güntner ADC HD 3-2-2P

Figure 17. Overview of the cooling solution

5) Two independent power lines were connected to the data centre (medium voltage)



Redundant Utility Power

- Current load: 1,45 MW
  - 2023: planned 5 MW
- Medium voltage switchgear
  - Three new circuit brakers with protection relays
- Primary MV line
  - 2,5 MW contracted –assured
- Secondary MV line
  - **dedicated for HPC**
  - **designed upto 5 MW**
  - **about 2,5MW contracted**

Figure 18. Power lines

The implementation of steps 1 to 5, although not necessary from the point of view of HPC services, significantly increases their reliability and security. Thus, it is indispensable from the business point of view as an element that guarantees the continuity of business operations.

3.4 Update on the PPI4HPC project

Presenter: *Dirk Pleiter, JSC and PDC*

(Summary provided by the presenter)

In the following, we describe the current status of the PPI4HPC project[19]. This project started in 2017 and was the first Public Procurement of Innovative Solutions (PPI) in the field of HPC. The objective of the project was to jointly buy innovative IT equipment such as supercomputers or high-end storage systems and make them available as production systems for scientists and engineers.

A PPI[20] is a co-funding instrument of the European Commission (EC) to stimulate innovation in the European economic zone. The strategy is to provide incentives for the public sector to act as a launching customer for innovative solutions to support their introduction on the market. These solutions must not yet be widely available

[19] https://ppi4hpc.eu/

[20] https://ec.europa.eu/digital-single-market/en/public-procurement-innovative-solutions

on the market. To enlarge the market impact and to mitigate risks the EC expects several public procurers to form a buyers group and run a joint procurement. In the case of PPI4HPC, the buyers group was comprised of BSC (Spain), CINECA (Italy), GENCI (France) and Forschungszentrum Jülich (Germany). The project furthermore involved CEA (France), who supported GENCI.

One of the goals of the PPI4HPC project was to foster science and engineering applications in Europe by providing more computing and/or storage resources through PRACE. The project furthermore aimed for promoting research and innovation in Europe in the area of HPC architectures and technologies in Europe, ideally leveraging the outcomes of the Pre-Commercial Procurement (PCP), which had been executed during the PRACE-3IP project[21]. A final goal was to have greater weight and more impact on common topics of innovation and the design of the solutions according to the needs of scientists and engineers in Europe by a coordinated approach. The PPI4HPC partners spent a significant amount of time on identifying technical topics of interest. The list of topics ranged from energy efficiency and power management, data management, programming environments, to security.

The official part of the procurement started in September 2017 with an open dialogue event, where the buyers group informed the market about its intentions, the technical needs as well as the planned organisation of the joint procurement. This allowed interested suppliers to provide feedback before the final version of the tender documents were published in May 2018. PPI4HPC's joint procurement did foresee a joint publication of the tender as well as a joint selection of the qualified candidates. Thereafter, each member of the buyers group continued within a separate lot. One of the advantages of this approach was that each of the lots could progress at different speeds depending on when the national funding was available. Meanwhile, contracts have been successfully awarded in all four lots and four solutions have been deployed or will be deployed by early 2021.

Within the French lot, a contract was awarded to Atos, who delivered a system comprising a major compute partition, an HPDA/AI converged partition as well as an exploratory partition, which will be based on Arm processors from Fujitsu. Atos was also successful in the German lot and will deliver a system optimised for data-intensive applications by the end of 2020, which will be the successor of the current JURECA Cluster. In Italy, a contract was awarded to IBM, who delivered Marconi-100, a system based on POWER9 processors and NVIDIA V100 GPUs, which in June 2020 was listed on position 9 of the TOP500 list. Also, in Spain, a contract was awarded to IBM, where the company delivered a consolidated central storage infrastructure.

All these solutions strictly conformed to an innovation criterion, which the buyers group formulated to ensure that the procedure would qualify for a PPI. The tender documents required from each supplier to identify the innovative components of the offered solution. Only components could be listed that were relevant for achieving performance targets or meeting key functional requirements. They had to be recent and should not have been introduced into the market long ago, as public procurers within a PPI are supposed to act as launch customers.

The offered innovations concerned among others solutions for energy and power management. Atos offered in both of their successful bids the Bull Energy Optimizer (BEO), a power data collection and analysis infrastructure, and Bull Dynamic Performance Optimizer (BDPO), a set of tools for dynamic steering of hardware settings to minimise energy consumption. IBM integrated Examon, an infrastructure for fine-grained power monitoring and power capping developed by University of Bologna and E4. BEO, BDPO and Examon had been partially developed within the already mentioned PRACE-3IP PCP.

Several innovations from the area of data management have been integrated in the procured solution. Both Atos and IBM offered new burst buffer solutions. In the Spanish lot, IBM offered a new solution for integrating a tape back-end into a single name-space.

After the completion of the procurement, PPI4HPC performed an initial analysis of how well suppliers responded to the common technical goals. Based on a self-assessment by each of the public procurers, the response to goals related to supporting node power and energy measurements was strong to very strong, but for the goal of having job energy accounting integrated with workload managers, the response was in parts very weak. In the area of data management, the response had been strong to very strong, while in case of programming environment and

[21] https://prace-ri.eu/infrastructure-support/pre-commercial-procurement/

productivity it was typically strong, but sometimes also borderline. The response in the area of system and application monitoring tended to be disappointing.

The status of PPI4HPC can be summarised as follows: The PPI4HPC procurement procedure has been successfully completed resulting in four new HPC or storage systems that have been or are being deployed in France, Germany, Italy and Spain. The focus on innovations, which was stronger than usual for HPC-related procurements, helped to push new technologies. Running a joint procurement does involve non-negligible additional efforts, but results in additional benefits. The strong interaction resulted in improved procurement documents, significant gains from shared knowledge and experiences as well as a greater weight on the market. The efforts within the PPI4HPC project also helped to prepare the recent procurements conducted by the new EuroHPC Joint Undertaking. The focus of this short article is on the technical aspects of PPI4HPC. We started also to collect lessons learned from a legal perspective, which have been published in a white paper that is now available for others to learn on how to run joint procurements[22].

3.5 Big Data & Machine Learning at TGCC - First approach with KAPSDATA

Presenters: *Frederic Souques and Jean-Marc Ducos, CEA*

(Summary provided by the presenters)

CEA is a French Research and Technology Organisation involved in low-carbon energies, technologies for industry, fundamental research, defense and global security.

The computing complex of CEA is located in the Paris area at Bruyères-le-Châtel site and encompasses two parts: the TERA facility for internal use and the TGCC facility open to external users from research and industry.

The monitoring and optimisation of this computing complex is currently performed with two tools: Panorama[23], provided by Codra[24], for overall monitoring and optimisation, and a set of tools (ECF[25] and DHT[26]) provided by MIV-Soft[27], focusing on energy monitoring and optimisation.

In 2019, CEA decided to explore with KAPSDATA[28] to see how an innovative platform developed by this French company, combining Big Data, analysis tools and connected industrial devices together with AI, could help to improve understanding, availability, and energy efficiency of the HPC facilities. It was also decided to first conduct the tests on the TGCC part of the complex, as it was more open, and with more historical data available than the TERA part.

The tests focused on the optimisation of the cooling equipment (air and water cooling devices) of the TGCC. It was organised in four steps: (1) define the objectives of the optimisation; (2) gather data related to these objectives; (3) analyse this data; and (4) identify possible optimisations based on this analysis. It started in November 2019 and it is about to finish – as of October 2020 - a little later than expected, partially due to the impact of the COVID-19 situation.

The ultimate goal was to integrate the main parameters related to the cooling system of the TGCC in a global model: IT power consumption (uninterruptable and non-uninterruptable), external temperature, computer room temperature, cold water temperature, electrical power used by the chillers, speed of fans and power consumption of the cooling towers. Such modelling would make possible not only to optimise the efficiency of the overall cooling system but also to identify faulty operation modes.

This goal was clarified during step 1, as well as the characteristics of the cooling system of the TGCC and its methods of operation, in order for KAPSDATA to be able to properly understand the technical context. Then, during step 2, the required data was gathered, including 18 months of historical electrical and fluid temperature measurements.

[22] https://ppi4hpc.eu/news/ppi4hpc-whitepaper-%E2%80%9Clessons-learned-legal-aspects%E2%80%9D-just-published

[23] https://codra.net/en/offer/panorama-suite-software-platform/

[24] https://www.codra.net/en/

[25] Energie - Fluide - CO2: https://www.miv-soft.com/miv-logiciel-energetique-iso50001-sme.html

[26] Data Historian Technical: https://www.miv-soft.com/miv-logiciel-historian-alerte-maintenance.html

[27] https://www.miv-soft.com/

[28] http://kapsdata.com/

Step 3 started in mid-December 2019. It involved numerous meetings between CEA and KAPSDATA teams as well as a lot of work to improve the quality of information gathered during the previous steps. This proved to be mandatory for the proper and relevant usage of this information for building a model. The main issues identified and which had to be corrected, whenever possible, were:

- Errors in diagrams of cabling, making for example difficult to distinguish measurement related to uninterruptable and non-uninterruptable power consumption.
- Missing information in terms of temperature and flows.
- Too few measurement points at the level of the cooling towers.
- Lack of accuracy of time stamps of the measurements.
- A lot of transient phenomena (lasting a few seconds) not adequately captured because of the low frequency of measurements (one every 10 min).
- Missing measurements due to communication errors.

This took a long time, partially due to the COVID-19 crisis, and lasted until September 2020.

Since a number of collected data issues encountered during step 3 could not be solved, it was decided to revise the final objective to something less ambitious. Therefore, a new goal for the final step (step 4), still on-going at of writing this report, was agreed between CEA and KAPSDATA. This new goal was to produce a tool enabling advanced analysis of data produced by the components of the cooling system of the TGCC.

This tool is now expected to provide the following features:

- Time display of a set of data
- Temporal analysis of several sets of data per topic (production, distribution, cooling towers)
- Cross-analysis in 3D (for example: IT power usage, PUE and external temperature)
- Modelling of the functioning of the cooling towers

In conclusion, this exploratory work conducted by CEA together with KAPSDATA company was not able to fully reach the initial goal and unable to experiment with AI for the operational monitoring and optimisation of infrastructure equipment. The main reason was the lack of necessary data for this purpose. This has shown the necessity to organise an acquisition of data suitable for this objective in the future (more reliable, diverse and frequent measurements).

However, the work conducted with KAPSDATA was very valuable for CEA since the work done during step 3 opened up the possibility to detect and understand several issues related to the functioning of the cooling system of the TGCC. Moreover, the tool developed by KAPSDATA will help monitoring and comparing the evolutions of this cooling system (especially regarding the cooling towers and the heat pumps).

4. 11th European Workshop on HPC Infrastructures

Presenter: *Herbert Huber, LRZ*

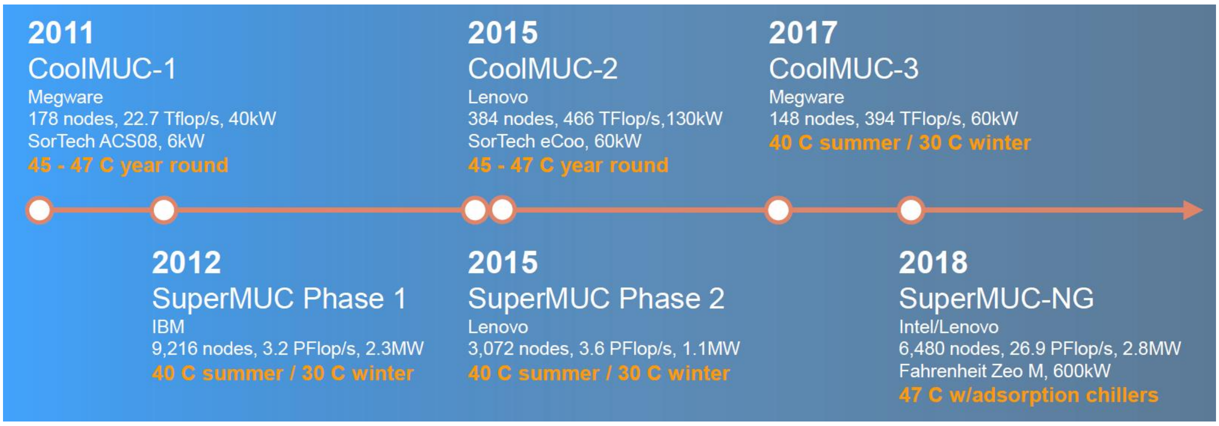
As mentioned in the introduction section the European Workshop on HPC Infrastructures was shifted from October 2020 to May 2021 due to COVID-19. It was initially scheduled for 18th to 20th May 2021 to be hosted at Leibniz Supercomputing Centre (LRZ) in Garching, Germany. This was however modified afterwards, again due to COVID-19, to an online event taking place from May 17th to 19th followed by the PRACE session on May 20th, 2021.

4.1 Leibniz Supercomputing Centre: A Pioneer of "Direct Warm Water Cooling"

LRZ's first direct warm water cooled system was installed in 2011. The system, referred to as CoolMUC-1 and built by Megware, consisted of 178 compute nodes each featuring two AMD Opteron 6128 HE (Magny-Cours) 8 core processors. CoolMUC-1 was the world's first AMD processor based direct warm water cooled system. The system was connected to a SorTech (now Fahrenheit GmbH[29]) adsorption chiller for the reuse of waste heat.

[29] https://fahrenheit.cool/en/

Figure 19 illustrates the history of the direct warm water cooled HPC systems, alongside with heat reuse technologies, deployed at LRZ starting from 2011.



2011
CoolMUC-1
Megware
178 nodes, 22.7 Tflop/s, 40kW
SorTech ACS08, 6kW
45 - 47 C year round

2012
SuperMUC Phase 1
IBM
9,216 nodes, 3.2 PFlop/s, 2.3MW
40 C summer / 30 C winter

2015
CoolMUC-2
Lenovo
384 nodes, 466 TFlop/s,130kW
SorTech eCoo, 60kW
45 - 47 C year round

2015
SuperMUC Phase 2
Lenovo
3,072 nodes, 3.6 PFlop/s, 1.1MW
40 C summer / 30 C winter

2017
CoolMUC-3
Megware
148 nodes, 394 TFlop/s, 60kW
40 C summer / 30 C winter

2018
SuperMUC-NG
Intel/Lenovo
6,480 nodes, 26.9 PFlop/s, 2.8MW
Fahrenheit Zeo M, 600kW
47 C w/adsorption chillers

Figure 19. LRZ: Pioneer in hot water cooling and reuse of waste heat

In 2017 LRZ installed the CoolMUC-3 system which was also built by Megware and was the first 100% direct warm water cooled system with a measured warm water cooling efficiency of 97%.

One of the requirements during the procurement of the current flagship system SuperMUC-NG Phase 1 at LRZ was a heat removal efficiency of 97% using direct warm water cooling or other cooling technologies that do not require the use of compressor cooling. Since the requested heat removal efficiency cannot be achieved with the direct warm water cooling technology design of SuperMUC-NG Phase 1 alone, Intel and Lenovo installed large adsorption chillers. These adsorption chillers allowed the reuse of the waste heat for the production of chilled water that is used for cooling the in-row coolers and rear door heat exchangers, which are still needed due to the presence of air cooled components such as power supplies, large Omni-Path switches, and some management nodes.

LRZ's current flagship system, SuperMUC-NG Phase 1, is an Intel Xeon Skylake processor based machine featuring 311,040 cores. The system has a peak performance of 26.9 PFLOPS and an HPL performance of 18.5 PFLOPS. It has a main memory capacity of 719 TB and 70 PB of disk space. The measured efficiency of the direct warm water cooling of the system is in the order of 75% - 85%. This is due to the fact that there is still a need for cooling air flowing over the hot processing components, since the air cooled power supplies are located on the back side of the nodes.

LRZ is currently in the procurement phase of SuperMUC-NG Phase 2.

5. Conclusions

The EWHPC 2020 online event, substituting the 11th EWHPC workshop due to COVID-19 pandemic, has been very successful and attracted experts from 23 different countries representing 40 different supercomputing sites.

This online event started with an overview on EuroHPC Joint Undertaking (JU) and was followed by updates regarding the three EuroHPC JU pre-exascale supercomputers and their supporting building infrastructures. In addition to technical specifications, the talks also shared certain insights on system procurement procedures. The presentation files of EWHPC 2020 Online event can be accessed via this link. [30]

The annually held EWHPC series, attended upon invitation only remain a unique platform offering specialists in the area of HPC/cloud data centre design and operations the possibility to discuss latest infrastructure trends and technologies as well as exchange on current challenges, lessons-learned, and best practices. **Figure 20** outlines the number of workshop attendees in a historical perspective.

[30] https://www.euhpcinfrastructureworkshop.org/?page_id=656.

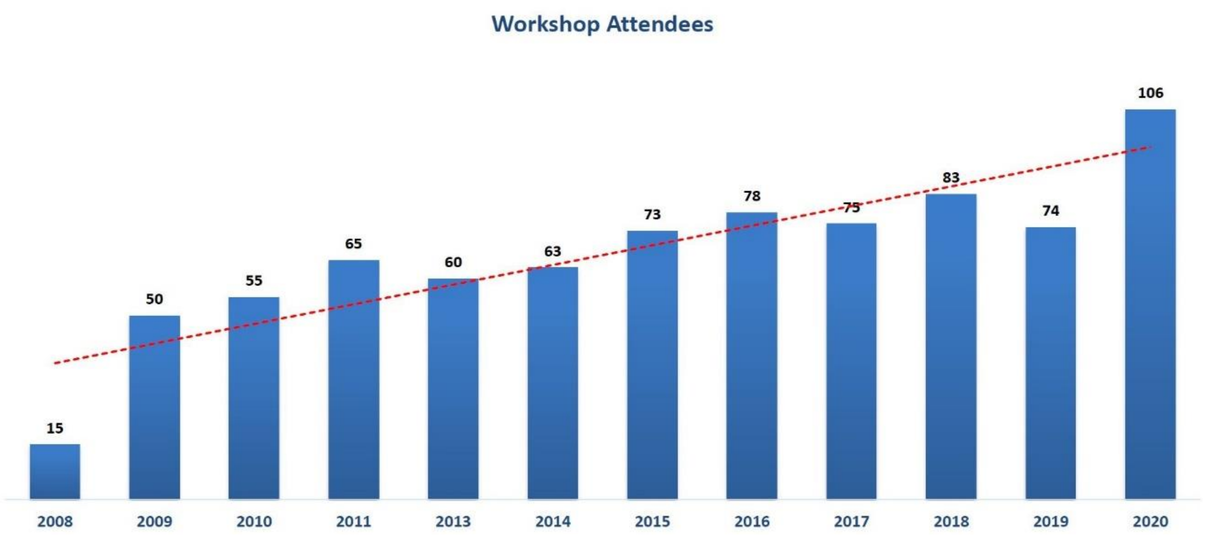


Figure 20. Number of Workshop on HPC Infrastructures attendees in a historical perspective

Table 6 provides a detailed view on EWHPC attendance starting from 2015. As indicated, there was an overall increase in the attendance during the EWHPC online event as compared to previous on-site meetings. As can be seen, this increase was mainly driven by the expansion in the number of participating PRACE sites, most probably due to the online nature of the event.

Workshop	**EWHPC #6**	**EWHPC #7**	**EWHPC #8**	**EWHPC #9**	**EWHPC #10**	**EWHPC 2020 Online Event**
Date	**2015 (May)**	**2016 (April)**	**2017 (April)**	**2018 (May)**	**2019 (May)**	**2020 (October)**
Location	Stockholm	Munich	Lugano	Bologna	Poznan	Online
Host	KTH	LRZ	CSCS	CINECA	PSNC	LRZ
Participants	**73**	**78**	**75**	**83**	**74**	**106**
Countries	17	17	17	15	16	23
EU countries	14	12	14	13	13	20
non EU countries	3	5	3	2	3	3
Sites	**30**	**28**	**27**	**34**	**30**	**40**
EU sites	26	22	23	25	23	33
PRACE sites	14	11	17	16	15	24
EU non PRACE sites	12	11	6	9	8	9
non EU sites	4	6	4	6	5	4
commercial DC	2	2	0	3	2	3

Table 6: tailed EWHPC attendance view starting from 2015

More information concerning the previous editions of the EWHPC series can be found online[31]. The next edition of the workshop, due to COVID-19 pandemic, will be an online event and will take place from May 18th to 19th, 2021 followed by the PRACE session on 20th of May, 2021.

Acknowledgements

The work described in this document was funded by the European Commission under the project PRACE-6IP, grant agreement number: 823767.

[31] https://prace-ri.eu/infrastructure-support/european-workshops-on-hpc-infrastructures/

List of Acronyms and Abbreviations

AI	Artificial Intelligence
BEO	Bull Energy Optimizer
BDPO	Bull Dynamic Performance Optimizer
BSC	Barcelona Supercomputing Centre
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3rd Party to GENCI)
CFD	Computational Fluid Dynamics
CINECA	CINECA Consorzio Interuniversitario, Italy
CPU	Central Processing Unit
DLC	Direct Liquid Cooling
EC	European Commission
EuroHPC JU	The European High-Performance Computing Joint Undertaking
EMAS	Eco-Management and Audit Scheme
EWHPC	European Workshops on HPC Infrastructures
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.
GENCI	Grand Equipement National de Calcul Intensif, France
FLOPS	Floating Point Operations Per Second
HEP	High Energy Physics
HPC	High Performance Computing
HPCG	High Performance Conjugate Gradient
HPDA	High Performance Data Analytics
HPL	High Performance LINPACK
ICT	Information and Communication Technologies
IT	Information Technology
LINPACK	Software library for Linear Algebra
LRZ	Leibniz Supercomputing Centre
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
MPI	Message Passing Interface
PCP	Pre-Commercial Procurement
PFLOPS	PetaFLOPS, i.e. 10^{15} FLOPS
PPI	Public Procurement of Innovative Solutions
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PSNC	Poznan Supercomputing and Networking Center

PUE	Power Usage Effectiveness
SME	Small and Medium-sized Enterprise
TCO	Total Cost Ownership
TFLOPS	TeraFLOPS, i.e. 10^{12} FLOPS
UPS	Uninterruptable Power Supply