E-Infrastructures
H2020-EINFRA-2016-2017

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

PRACE-5IP

PRACE Fifth Implementation Phase Project

Grant Agreement Number: EINFRA-730913

D5.5

Requirements of new user communities for the use of next generation computing systems evolving towards Exascale

Final

Version: 1.1
Author(s): Hayk Shoukourian, BADW-LRZ
Date: 18.04.2018

Project and Deliverable Information Sheet

PRACE Project	Project Ref. №: EINFRA-730913	
	Project Title: PRACE Fifth Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: D5.5	
	Deliverable Nature: Report	
	Dissemination Level: PU*	Contractual Date of Delivery: 30 / April / 2018
		Actual Date of Delivery: 27 / April / 2018
	EC Project Officer: Leonardo Flores Añover	

* **PU** – Public

Document Control Sheet

Document	Title: Requirements of new user communities for the use of next generation computing systems evolving towards Exascale	
	ID: D5.5	
	Version: 1.1	Status: Final
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2013	
	File(s): PRACE-5IP-D5.5.docx	
Authorship	Written by:	Hayk Shoukourian, BADW-LRZ
	Contributors:	Carlo Cavazzoni, CINECA Radosław Januszewski, PSNC Giannis Koutsou, CaSToRC Volker Weinberg, BADW-LRZ
	Reviewed by:	Vit Vondrak, IT4I Thomas Eickermann, FZJ
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	29/November/2017	Draft	Skeleton, the very first draft
0.2	23/January/2018	Draft	Compilation of survey results and distribution to T3 partners with an input request
0.3	12/March/2018	First complete draft	First complete draft sent to WP leader and co-leaders

0.4	23/March/2018	Semi-final complete version	Ready for final work package internal review
1.0	03/April/2018	Draft	Submission to PMO for review
1.1	18/April/2018	Final version	Includes comments from reviewers

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, User Prototyping Requirements

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] PRACE-4IP Deliverable D5.1 “Market and Technology Watch Report Year”, 2016

[2] PRACE-4IP Deliverable D5.2 “Market and Technology Watch Report Year 2. Final summary of results gathered”, 2017

[3] [Online]. Available: http://www.prace-ri.eu/prace-4ip/
[4] [Online]. Available: http://www.prace-project.eu
[5] [Online]. Available: http://www.prace-ri.eu/prace-pp/
[6] [Online]. Available: http://www.prace-ri.eu/prace-1ip/
[7] [Online]. Available: http://www.prace-ri.eu/prace-2ip/
[8] [Online]. Available: http://www.prace-ri.eu/prace-3ip/
[9] [Online]. Available: http://montblanc-project.eu/
[10] [Online]. Available: http://www.deep-project.eu/deep-project/EN/Home/home_node.html
[11] [Online]. Available: http://hpc.desy.de/qpace/
[12] PRACE-4IP Deliverable D5.5; 2016
[13] PRACE-4IP Deliverable D5.6 “Best Practices for Prototype Planning and Evaluation”, 2017
[14] [Online]. Available: https://bioexcel.eu/
[15] [Online]. Available: https://www.cecam.org/
[16] [Online]. Available: www.compbiomed.eu
[17] https://www.cmcc.it/projects/esiwace-centre-of-excellence-in-simulation-of-weather-and-climate-in-europe
[18] [Online]. Available: http://www.nanogune.eu/
[19] [Online]. Available: https://www.e-cam2020.eu/
[20] [Online]. Available: http://www.eocoe.eu/
[21] [Online]. Available: https://www.nomad-coe.eu/
[22] [Online]. Available: https://www.cineca.it/
[23] [Online]. Available: https://www.cyi.ac.cy/index.php/castorc/about-the-center/castorc-center-overview.html
[24] [Online]. Available: https://www.csc.fi/
[25] [Online]. Available: https://www.lrz.de/english/
[26] [Online]. Available: https://grnet.gr/en/
[27] [Online]. Available: https://www.edu.unideb.hu/
[28] [Online]. Available: http://www.man.poznan.pl/online/en/
[29] [Online]. Available: https://pg.edu.pl/welcome?p_l_id=52858455&p_l_id=2601414&p_v_l_s_g_id=0&p_v_l_s_g_id=0&
[30] [Online]. Available: https://www.hartree.stfc.ac.uk/Pages/home.aspx
[31] [Online]. Available: https://www.top500.org/

List of Acronyms and Abbreviations

AT	Advanced Technology
CISC	Complex Instruction Set Computer
CoE	Center of Excellence
CPU	Central Processing Unit
DDR	Double Data Rate

FLOPS	Floating Point Operations Per Second
FPGA	Field-programmable gate array
GB	Giga (= 230 ~ 109) Bytes (= 8 bits), also Gbyte
Gb/s	Giga (= 109) bits per second, also Gbit/s
GB/s	Giga (= 109) Bytes (= 8 bits) per second, also Gbyte/s
GDDR	Graphics DDR
GHz	Giga (= 109) Hertz, frequency =109 periods or clock cycles per second
GPU	Graphic Processing Unit
HBM	High Bandwidth Memory
HDD	Hard Disk Drive
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
I/O	Input/Output
ISA	Instruction Set Architecture
MB	Mega (= 220 ~ 106) Bytes (= 8 bits), also Mbyte
MB/s	Mega (= 106) Bytes (= 8 bits) per second, also Mbyte/s
MIC	Many Integrated Core
MPI	Message Passing Interface
NFS	Network File System
NVM	Non-Volatile Memory
NVMe	NVM Express
NVRAM	Non-Volatile Random-Access Memory
OpenCL	Open Computing Language
OS	Operating System
PAPI	Performance Application Programming Interface
PTX	Parallel Thread Execution
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
RAID	Redundant Array of Independent Disks, originally Redundant Array of Inexpensive Disks
RISC	Reduced Instruction Set Computer
SATA	Serial AT Attachment
SIMD	Single Instruction, Multiple Data
SSD	Solid-State Drive
TB	Tera (= 240 ~ 1012) Bytes (= 8 bits), also Tbyte
TB/s	Tera (= 240 ~ 1012) Bytes (= 8 bits) per second, also Tbyte/s
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
UEABS	Unified European Applications Benchmark Suite
VM	Virtual Machine

List of Project Partner Acronyms

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

BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion, Spain
CaSToRC	Computation-based Science and Technology Research Center, Cyprus
CCSAS	Computing Centre of the Slovak Academy of Sciences, Slovakia
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3 rd Party to GENCI)
CESGA	Fundacion Publica Gallega Centro Tecnológico de Supercomputación de Galicia, Spain, (3rd Party to BSC)
CINECA	CINECA Consorzio Interuniversitario, Italy
CINES	Centre Informatique National de l'Enseignement Supérieur, France (3 rd Party to GENCI)
CNRS	Centre National de la Recherche Scientifique, France (3 rd Party to GENCI)
CSC	CSC Scientific Computing Ltd., Finland
CSIC	Spanish Council for Scientific Research (3rd Party to BSC)
CYFRONET	Academic Computing Centre CYFRONET AGH, Poland (3rd party to PNSC)
EPCC	EPCC at The University of Edinburgh, UK
ETHZurich (CSCS)	Eidgenössische Technische Hochschule Zürich – CSCS, Switzerland
FIS	FACULTY OF INFORMATION STUDIES, Slovenia (3rd Party to ULFME)
GCS	Gauss Centre for Supercomputing e.V.
GENCI	Grand Equipement National de Calcul Intensiv, France
GRNET	Greek Research and Technology Network, Greece
INRIA	Institut National de Recherche en Informatique et Automatique, France (3 rd Party to GENCI)
IST	Instituto Superior Técnico, Portugal (3rd Party to UC-LCA)
IUCC	INTER UNIVERSITY COMPUTATION CENTRE, Israel
JKU	Institut fuer Graphische und Parallele Datenverarbeitung der Johannes Kepler Universitaet Linz, Austria
JUELICH	Forschungszentrum Juelich GmbH, Germany
KTH	Royal Institute of Technology, Sweden (3 rd Party to SNIC)
LiU	Linkoping University, Sweden (3 rd Party to SNIC)
NCSA	NATIONAL CENTRE FOR SUPERCOMPUTING APPLICATIONS, Bulgaria
NIIF	National Information Infrastructure Development Institute, Hungary
NTNU	The Norwegian University of Science and Technology, Norway (3rd Party to SIGMA)
NUI-Galway	National University of Ireland Galway, Ireland
PRACE	Partnership for Advanced Computing in Europe aisbl, Belgium
PSNC	Poznan Supercomputing and Networking Center, Poland
RISCSW	RISC Software GmbH
RZG	Max Planck Gesellschaft zur Förderung der Wissenschaften e.V., Germany (3 rd Party to GCS)
SIGMA2	UNINETT Sigma2 AS, Norway
SNIC	Swedish National Infrastructure for Computing (within the Swedish Science Council), Sweden
STFC	Science and Technology Facilities Council, UK (3rd Party to EPSRC)

SURFsara	Dutch national high-performance computing and e-Science support center, part of the SURF cooperative, Netherlands
UC-LCA	Universidade de Coimbra, Labotatório de Computação Avançada, Portugal
UCPH	Københavns Universitet, Denmark
UHEM	Istanbul Technical University, Ayazaga Campus, Turkey
UiO	University of Oslo, Norway (3rd Party to SIGMA)
ULFME	UNIVERZA V LJUBLJANI, Slovenia
UmU	Umea University, Sweden (3 rd Party to SNIC)
UnivEvora	Universidade de Évora, Portugal (3rd Party to UC-LCA)
UPC	Universitat Politècnica de Catalunya, Spain (3rd Party to BSC)
UPM/CeSViMa	Madrid Supercomputing and Visualization Center, Spain (3rd Party to BSC)
USTUTT-HLRS	Universitaet Stuttgart – HLRS, Germany (3rd Party to GCS)
VSB-TUO	VYSOKA SKOLA BANSKA - TECHNICKA UNIVERZITA OSTRAVA, Czech Republic
WCNS	Politechnika Wroclawska, Poland (3rd party to PNSC)

Executive Summary

High Performance Computing (HPC) is experiencing vast amount of changes in the road towards Exascale computing capability. These changes stretch throughout different levels: from technology and architectures to use cases. In order to attain the best performing HPC system, it is imperative that the underlying technology and architecture match the requirements of the current and emerging applications.

This document aims to provide an overview of these requirements by assessing the needs of user communities and of HPC centres in terms of technologies and architectures for next generation HPC systems evolving towards Exascale. For this purpose, surveys have been conducted among recently started Centres of Excellences (CoEs) in Europe for collecting the requirements from HPC user communities. A different survey has been distributed to all PRACE Tier-0/Tier-1 HPC sites to understand how these requirements differ from the current state of the art, to determine the requirements of HPC centres, and possibly motivate related prototyping efforts.

This deliverable summarizes the results of the two surveys. The most important points to note are indicated in the list below:

- a need for prototype systems involving heterogeneous system architectures that include new kinds of memory and parallel I/O file systems is seen by the user communities as well as by PRACE Tier-0/Tier-1 HPC centres;
- Graphic Processing Units (GPUs) are the most appealing accelerator systems for the user communities – a requirement which is already fulfilled by 45% of PRACE Tier-0/Tier-1 HPC sites;
- a shift from conventional x86 based processing technologies (which is currently dominating at PRACE HPC sites) to alternatives such are ARM, IBM Power Architecture, PTX (Parallel Thread Execution) processing technologies is foreseen for the surveyed HPC sites;
- containers, which are instances of an Operating System (OS) level virtualization, are getting more appealing due to their higher efficiency as compared to the full, hardware-level, virtualization;
- growing power density for the required heterogeneous compute nodes further motivates the need for the adoption of water cooling technologies.

1 Introduction

Throughout generations, the processors have been primarily improved with the help of smaller and faster circuitry, which brought continuously increasing processing speeds allowing various complex computations to be solved at a faster clip without major changes in architectures and in applications. However, for a number of years, the clock rate of processors has been stable, mainly because of the limit of acceptable power consumption (in terms of cost and heat dissipation). Therefore, major chip manufacturers are transitioning from multi-core processors (typically involving a small amount of independent processing cores) to many-core processors, possibly with hardware accelerators (such as, GPUs, FPGAs, etc.) and to more complex memory hierarchy. Similarly, the architecture of supercomputers is getting more complex with a large number of possibly heterogeneous nodes. A survey of trends in terms of technologies and architectures can be found in the deliverables D5.1 [1] and D5.2 [2] produced by PRACE-4IP [3] WP5.

In this context, optimizing and mapping computationally intensive tasks to suitable processing resources is needed for making the overall computations more time and energy-efficient. A similar effort is needed for I/O intensive tasks.

Therefore, there is a challenge for HPC application developers, requiring moving away from the currently used application programming paradigms. For example, the majority of currently existing large-scale HPC applications rely only on the MPI communication protocol, which implies a distribution of computational problems to individual compute units (cores) and a large number of communications between cores. The ever-increasing number of computational resources, foreseen with next generation HPC systems, will make the management of this type of communication traffic even more complex and error prone. This means, for example, using a multi-level parallelism both at the node level (shared-memory) and across nodes (message passing) in addition to vectorization or SIMD parallelism at the code level.

All these aspects make the co-design activities even more important, bringing together application developers and hardware manufacturers to understand and design complex software and hardware architectures in the most efficient way,.

HPC prototyping allows the evaluation of new concepts and technologies that aim to address the functionality shortages present in the existing state of the art solutions. It was one of the main activities for various former PRACE [4] projects, such are PRACE-PP [5], PRACE-1IP [6], PRACE-2IP [7], PRACE-3IP [8] - activities, which were later moved to separate, EC funded, technology projects such as Mont-Blanc [9], DEEP/DEEP-ER [10], and QPACE [11].

The previous PRACE-4IP [3] WP5 efforts looked at the requirements of HPC application developers and supercomputing centres for a typical hardware prototyping project [12] as well as provided a comprehensive overview on the individual phases of HPC prototyping project [13].

This document covers the next step, by providing an overview on the requirements (from the HPC user community as well as from HPC centre perspective), in terms of technologies and architectures, for the use of next generation computing systems evolving towards Exascale. This document also provides a synopsis on the foreseen prototyping activities within PRACE Tier-0 and Tier-1 sites and draws a comparison between user expectations and the planned prototyping

projects at PRACE HPC sites. Additionally, it provides a short outline of the current state of the art architectures/technologies that help in understanding how the mentioned requirements arise from the current state of the art and possibly motivate related prototyping efforts.

The rest of document is organised as follows: Section 2 lists the entities that completed the surveys on which this deliverable is based; Section 3 summarises the expectations of current user communities; Section 4 outlines the foreseen activities of PRACE partners in prototyping projects. Section 5 presents the state of the art at PRACE Tier-0/Tier-1 sites in reference to user requirements. Section 6 provides outlook, delineates future work, and concludes this report. The surveys and raw data of the obtained results have been uploaded to the PRACE repository, and can be accessed via https://repository.prace-ri.eu/git/hayk.shoukourian/5IPT3.git link (access restricted to PRACE-IP partners).

2 Surveys

The analyses presented in this document are mostly based on the results of online surveys that were distributed among PRACE Tier-0 and Tier-1 sites and Centres of Excellence (CoEs). Three surveys were created: two surveys intended for CoEs (a very short survey with 3 questions, and an optional and longer one with 19, mainly multiple-choice, questions)[1], and one, with overall 53 questions, for PRACE Tier-0 and Tier-1 sites. The CoE related surveys were prepared in cooperation with the PRACE-5IP WP7 application-focused work package, which is among other activities in charge of code enabling activities, publication of Best Practice Guides and the development of the Unified European Applications Benchmark Suite (UEABS). All three surveys were distributed with the help of an open-source survey tool, LimeSurvey hosted at BADW-LRZ.

Eleven participants from the following CoEs participated in the short survey:

CoE	Coordinating Country
BioExcel [14]	Sweden
CECAM [15]	Switzerland
CompBioMed [16]	UK
ESiWACE () [17]	Germany
CIC nanoGUNE [18]	Spain
E-CAM [19]	Switzerland
EoCoE [20]	France
NOMAD [21]	Germany

Table 1: List of CoEs that participated in the short survey.

[1] Two separate surveys were created since there was less incentive from CoEs in time investment for survey completion

These CoEs cover a wide range of HPC application domains and developers, and therefore provide a good view on the requirements of the diverse European HPC user community.

Representatives of the following CoEs participated in the long survey:

CoE	Coordinating Country
Centre of Excellence in Simulation of Weather and Climate in Europe (ESiWACE)	Germany
E-CAM	Switzerland

Table 2: List of CoEs that participated in the long survey.

The following PRACE Tier-0/Tier-1 sites participated in the survey:

PRACE Tier-0/Tier-1 site	Name of the flagship system	Country
CINECA [22]	MARCONI	Italy
Computation-based Science and Technology Research Center (CaSToRC), The Cyprus Institute [23]	Cy-Tera	Cyprus
CSC - IT Center for Science Ltd. [24]	Sisu	Finland
Leibniz Supercomputing Centre of the Bavarian Academy of Sciencies (BAdW-LRZ) [25]	SuperMUC Phase 2	Germany
Greek Research and Technology Network (GRNET) [26]	ARIS	Greece
University of Debrecen [27]	VGGD	Hungary
Poznan Supercomputing and Networking Center (PSNC) [28]	Eagle / Hetman	Poland
Gdansk University of Technology [29]	Tryton	Poland
The Hartree Centre [30]	Scafell Pike	UK

Table 3: List of PRACE Tier-0/Tier-1 sites that participated in the survey.

Most of these HPC sites were involved in the previously mentioned PRACE prototyping projects, deploy prototype systems on a regular basis, and thus bring in significant expertise in terms of HPC prototyping [12] [13].

3 Expectations of user communities from an HPC prototyping project

Most user communities are not much involved in activities related to prototypes. The main reason is that these activities are mostly about assessing and validating a technology that is not able to perform useful work yet or contribute to their projects. For example, when dealing with a prototype, a user is faced with a lot of tedious work (compared to a production system) since a user has to complete trivial tasks, like setting up the application environment or managing executions if the compilation of the application is at the end successful.

But now, with the current trends in HPC technologies and architectures (as reported in the introduction) more and more user communities understand the benefit of taking part in a co-design cycle and therefore face the challenge of porting their use cases on prototypes. In fact, with specialized hardware and heterogeneous architectures applications need to be re-factored and adapted. If this effort is done when a prototype is available, then the applications will be ready when the production system is deployed.

In anticipation of this trend the European CoEs for computing applications have been established through targeted calls. CoEs are tasked to develop the next generation of community codes, capable of Exascale, in anticipation of future computing technologies. PRACE-4IP WP5 investigated the readiness of user communities in engaging in prototyping activities towards Exascale and in this deliverable targets CoE researchers and developers.

In this chapter, we analyze the most urgent requirements from the CoEs in terms of technologies and architectures and outline the activities that can be done at the level of PRACE for covering these requirements.

3.1 [COE] Q1 - Please prioritize your requirements for next generation HPC systems (in terms of hardware and system software perspective) from 1 to 12, with 12 being as "the most required", and 1 being "the least required" (specify 0, if irrelevant)

The intention of this question was to obtain some quantification of user requirements in terms of system hardware and system software for next generation HPC systems.

Twelve main items were selected. The bar chart in Figure 1 illustrates the average scoring for the mentioned 12 items, namely (items sorted according to the received averaged scores, from highest to lowest):

- I/O performance for the file system (average score: 8.72)
- GPU accelerators (average score: 8.27)
- I/O performance for the interconnect network (average score: 7.45)
- Memory size (average score: 7.36)
- Programmability (average score: 7)

- Memory bandwidth (average score: 6.81)
- Support for long term data archiving (average score: 6.81)
- Persistent storage on the node (average score: 6.72)
- Increased memory/core ratio (average score: 6.45)
- Performance monitoring tools (average score: 6.27)
- Intel Xeon Phi (average score: 4.54)
- FPGA accelerators (average score: 4)

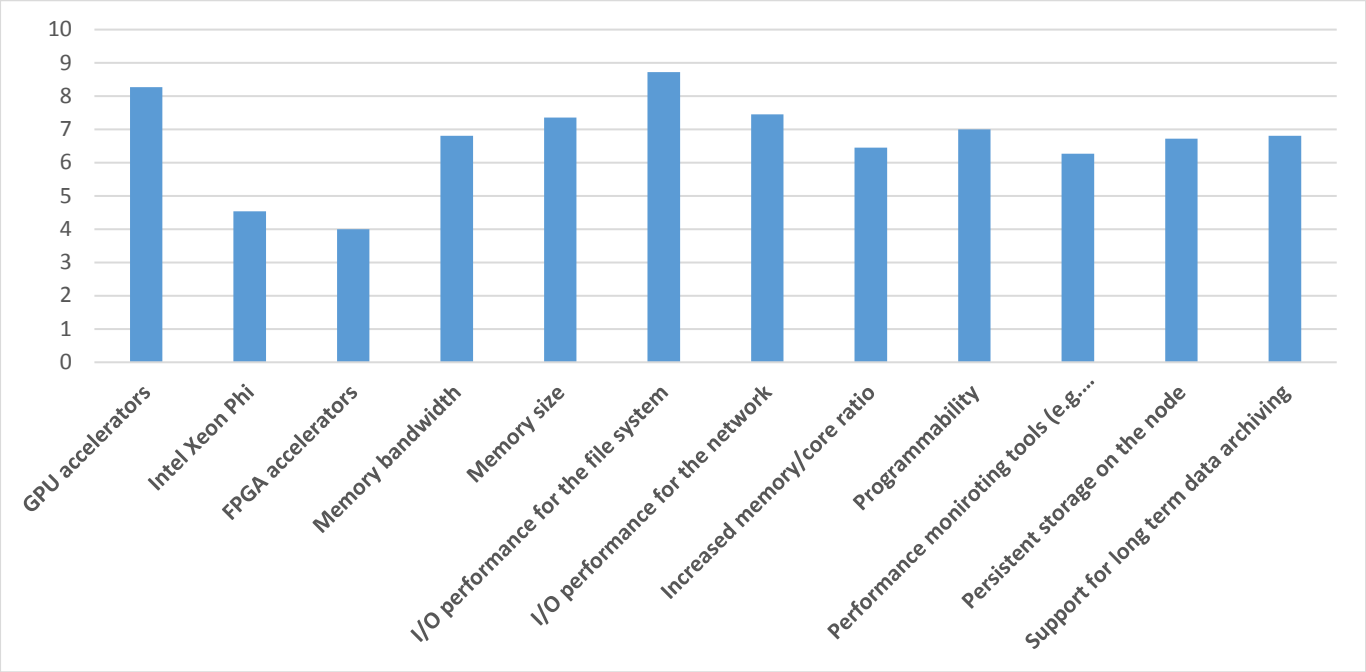

Figure 1: Average scoring of requirements for next generation HPC systems (from CoEs).

Analysis

The answers from the users reveal a prioritization in the requirement of GPUs and higher I/O for the file system. There are several reasons that GPUs (that work as accelerators alongside with the CPUs of the compute nodes to accelerate certain application regions requiring a large amount of numerical operations) are becoming more broadly used and adopted in HPC. First, due to the slowdown in Moore's law, manufacturers need to find new ways for delivering the required, ever-increasing, computational power more efficiently – that is one of the reasons that the current TOP500 [31] list includes more than 100 accelerated systems. The survey results clearly show that users want to test and obtain more experience in using accelerated systems in order to benefit from the massive parallelism offered by an accelerated system. As can be seen, Intel's Xeon Phi's have much less appeal with respect to GPUs, presumably since most users by the time of these surveys had anticipated Intel's reluctance in further development of this architecture. FPGAs appear to be

the least required accelerator architecture, most probably due to the lack of options in programming models and tools.

The survey has also revealed a high prioritization in I/O to the filesystem. A possible explanation is that as the HPC community approaches Exascale, it is anticipated that the analysis of simulation outputs may become the bottleneck. In some fields such as climate modelling and weather forecasting, ingestion of data into the simulation will become possible during Exascale and will require high bandwidth to storage. Furthermore, the CoEs with applications in life sciences are expected to be more data-driven, and therefore the performance of their applications is more susceptible to storage I/O.

Additionally, responses to memory related questions have scored a high interest, reflecting the fact that most user community applications, which are memory bound, struggle the most in achieving good performance on architectures with a high FLOPS / Bytes ratio. The survey reveals that users are keen to evaluate possible solutions that could mitigate that issue. Programmability and tools for analysing code and performance are of high interest as well - this is probably connected to the fact that more complex architectures require more insights to be fully understood and exploited.

3.1.1 Please provide other critical requirements (if any) not mentioned in above question with corresponding ranking (from the "least important" to "most important")

Out of 11 survey participants only 4 indicated 5 additional (to the above mentioned 12) requirements that should be considered in a prototyping project. The following list summarizes these requirements:

1. Early insights into future technologies: “To have a hardware roadmap soon enough to prepare applications” (1 answer);
2. Network: “Network latency is critical for massive particle simulations (being ranked as “most required”)” (2 answer);
3. Data analysis: “NOMAD needs fast random access file I/O. Hadoop-like solutions will be necessary” (1 answer);
4. Compilers: “Mature optimising compilers including for Fortran 2003 & 2008” (1 answer).

Analysis

These answers were spontaneous (were not chosen from a predefined list) and therefore reflect specific needs of certain use cases. Four out of the five responses are covered by the previous question, namely “I/O performance for the file system” (answer 3.), “I/O performance for the network” (the two answers in 2.), and “Programmability” (answer in 4.).

3.2 [COE] Q2 - Please indicate which technologies it would be useful to investigate with prototypes in the next 2 years?

The following suggestions were obtained from the survey participants:

- Fortran 2015 on MIC

- Porting existing OpenCL codes to FPGA
- Intel MIC
- Nvidia GPUs (with increased capabilities of file I/O)
- Low-power processors (e.g. ARM, FPGA)
- NVRAM

Analysis

The majority of the obtained suggestions relate to new architecture and code design paradigms, indicating that the users are aware of foreseen modifications and ready to diverge from well-established architectures (such are mainstream CISC and RISC processors) and programming models (such as MPI).

4 Technologies to be assessed with future prototype systems according to PRACE Tier-0/Tier-1 sites

In this chapter, we analyse the answers from the HPC centres about technologies that they consider as important for the future and that, as such, should be tested if not already used in production systems. In general, the answers depend on the technology already deployed in the centre, with a clear tendency to be interested in alternatives, unless the technology is indeed new on the market.

4.1 ISA of processing units

Figure 2 shows the Instruction Set Architectures (ISAs) of the processing technologies that, according to the PRACE HPC sites, should be tested with future HPC prototype systems.

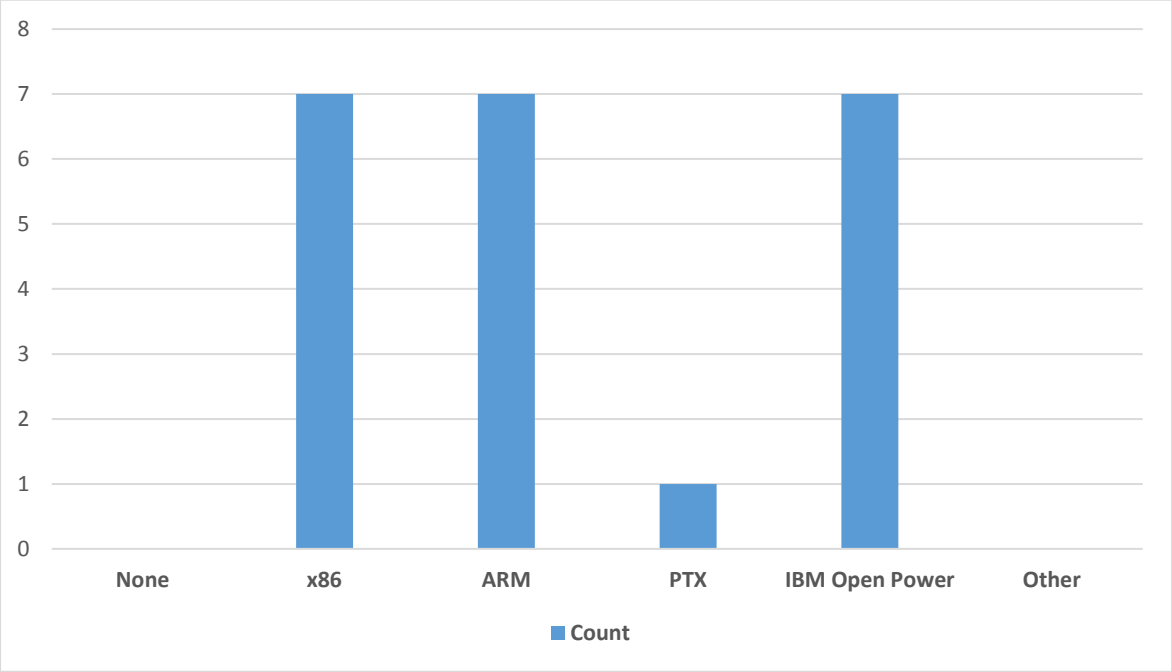

Figure 2: Answers from PRACE Tier-0/Tier-1 HPC sites regarding the instruction set architectures, which the processing technologies of future prototype systems should be compatible with.

Analysis

This bar chart indicates two main alternatives to the mainstream x86 ISA: ARM and IBM Open Power. According to the long user surveys that were also distributed to CoEs, users require that the compute nodes are mainly compatible with x86 and IBM Open Power instruction set architectures - a requirement that (as will be shown in Section 5.1) is not currently fulfilled by PRACE Tier-0/Tier-1 HPC sites. This additionally indicates the need for testing processing units different from mainstream with future prototype systems.

4.2 Accelerators

PRACE HPC sites were asked to indicate which accelerator systems should be investigated. Figure 3 summarizes this survey results.

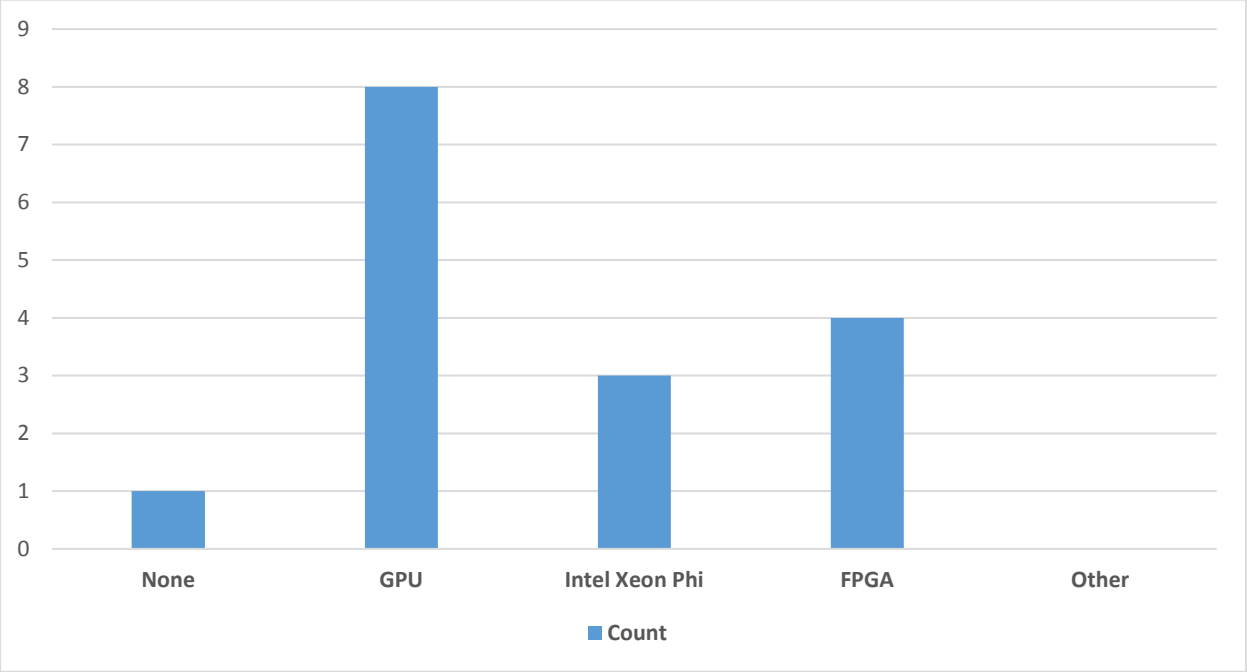

Figure 3: Answers from PRACE Tier-0/Tier-1 HPC sites regarding the accelerator technologies that should be assessed in future.

Analysis

The strongest interest is clearly for GPU. Several reasons can be found for such interest:

- the interest, prevalence, and adoption from/by the user communities;
- the performance potential in application domains, including new ones such are machine learning and data mining;
- the significant number of applications already ported HPC applications;
- energy-efficiency.

The last point can be attributed to FPGAs as well, which explains its relative high rate being the second in the list.

The low interest in Intel Xeon Phi processors is likely related to the recent change of roadmap of Intel announcing the end of the Xeon Phi line.

It is worth mentioning that one site answered “None” as it believes that accelerators are not useful for a typical user, since most of the existing HPC applications would require significant modifications for the porting.

4.3 Storage technologies

Participants were requested to indicate which storage technologies they are interested in, and to justify their choices.

Six HPC sites have shown interest in testing new storage technologies. Below we report the list of technologies entered by the sites and the motivation they gave:

- LUSTRE alternatives ("We are not completely happy with LUSTRE and having alternatives is a good thing");
- non-volatile memory technology (such as 3D XPoint - "Could be a replacement for applications being not latency bound (e.g. machine learning or big data applications)"; "test new possibilities");
- partitionable storage technology ("We are satisfied with LUSTRE FS, but there is on demand portioning missing"); storage class memory ("Most interesting");
- SSD ("test new possibilities").

Analysis

Apart from the first one, the majority of the listed technologies and the motivations reflect the need to have a better understanding of new memory devices that are being introduced in the market, and the impact they may have in the exploitation of HPC facilities. The first and the third answers are contradicting each other with one HPC site being not satisfied with LUSTRE. Nonetheless, both agree that having alternatives is good. It can be concluded that there is a need to consider a prototype system with novel kinds of file systems that can leverage new device memory.

4.4 Cooling and heat reuse technologies

Participants were asked to specify their interest among listed seven different cooling technologies, or indicate any other cooling solution not listed in the questionnaire. All sites indicated some interest in the listed technologies, and no other technology has been mentioned. Figure 4 presents the responses concerning cooling technologies.

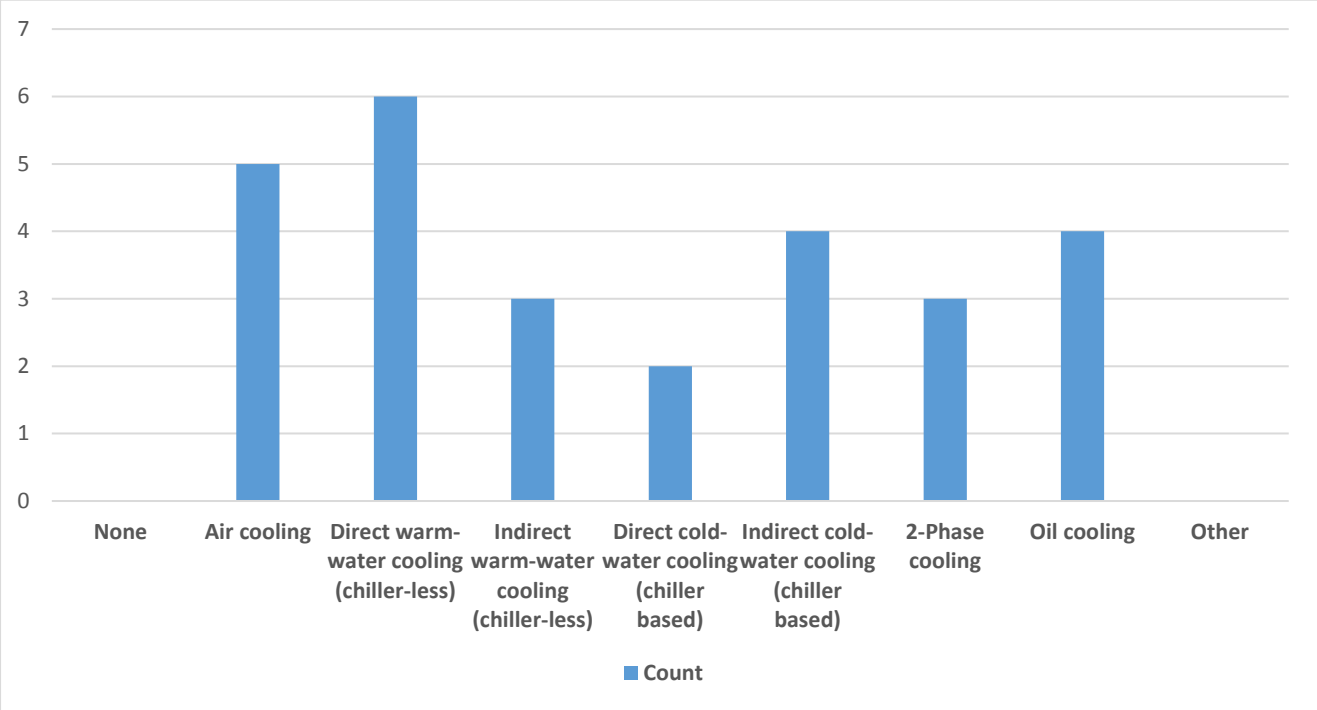

Figure 4: Cooling technologies to be assessed in future.

Participants were requested to justify their choices by answering the question "Why should (or should not) the above selection for cooling technologies be tested?"

Table 4 summarizes the motivations of PRACE Tier-0/Tier-1 HPC sites for using/testing certain cooling technology.

Cooling technology	Motivation for using/testing
Air cooling	current data centre cooling infrastructure setup; exploring new possibilities; most interesting for testing
Direct-warm water cooling (chiller-less)	runtime costs reduction; cooling efficiency; need of direct liquid cooling on chip stipulated by future processors; promising for heat reuse; exploring new possibilities; most interesting for testing
Indirect-warm water cooling (chiller-less)	promising for heat reuse; exploring new possibilities; most interesting for testing
Direct-cold water cooling (chiller based)	current data centre cooling infrastructure setup; promising for heat reuse; exploring new possibilities; most interesting for testing
Indirect cold-water cooling (chiller based)	current data centre cooling infrastructure setup; explore new possibilities; most interesting for testing
2-Phase cooling	runtime costs reduction; exploring new possibilities
Oil cooling	promising for heat reuse; exploring new possibilities; most interesting for testing

Table 4: Motivation for usage of certain cooling technology

The participants were further asked to indicate if the choice of a new cooling solution would have some impact on the current building infrastructure. Figure 5 presents the obtained results.

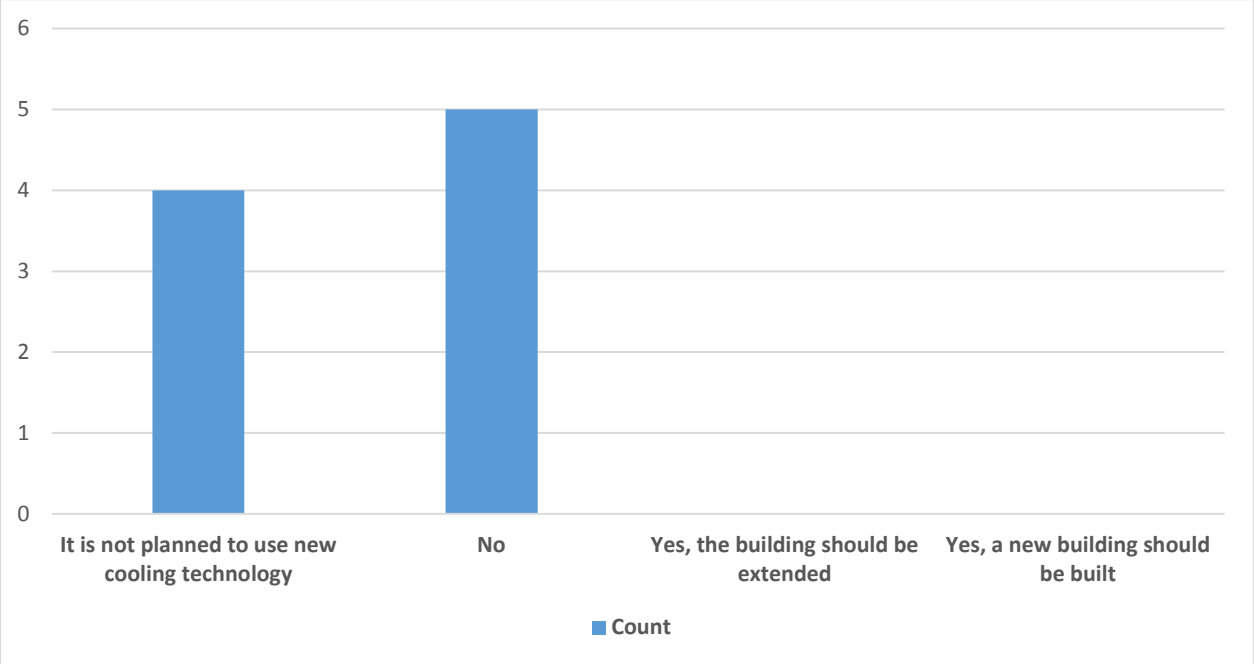

Figure 5: Answers to the question "Does the use of new cooling technology imply a change in the current building infrastructure (i.e. will require constructing a new building or extending the existing one)?".

Analysis

The fact that all sites have answered, testifies the interest in cooling technologies, but in contrast to previous questions regarding storage technologies, this section explicitly listed the technologies. Thus, it cannot be taken as a measure of the fact that cooling is more interesting than the former one. The higher number of answers probably can be attributed to the fact that for cooling related questions there was a pre-defined list of answers to select from, whereas for storage questions the participants were asked to specify their preferred technology.

The majority of received answers expressed interest in direct warm-water cooling technology, being felt as the most efficient technique for cooling future power hungry sockets, allowing for higher node density and heat reuse.

On the other hand the second most popular cooling technology indicated by the participants (Figure 4), is the plain air cooling, on the opposite end in terms of efficiency with respect to the most selected solution. This shows a polarization between those that would like to innovate, and those that would like to have a more standard setup not requiring new skills and competence. This selection might also be related to the power consumption of the hosted flagship systems. According to the survey results, the average power consumption of the main machine of the former group (i.e. the group of PRACE Tier-0/Tier-1 HPC sites expressing interest in direct warm-water cooling) is 1100 kW during normal operational modes, whereas the average power consumption of the main

machine for the latter group is only 120 kW. Finally, some interest is expressed also in oil cooling technology.

Interestingly, most of the cooling technologies do not require a change in the current building infrastructure. This will, in principle, allow to design various prototype systems based on innovative cooling technologies without introduction of major modifications and costs to the building and the data centre's facility.

5 State of the art at PRACE Tier-0/Tier-1 sites in reference to user requirements

This section aims to assess how far the needs and requirements of the user communities (CoEs) are from the current state of the art present at PRACE Tier-0/Tier-1 sites. The section considers the flagship systems of the surveyed supercomputing sites in order to provide a complete view on the current state of the art deployments of large-scale production systems.

5.1 [HPC sites] Q1 – Which of the following instruction set architecture(s) are the compute nodes compatible with?

Six options for possible answers were specified. Figure 6 outlines the survey results.

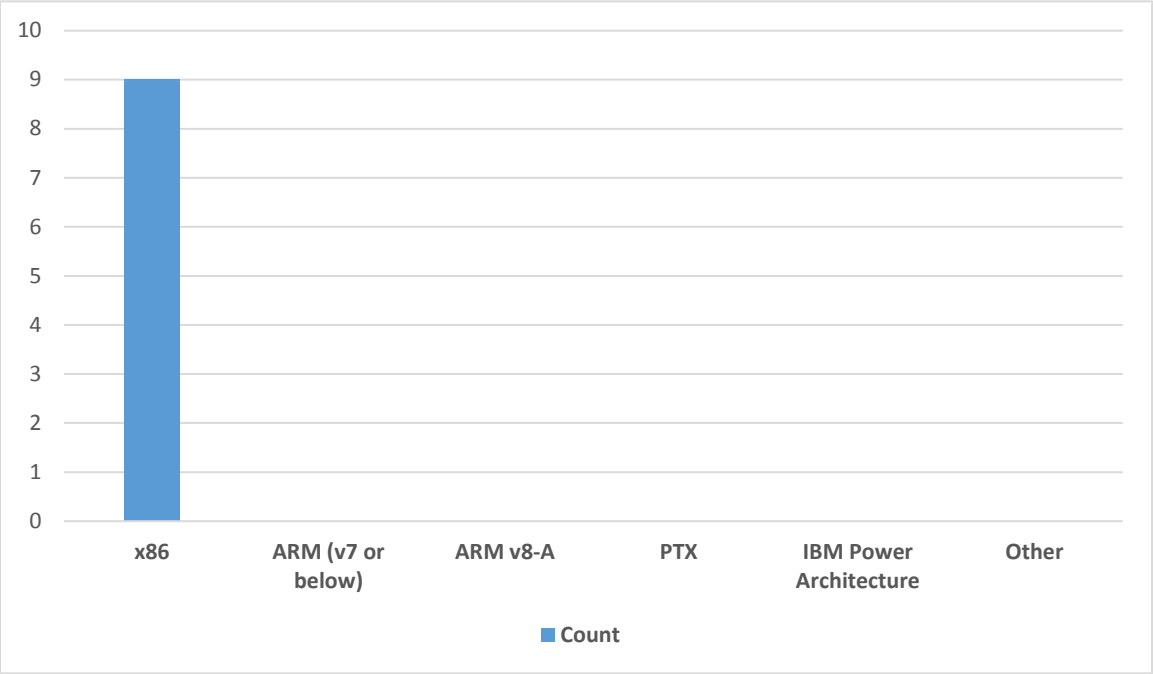

Figure 6: Instruction set architectures supported by PRACE Tier-0/Tier-1 sites.

Analysis

The answers collected from PRACE HPC Tier-0/Tier-1 sites suggest that the x86 architecture is the only one present on the market. These nine responses reflect the general state of the market represented by the TOP500 [31] list – the market is dominated by single technology to the point where other than x86 technologies for CPUs may be considered as peculiar. The reason for this situation is the better price to performance ratio of x86 compared to other architectures in the past. This situation may be changing: we are observing a gradual emergence of different architectures that are receiving attention from both HPC sites and user communities with high expectations regarding alternative technologies (see Section 4.1).

5.2 [HPC sites] Q2 – Accelerator type

Figure 7 shows the responses received regarding the accelerator types available at PRACE Tier-0/Tier-1 sites.

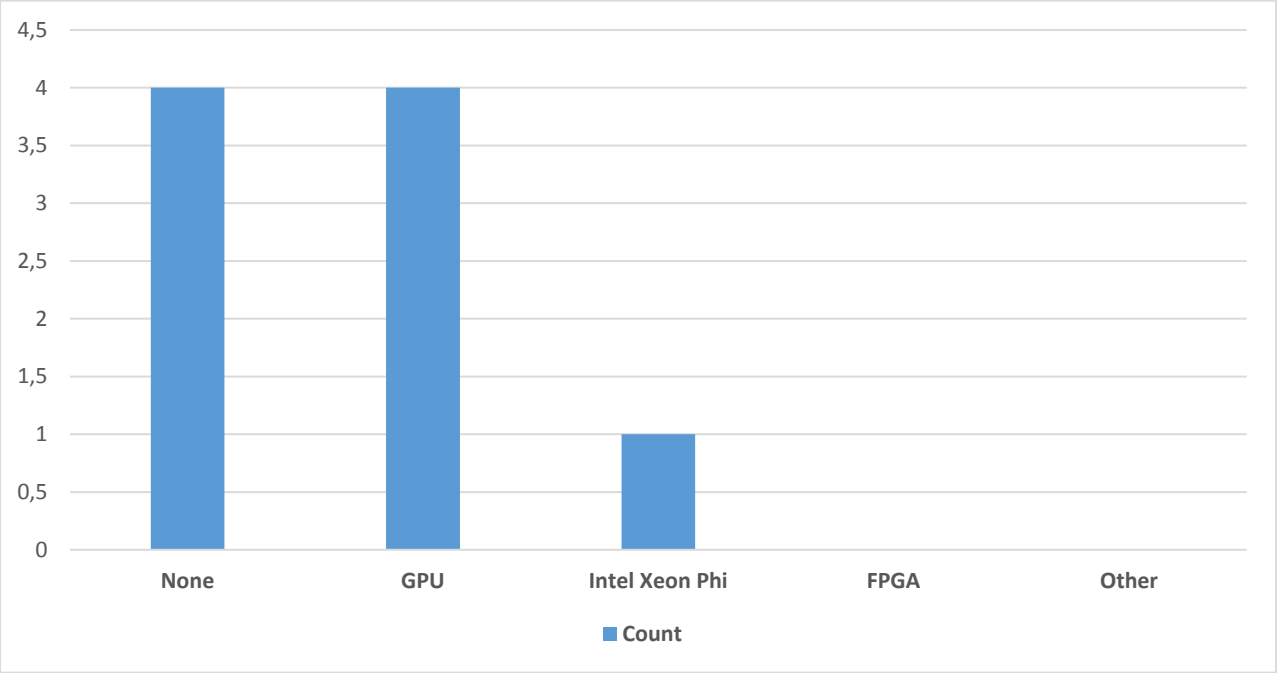

Figure 7: Accelerator types currently used at PRACE Tier-0/Tier-1 sites.

Analysis

The feedback obtained from CoEs suggests that the majority of users (63% of the questionnaires gave value 8 or more, see Section 3.1) indicates a need for GPUs. There is significantly less demand for Intel's Xeon Phis and FPGAs – 18% and 9% correspondingly. The current situation in HPC centres seem to reflect the demand of accelerators with exception of FPGAs, most probably due to the limited availability of supporting system software and the difficulty regarding application porting/software development.

5.3 [HPC sites] Q3 – Size of main memory per node (in GByte)

Figure 8 illustrates the results obtained from nine PRACE HPC sites regarding the size of main memory per node.

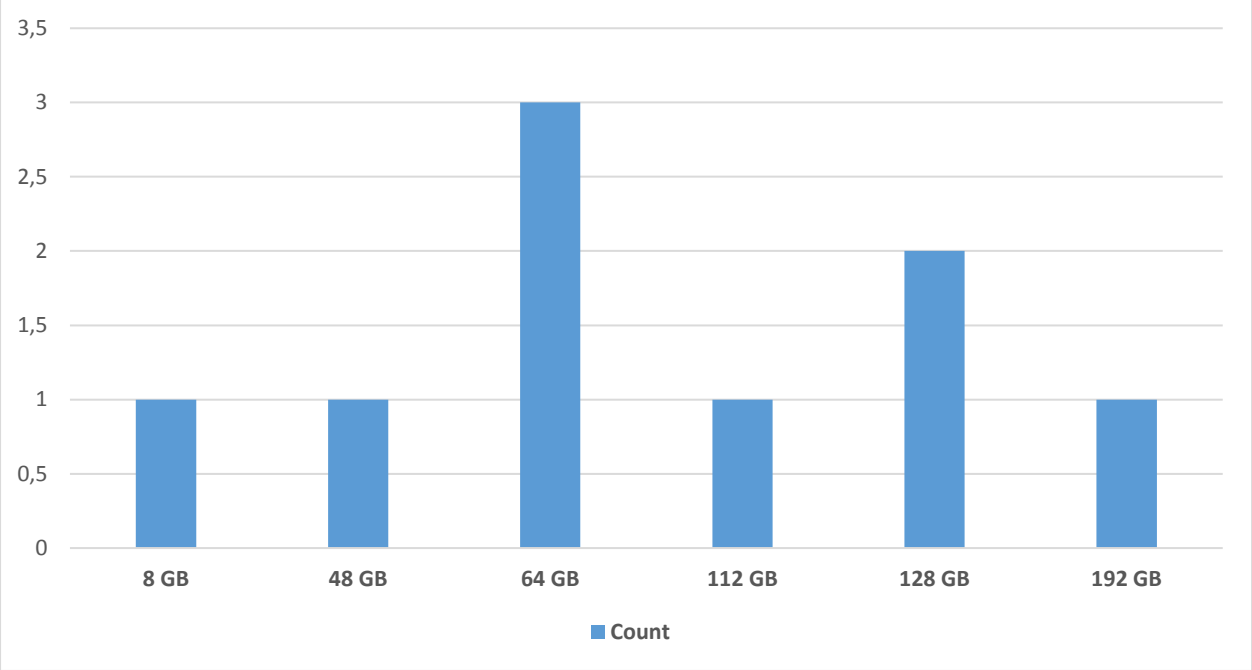

Figure 8: Size of main memory per node at PRACE Tier-0/Tier-1 sites.

Analysis

Memory size seems to be an important but not critical requirement for HPC machines as 90% of answers obtained from HPC user communities (see Section 3.1) put values between 6 and 8 on scale 1-12 with 1 being least important. This is not surprising, since the majority of traditional HPC applications scale in a way that one can distribute the problem to more compute nodes if more total memory is required. This is also well reflected in the current state of the Tier-0/Tier-1 systems - a balance between compute power and memory capacity can be identified. Most of these systems have at least 64GB of memory but there are none with 256GB or more.

5.4 [HPC sites] Q4 - Which storage technologies are you using?

Four options for possible answers were specified.

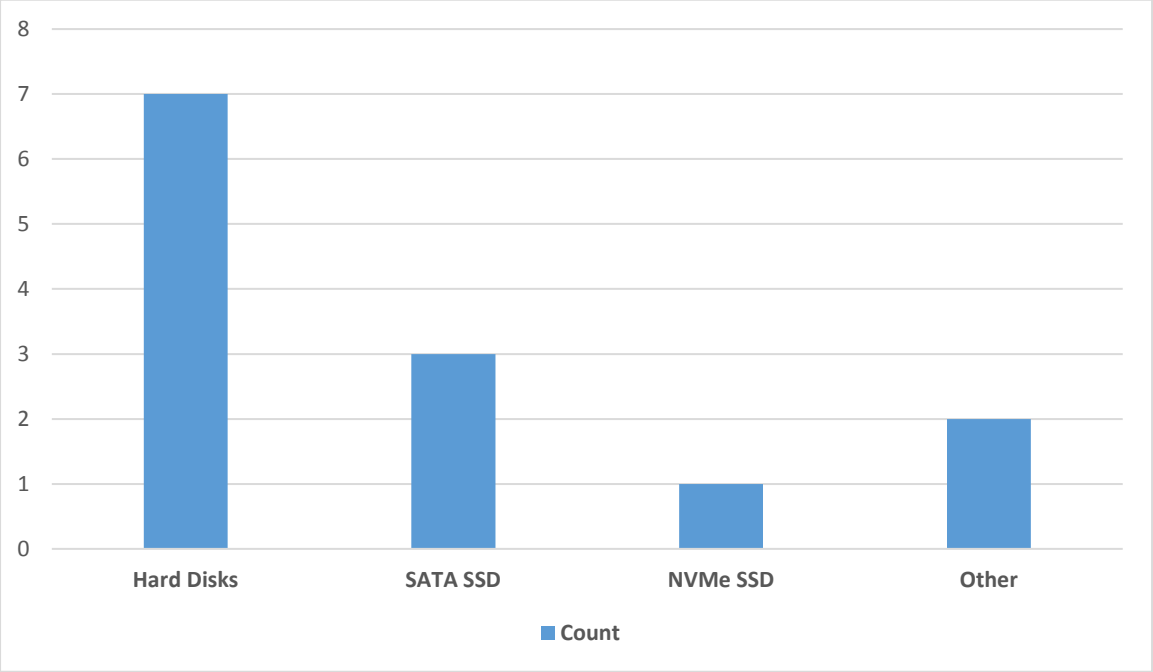

Figure 9: Storage technologies used at PRACE Tier-0 and Tier-1 sites.

Figure 9 presents the obtained answers. The "Other" choice gained the following options:

- shared storage from disk arrays;
- LUSTRE, NFS.

All sites but CINECA reported disk-less compute nodes, i.e. without disk drives.

Analysis

While I/O performance seems to be important for the users, the majority (63%) of the PRACE Tier-0/Tier-1 systems are using storage that is based on traditional hard drives due to good price/performance and capacity ratios. It is possible to achieve an I/O performance out of relatively slow HDD drives by merging into RAID arrays and exposing it to users as shared file system thus granting a good user experience at moderate cost. This is currently a dominating paradigm – all but one centre build the clusters without any local storage in the compute nodes. However, as can be observed, faster storage technologies are also being introduced – dropping costs of fast SSD storage makes it perfect as cache for slower HDDs.

While there is a demand from the users for "persistent storage on the node" (36% of answers gave score between 10 and 12, see Section 3.1) it is unclear if this requirement is fulfilled by shared storage or it really suggests the importance of real local per-node storage.

5.5 [HPC sites] Q5 – Memory bandwidth per node (in GByte/s)

Figure 10 presents the per node memory bandwidth of nine PRACE Tier-0/Tier-1 sites.

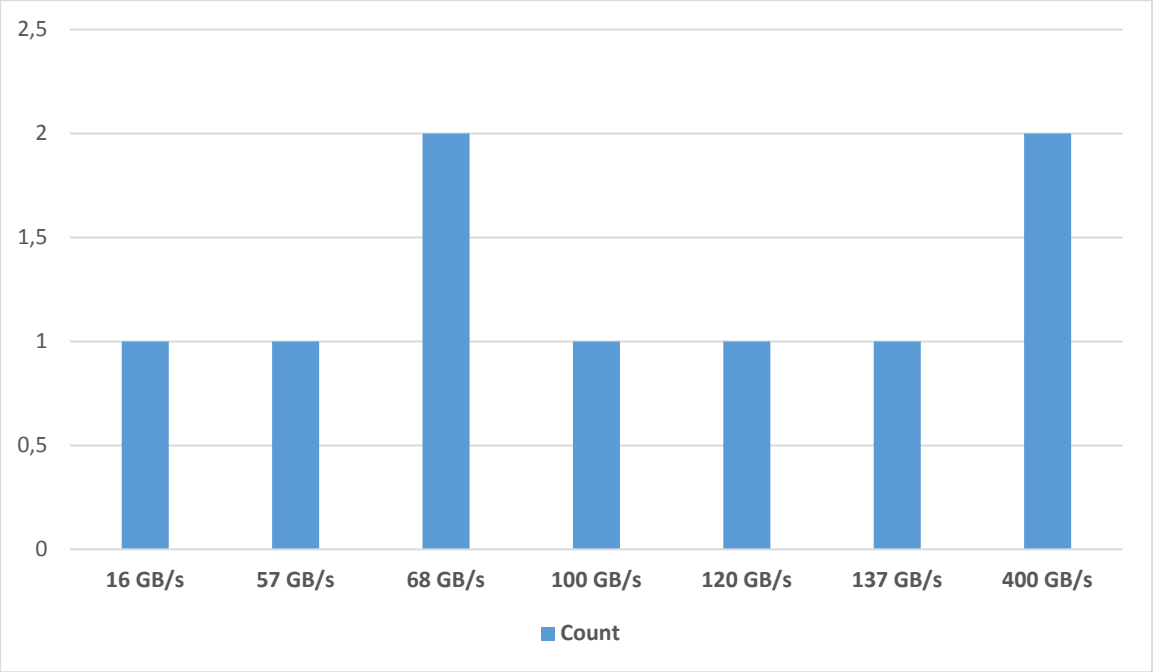

Figure 10: Memory bandwidth per node at PRACE Tier-0/Tier-1 sites.

Analysis

The importance of the memory bandwidth relies heavily on the problem characteristics that is to be solved on the system – there are both "1" scores (it is not important) and "12" (very important) regarding the memory bandwidth from CoEs (see Section 3,1). Unfortunately, HPC centres have very little influence on this aspect of their machines – due to the x86 monopoly, the bandwidth seems to be reflecting generation of the CPU that is installed in the cluster. Values above 200 GB/s apply to accelerators where small, but expensive, HBM or GDDR memory grants significant benefits to applications that can use this hardware.

5.6 [HPC sites] Q6 – Is node level or/and application level isolation supported

In past, many HPC architectures didn't take virtualization into consideration. Current virtualization technologies allow HPC workloads to leverage resources more efficiently, making a virtual HPC architecture appealing. Containers and Virtual Machines (VMs) are the most common virtualization techniques.

Containers are an abstraction at the application layer which combines code and dependencies into one package – several containers can run on a single node/server and share the OS kernel between each other, where each runs as an isolated process in the user space.

VMs are an abstraction of a physical hardware layer, which turns one node/server into multiple virtual ones. Each VM contains the complete copy of underlying OS, necessary binary and libraries, etc. They are usually larger (in memory size) as compared to containers, and are much slower to boot.

Figure 11 shows the current node/application level isolation at PRACE Tier-0/Tier-1 sites.

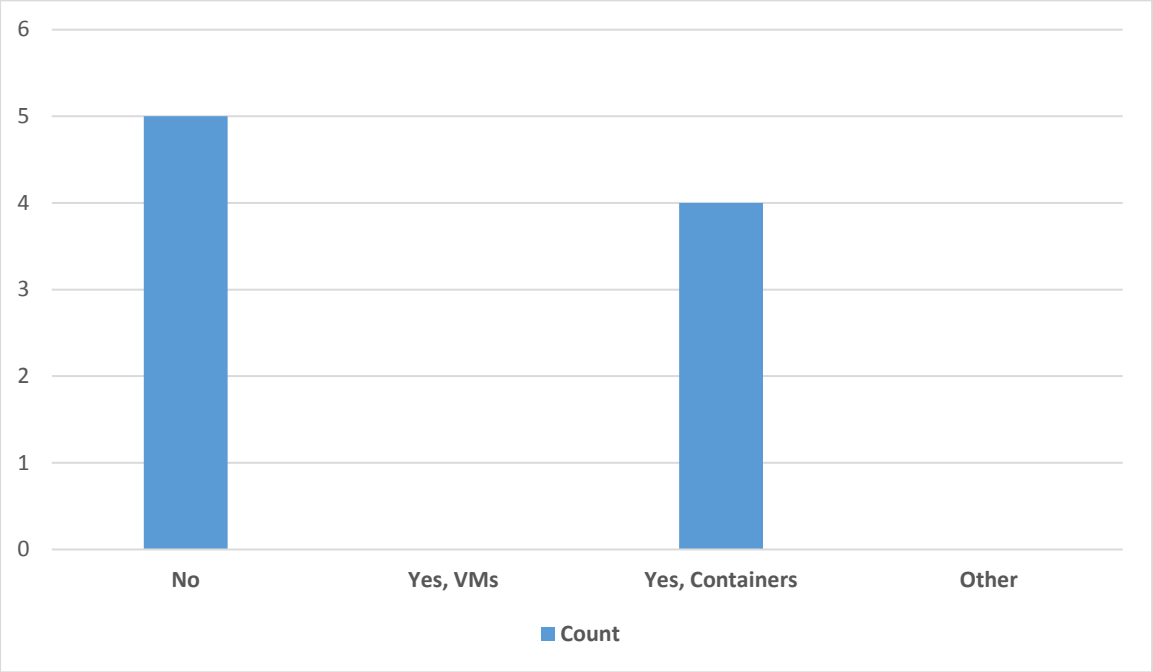

Figure 11: Node/Application level isolation at PRACE Tier-0/Tier-1 sites.

Analysis

Traditionally, HPC machines are very monolithic – there is unified OS version and libraries are provided on cluster scale by management teams. Modern software development methodologies are influenced by the tools available for cloud application development where the author of the software has big influence on which libraries and even OS is used. This situation forced adoption of some virtualization techniques to HPC environments. Currently container technology seems to be gaining popularity as it allows for more flexibility for the users on one hand, and on the other by simplifying life of HPC system administrators. Full virtualization is however not adopted as it introduces performance overheads, disrupts HPC cluster security model and is more suited to cloud environments where single server is rather a persistent entity.

5.7 [HPC sites] Q7 - Network topology

One can see that the need for fast interconnects is very important for the users – 45% of the answers scored 10-12 (see Section 3.1) so it should be reflected in all aspects of the HPC network design and technology selection. The bar chart below presents the main network topologies used at

PRACE Tier-0/Ter-1 sites and the Figure 12 shows the bisection bandwidth of interconnect in TByte/s per node.

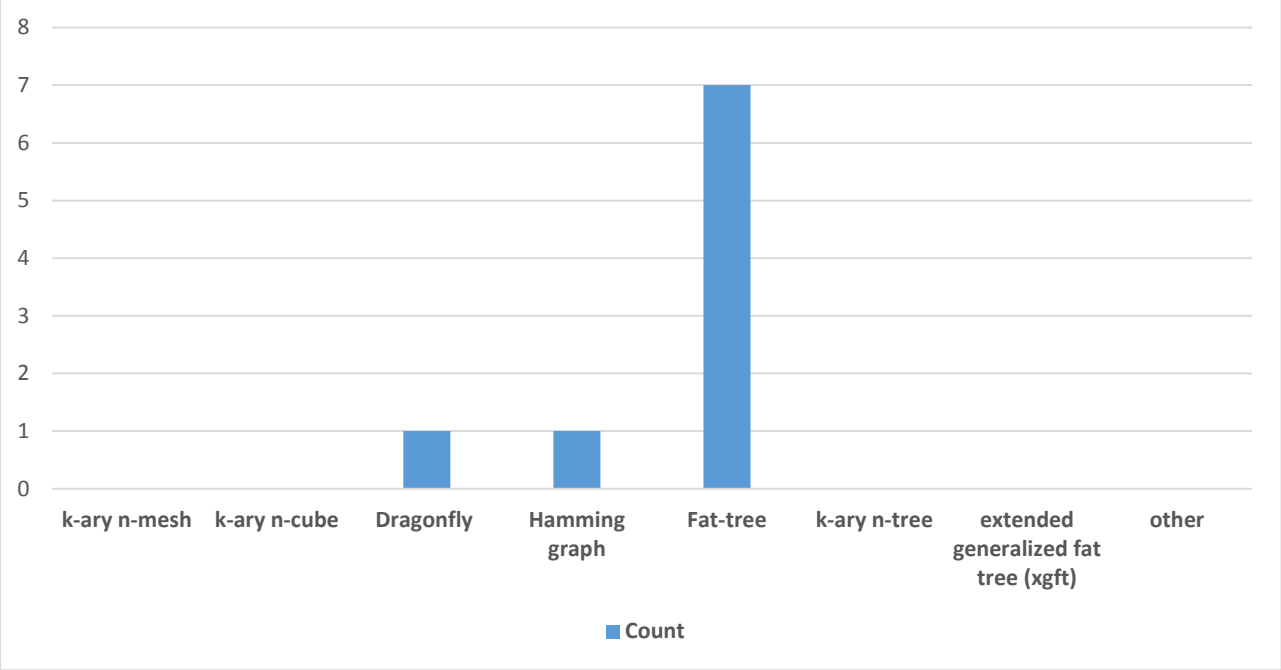

Figure 12: Network topologies at PRACE Tier-0/Tier-1 sites.

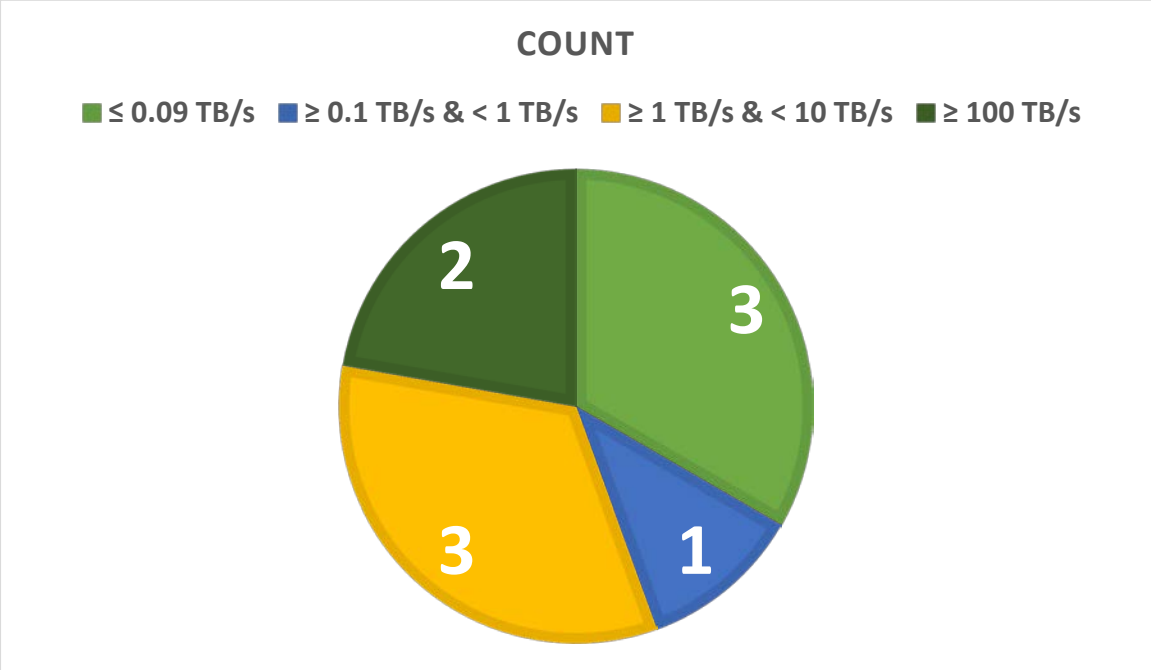

Figure 13: Clustering of bisection bandwidth per node.

Analysis

Currently the Fat-tree topology is used in most of the clusters with the exceptions of HPC machines having vendor-specific topologies. Most probably, the popularity of this topology is caused by its relatively simple design, implementation, and usage (ease of application placement).

The spread of bisection bandwidth of the interconnect in the answers illustrated in Figure 13 is 4 orders of magnitude wide. The top two systems with 100 TByte/s and 360 TByte/s bisection bandwidth are non-Fat tree topologies because both Dragonfly and Hamming graph topologies are usually characterized by higher bisection bandwidth in relation to cost of interconnect. While non-blocking fat tree topologies are featured by the maximum bisection bandwidth, due to high cost of network equipment it is uncommon to deploy this topology in large scale installations. Looking at the values one can deduct (while there is no direct data in the survey supporting directly this interpretation) that in most cases there are few or no applications that span the entire machine or these applications are not bandwidth sensitive. We can assume that the owners of the compute clusters keep track of the typical job requirements and collaborates with local user communities when preparing technical specifications for new clusters. This might explain the spread of the values for the bisection bandwidth parameter – whenever it is needed it is provided but implemented in a way that is a compromise between performance and networking infrastructure cost.

6 Conclusions and outlook

This document aims to provide an overview of the requirements, in terms of technologies and architectures, for the next generation computing systems evolving towards Exascale. It includes the vision of the user communities, the vision of the HPC centres and the correlation between the two.

The need for prototype systems involving a heterogeneous system architecture is seen by the user communities as well as by HPC centres.

The accelerator-assisted computing is becoming essential for performance improvement of not only traditional HPC applications, but also for various visualization, big data, data analytics, and machine learning related challenges. According to the survey results, GPUs are the most appealing accelerator systems for the European HPC user community. Interestingly enough, 45% of PRACE Tier-0/Tier-1 HPC sites cover this requirement, at least as part of the system. These HPC sites reported to have a configuration of 4 or 2 GPU accelerators per node.

Another interesting observation is the foreseen shift for HPC sites from conventional x86 based processing technologies (which as was seen is currently dominating at PRACE HPC sites) to alternatives such are ARM, IBM Power Architecture, PTX.

The conducted surveys also showed that containers are getting more appealing due to their higher efficiency as compared to the full, hardware-level, virtualization. Containers, being a form of virtualization, offer a better performance by placing applications closer to the host system. The DevOps (Development and Operations) workflow support of containers, allowing to move a tested application from one environment to another without any porting or re-testing efforts, makes the containers very useful also from a user's perspective.

In summary, the features that are most wanted for testing in future prototype systems comprise from heterogeneous architectures that include new kinds of memory and parallel I/O file systems.

From the data centre infrastructure point of view, the growing power density for the required heterogeneous nodes further motivates the need for the adoption of water cooling technologies.

There will be a further deliverable in WP5 at M27 on "Extended best practice guide for prototypes and demonstrators" that will extend the previously developed best practice guide by PRACE-4IP WP5 with tools to evaluate prototype systems with regard to their usability and fit for purpose.