E-Infrastructures
H2020- INFRAEDI-2018-2020

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

PRACE-6IP

PRACE Sixth Implementation Phase Project

Grant Agreement Number: INFRAEDI-823767

D5.2

Worldwide HPC technology and market landscape

Final

Version:	1.1
Author(s):	Evangelia Athanasaki, GRNET; Norbert Meyer, PSNC; Andreas Johansson, SNIC; Christelle Piechurski, GENCI; Dirk Pleiter, KTH; Ezhilmathi Krishnasamy, U Luxembourg; Mikael Johansson, CSC
Date:	30.11.2021

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: INFRAEDI-823767**	
	Project Title: PRACE Sixth Implementation Phase Project	
	Project Web Site: http://www.prace-ri.eu/about/ip-projects	
	Deliverable ID: D5.2	
	Deliverable Nature: Report	
	Dissemination Level: PU*	**Contractual Date of Delivery:** 30 / November / 2021
		Actual Date of Delivery: 30 / November / 2021
	EC Project Officer: Leonardo Flores Añover	

* - The dissemination level is indicated as follows: **PU** – Public, **CO** – Confidential, only for members of the consortium (including the Commission Services) **CL** – Classified, as referred to in Commission Decision 2005/444/EC.

Document Control Sheet

Document	**Title: Worldwide HPC technology and market landscape**	
	ID: D5.2	
	Version: 1.1	**Status:** ***Final***
	Available at: http://www.prace-ri.eu/about/ip-projects	
	Software Tool: Microsoft Word 2016	
	File(s): D5.2-v1.1.docx	

Authorship	**Written by:**	Andreas Johansson, SNIC;
		Christelle Piechurski, GENCI;
		Dirk Pleiter, KTH;
		Ezhilmathi Krishnasamy, U Luxembourg;
		Evangelia Athanasaki, GRNET;
		Mikael Johansson, CSC;
		Norbert Meyer, PSNC

	Contributors:	Adem Tekin, UHEM;
		Ahmet Tuncer Durak, UHEM;
		Damian Kaliszan, PSNC;
		Dirk Pleiter, JSC-GCS;
		Enver Özdemir, UHEM;
		Fethiye Aylin Sungur, U Luxembourg;
		Filip Blicharczyk, PSNC;
		Fredrik Robertsén, CSC;
		Krzysztof Wadówka, PSNC;
		Michael Mucciardi, BSC;
		Michal Pilc, PSNC;
		Mirco Cestari, CINECA;
		Philipp Gschwandtner, UIBK;
		Sebastien Varrette, U Luxembourg

	Reviewed by:	Walter Lioen, SURF; Veronica Teodor, JUELICH
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	6/October/2021	1st draft	Initial draft and setup of the document structure
0.2	2/October/2021	2nd draft	2nd compiled version with additions to section 2
0.3	27/October/2021	3rd draft	Executive Summary and additional updates
0.4	3/November/2021	4th draft	Ready for WP5 internal review
0.5	8/November/2021	5th draft	Draft with WP5 and PMO internal review comments
0.6	15/November/2021	6th draft	Integration of reviewers' comments
0.7	22/November/2021	7th draft	Draft with 2nd review comments
1.0	23/November/2021	Final	Integration of reviewers' comments. Addition of technical reports as an annex in pdf version
1.1	30/November/2021	Final	Final adjustments

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Technology landscape, Market landscape

Disclaimer

This deliverable has been prepared by the responsible work package of the project in accordance with the Consortium Agreement and the Grant Agreement n° INFRAEDI-823767. It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the project and to the extent foreseen in such agreements. Please note that even though all participants to the project are members of PRACE aisbl, this deliverable has not been approved by the Council of PRACE aisbl and therefore does not emanate from it nor should it be considered to reflect PRACE aisbl's individual opinion.

Copyright notices

Table of Contents

List of Figures

References and Applicable Documents

[1] https://prace-ri.eu/infrastructure-support/market-and-technology-watch/

[2] https://prace-ri.eu/wp-content/uploads/Edge-Computing-An-Overview-of-Framework-and-Applications-1.pdf

[3] https://doi.org/10.5281/zenodo.5534072

[4] https://prace-ri.eu/wp-content/uploads/Data-Management-Services-and-Storage-1.pdf

[5] https://doi.org/10.5281/zenodo.5534064

[6] https://prace-ri.eu/wp-content/uploads/State-of-the-Art-and-Trends-for-Computing-and-Interconnect-Network-Solutions-for-HPC-and-AI-1.pdf

[7] https://doi.org/10.5281/zenodo.5534080

[8] https://prace-ri.eu/wp-content/uploads/TR-Quantum-Computing-A-European-Perspective.pdf

[9] https://doi.org/10.5281/zenodo.5547408

[10] https://prace-ri.eu/wp-content/uploads/Security-in-an-Evolving-European-HPC-Ecosystem.pdf

[11] https://doi.org/10.5281/zenodo.5638456

List of Acronyms and Abbreviations

ACID	Atomicity, Consistency, Isolation, Durability
AI	Artificial Intelligence
API	Application Programming Interface
APU	Accelerated Processing Unit
CPU	Central Processing Unit
DDR	Double Data Rate
DL	Deep Learning
DNS	Domain Name System
DOI	Digital Object Identifier
EC	European Commission
EPI	European Processor Initiative
EuroHPC JU	The European High Performance Computing Joint Undertaking
EXDCI-2	European Extreme Data & Computing Initiative
FPGA	Field Programmable Array
GPPs	General Purpose Processors
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
HPC	High Performance Computing
IDS	Intrusion Detection Systems
INFRAG	Infrastructure Advisory Group
I/O	Input/Output
IoT	Internet of Things

MCM	Multi-Chip Module
ML	Machine Learning
NVMe	Non-Volatile Memory express
POSIX	Portable Operating System Interface
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
QC	Quantum computing
RIAG	Research & Innovation Advisory Group (RIAG)
SSH	Secure Socket Shell
SVE	Scalable Vector Extension
WP	Work Package

List of Project Partner Acronyms

BADW-LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Germany (3rd Party to GCS)
BILKENT	Bilkent University, Turkey (3rd Party to UHEM)
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion, Spain
CaSToRC	The Computation-based Science and Technology Research Center (CaSToRC), The Cyprus Institute, Cyprus
CCSAS	Computing Centre of the Slovak Academy of Sciences, Slovakia
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3rd Party to GENCI)
CENAERO	Centre de Recherche en Aéronautique ASBL, Belgium (3rd Party to UANTWERPEN)
CESGA	Fundacion Publica Gallega Centro Tecnológico de Supercomputación de Galicia, Spain, (3rd Party to BSC)
CINECA	CINECA Consorzio Interuniversitario, Italy
CINES	Centre Informatique National de l'Enseignement Supérieur, France (3rd Party to GENCI)
CNRS	Centre National de la Recherche Scientifique, France (3rd Party to GENCI)
CSC	CSC Scientific Computing Ltd., Finland
CSIC	Spanish Council for Scientific Research (3rd Party to BSC)
CYFRONET	Academic Computing Centre CYFRONET AGH, Poland (3rd Party to PNSC)
DTU	Technical University of Denmark (3rd Party of UCPH)
EPCC	EPCC at The University of Edinburgh, UK
EUDAT	EUDAT OY
ETH Zurich (CSCS)	Eidgenössische Technische Hochschule Zürich – CSCS, Switzerland
GCS	Gauss Centre for Supercomputing e.V., Germany
GÉANT	GÉANT Vereniging
GENCI	Grand Equipement National de Calcul Intensif, France
GRNET	National Infrastructures for Research and Technology, Greece
ICREA	Catalan Institution for Research and Advanced Studies (3rd Party to BSC)
INRIA	Institut National de Recherche en Informatique et Automatique, France (3rd Party to GENCI)
IST-ID	Instituto Superior Técnico for Research and Development, Portugal (3rd Party to UC-LCA)

IT4I	Vysoka Skola Banska - Technicka Univerzita Ostrava, Czech Republic
IUCC	Machba - Inter University Computation Centre, Israel
JUELICH	Forschungszentrum Jülich GmbH, Germany
KIFÜ (NIIFI)	Governmental Information Technology Development Agency, Hungary
KTH	Royal Institute of Technology, Sweden (3rd Party to SNIC-UU)
KULEUVEN	Katholieke Universiteit Leuven, Belgium (3rd Party to UANTWERPEN)
LiU	Linkoping University, Sweden (3rd Party to SNIC-UU)
MPCDF	Max Planck Gesellschaft zur Förderung der Wissenschaften e.V., Germany (3rd Party to GCS)
NCSA	NATIONAL CENTRE FOR SUPERCOMPUTING APPLICATIONS, Bulgaria
NTNU	The Norwegian University of Science and Technology, Norway (3rd Party to SIGMA2)
NUI-Galway	National University of Ireland Galway, Ireland
PRACE	Partnership for Advanced Computing in Europe aisbl, Belgium
PSNC	Poznan Supercomputing and Networking Center, Poland
SDU	University of Southern Denmark (3rd Party to UCPH)
SIGMA2	UNINETT Sigma2 AS, Norway
SNIC-UU	Uppsala Universitet, Sweden
STFC	Science and Technology Facilities Council, UK (3rd Party to UEDIN)
SURF	SURF is the collaborative organisation for ICT in Dutch education and research
TASK	Politechnika Gdańska (3rd Party to PNSC)
TU Wien	Technische Universität Wien, Austria
UANTWERPEN	Universiteit Antwerpen, Belgium
UC-LCA	Universidade de Coimbra, Labotatório de Computação Avançada, Portugal
UCPH	Københavns Universitet, Denmark
UEDIN	The University of Edinburgh
UHEM	Istanbul Technical University, Ayazaga Campus, Turkey
UIBK	Universität Innsbruck, Austria (3rd Party to TU Wien)
UiO	University of Oslo, Norway (3rd Party to SIGMA2)
UL	UNIVERZA V LJUBLJANI, Slovenia
ULIEGE	Université de Liège; Belgium (3rd Party to UANTWERPEN)
U Luxembourg	University of Luxembourg
UM	Universidade do Minho, Portugal, (3rd Party to UC-LCA)
UmU	Umeå University, Sweden (3rd Party to SNIC-UU)
UnivEvora	Universidade de Évora, Portugal (3rd Party to UC-LCA)
UnivPorto	Universidade do Porto, Portugal (3rd Party to UC-LCA)
UPC	Universitat Politècnica de Catalunya, Spain (3rd Party to BSC)
USTUTT-HLRS	Universitaet Stuttgart – HLRS, Germany (3rd Party to GCS)
WCSS	Politechnika Wroclawska, Poland (3rd Party to PNSC)

Executive Summary

The work package 5 (WP5) – "HPC Planning and Commissioning" of PRACE-6IP builds on the work of previous PRACE-IP projects in terms of technology watch, know-how and best practices for energy-efficient High Performance Computing (HPC) Centre Infrastructures design and operations, and prototyping of HPC systems. The main aim is to deliver, in the general perspective of the EU HPC changing landscape and EuroHPC, information and guidance which could be useful for decision makers at different levels: PRACE aisbl, PRACE members, EuroHPC Joint Undertaking (JU) and other European HPC sites. On this direction, Task 5.1 – "Europe-Centred View of the Worldwide HPC Technology and Market Landscape" studies worldwide projects and strategies towards Exascale and compares what is happening in different parts of the world, in perspective of efforts in Europe. It also carries out market studies and overviews, focusing in particular on the market shares of European and non-European suppliers and the adoption of European technologies and complementary paradigms to HPC in Europe.

After extended discussions with EC, the Infrastructure Advisory Group (INFRAG), the Research & Innovation Advisory Group (RIAG) and European Extreme Data & Computing Initiative (EXDCI-2), it was decided to focus the work within Task 5.1 on specific topics. Six topics were chosen and presented separately through technical reports: "Edge Computing: An Overview of Framework and Applications", "Data Management Services and Storage", "State-of-the-Art and Trends for Computing and Interconnect Network Solutions for HPC and AI", "Quantum Computing – A European Perspective" and "Security in an Evolving European HPC Ecosystem", "User requirements influencing HPC technologies".

This document is the final deliverable of WP5, Task 5.1, which summarises the work done within this task during the project. It outlines the details of the technical reports selection process and briefly describes the conclusions derived within the reports.

1 Introduction

Task 5.1 “Europe-Centred View of the Worldwide HPC Technology and Market Landscape” of PRACE-6IP Work Package 5 (WP5), is the continuation of a well-established effort of previous PRACE Implementation Phase projects (IP) to carry out an assessment of the HPC market based on market surveys, Top500 and Green500/HPCG lists analyses, supercomputing conferences (mainly ISC, SC and the European Workshops on HPC Infrastructures) and information exchange between vendors and WP5 experts who are involved in the work package. Since PRACE is the largest independent European HPC community, it can provide a neutral Technology Watch vision. The PRACE community has the expertise on how to introduce new technologies in advance. Early technology adoption ensures successful massive deployment and optimal exploitation of resources by the users communities.

Within the EuroHPC era, while PRACE-6IP WP5 is providing an exhaustive view on technologies that will fly on the market over the future years, the new scope defined after several discussions also focuses on worldwide market shares to understand how Europe is positioned compared to the worldwide landscape, i.e. at least US, China, and Japan.

The first three technical reports, published in December 2020, were:

- *“Edge Computing: An Overview of Framework and Applications”* explains why Edge Computing is needed and how the edge architecture is typically structured, presenting the technologies that help this cutting-edge model to function properly.
- *“Data Management Services and Storage”* summarises the history of this development and examines some of the technologies that are building blocks of near-future storage systems, both hardware and software required to manage the large amounts of data.
- *“State-of-the-Art and Trends for Computing and Interconnect Network Solutions for HPC and AI”* provides a consolidated view on the current and mid-term technologies (2019-2022+) for two important components of an HPC/AI system: computing (general purpose processor and accelerators) and interconnect capabilities and provides an outlook on future trends in terms of mid-term projections about what users may expect in the coming years.

Two technical reports were published in October-November 2021:

- *“Quantum Computing – A European Perspective”* gives an overview of quantum computing, the present state of affairs, and future scenarios.
- *“Security in an Evolving European HPC Ecosystem”* analyses challenges and requirements related to security in the context of an evolving European HPC ecosystem, to provide selected strategies on how to address them, and to come up with a set of forward-looking recommendations.

The final technical report is expected in January 2022:

- *“User Requirements influencing HPC Technologies”* will present how both software and hardware architectures are affected and can be adjusted to follow new computing trends and meet future needs. It is still under development given the timelines of PRACE-6IP PMO.

The remainder of this document is organised as follows. Chapter 2 describes the selection process of technical report topics, discusses the organisational aspects and the review process.

Chapter 0 presents brief descriptions of the technical reports that are (and will be) developed within PRACE-6IP. Chapter 4 provides the conclusion. Annex 1 gives the official published version of the five technical reports.

2 Approach of technical reports

2.1 Selection of topics

The outcome of Task 5.1 activities was originally planned to be published in 2 white papers on Market and Technology Watch in 2020 and 2021 and to be summarised in the final deliverable D5.2, following a format similar to previous PRACE market and technology watch white papers and deliverables. The participants had already agreed on the table of contents of the first white paper in May 2019 and authors and chapter editors were assigned.

However, after the final review of PRACE-5IP and EC recommendations, the scope of both the white papers and the deliverable has changed and turned to be more European centric, covering new, emergent and complementary paradigms to HPC. The scope and focus of the technology watch was also refined in collaboration with the INFRAG, the RIAG and EXDCI-2, to include topics of their interest and avoid overlaps with their work.

Finally, after several discussions, WP5 and PMO have decided to switch from an extensive monolithic annual market watch report to a series of technical reports that aim to describe the state-of-the-art and mid-term trends of the technology and market landscape in the context of HPC and Artificial Intelligence (AI), edge-, cloud- and interactive computing, big data and other related technologies. The series provides information and guidance useful for various European sites, members of PRACE and EuroHPC, and the EuroHPC Advisory Groups INFRAG and RIAG. Especially INFRAG is interested in a long-term analysis of the HPC market, user requirements, trends and comprehensive recommendations.

2.2 Organisation

Monthly teleconferences were organised that involved discussions on progress, pending issues, timelines, updates from partners involved, etc. Separate discussions were also organised by WP5 leader, Task 5.1 leader and PMO with INFRAG, RIAG and EXDCI-2 in order to be aligned on the topics selection and the final form and content of technical reports. Meeting minutes and presentation slides are uploaded to the project internal collaboration platform, BSCW.

The development activity was led by the Task 5.1 leader Norbert Meyer (PSNC) assisted by WP5 leader Volker Weinberg (BADW-LRZ), with the support of the technical reports lead authors:

- Andreas Johansson (SNIC): *“Data Management Services and Storage”*
- Christelle Piechurski (GENCI): *“State-of-the-Art and Trends for Computing and Interconnect Network Solutions for HPC and AI”*
- Dirk Pleiter (KTH): *“Security in an Evolving European HPC Ecosystem”*
- Ezhilmathi Krishnasamy (U Luxembourg): *“Edge Computing: An Overview of Framework and Applications”*
- Evangelia Athanasaki (GRNET)/Norbert Meyer (PSNC): *“User Requirements influencing HPC Technologies”*
- Mikael Johansson (CSC): *“Quantum Computing – A European Perspective”*

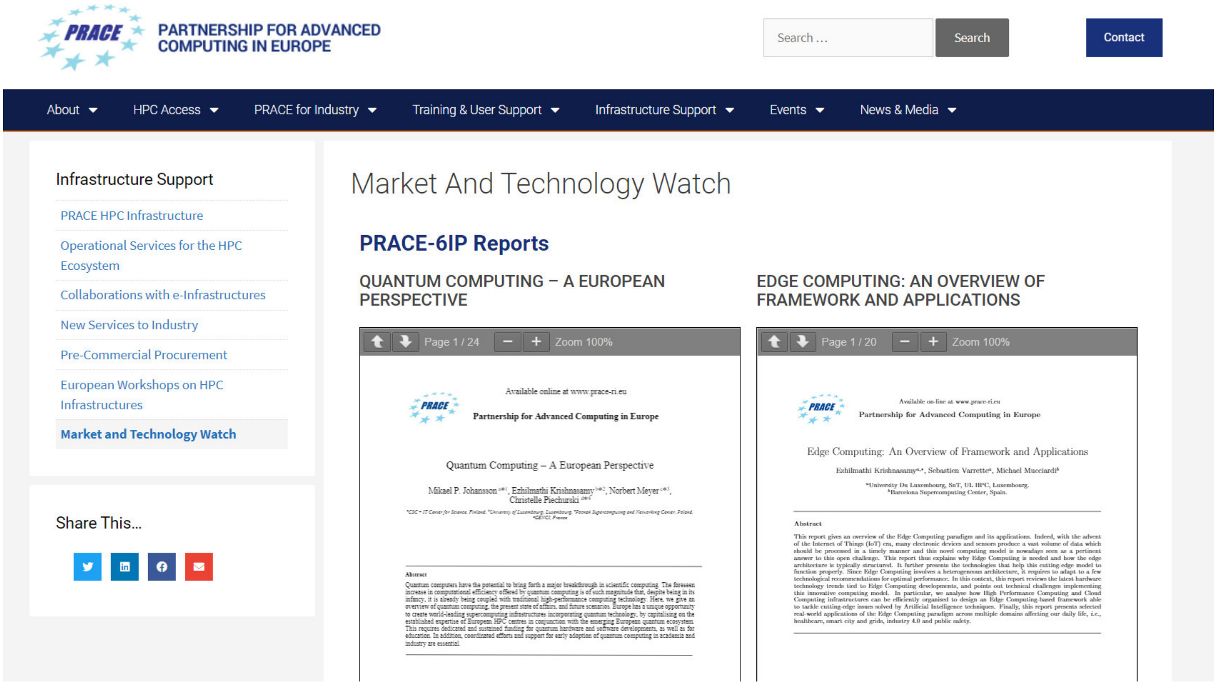
2.3 Review and Publication

The internal review of these technical reports followed the quality assurance process in place for all PRACE-IP deliverables and white papers – each report was internally reviewed by one member of the PMO and one project representative. This has further improved the quality of the published reports and deliverables. Prior to publication, the first set of three technical reports ready in December 2020 was sent for feedback also to INFRAG and RIAG, which confirmed the good quality and importance of these documents.

To facilitate further outreach of the technical reports, they are published on Zenodo and thus, have been assigned with a DOI.

3 Summaries of Technical Reports

Technical reports are publicly available, so this section only provides a brief overview on the six reports that have been developed within PRACE-6IP. The complete list of all market and technology watch reports published so far within the PRACE-IP projects, starting from 2016 within PRACE-4IP, can be found via [1] (Figure 1).



Market And Technology Watch

PRACE-6IP Reports

QUANTUM COMPUTING – A EUROPEAN PERSPECTIVE

Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

Quantum Computing – A European Perspective

Mikael P. Johansson a*1, Ezhilmathi Krishnasamy b*2, Norbert Meyer c*3, Christelle Piechurski d*4

aCSC – IT Center for Science, Finland, bUniversity of Luxembourg, Luxembourg, cPoznań Supercomputing and Networking Center, Poland, dGENCI, France

Abstract

Quantum computers have the potential to bring forth a major breakthrough in scientific computing. The foreseen increase in computational efficiency offered by quantum computing is of such magnitude that, despite being in its infancy, it is already being coupled with traditional high-performance computing technology. Here, we give an overview of quantum computing, the present state of affairs, and future scenarios. Europe has a unique opportunity to create world-leading supercomputing infrastructures incorporating quantum technology, by capitalising on the established expertise of European HPC centres in conjunction with the emerging European quantum ecosystem. This requires dedicated and sustained funding for quantum hardware and software developments, as well as for education. In addition, coordinated efforts and support for early adoption of quantum computing in academia and industry are essential.

EDGE COMPUTING: AN OVERVIEW OF FRAMEWORK AND APPLICATIONS

Available on-line at www.prace-ri.eu

Partnership for Advanced Computing in Europe

Edge Computing: An Overview of Framework and Applications

Ezhilmathi Krishnasamy a,*, Sebastien Varrette a, Michael Mucciardi b

aUniversity Du Luxembourg, SnT, UL HPC, Luxembourg.
bBarcelona Supercomputing Center, Spain.

Abstract

This report gives an overview of the Edge Computing paradigm and its applications. Indeed, with the advent of the Internet of Things (IoT) era, many electronic devices and sensors produce a vast volume of data which should be processed in a timely manner and this novel computing model is nowadays seen as a pertinent answer to this open challenge. This report thus explains why Edge Computing is needed and how the edge architecture is typically structured. It further presents the technologies that help this cutting-edge model to function properly. Since Edge Computing involves a heterogeneous architecture, it requires to adapt to a few technological recommendations for optimal performance. In this context, this report reviews the latest hardware technology trends tied to Edge Computing developments, and points out technical challenges implementing this innovative computing model. In particular, we analyse how High Performance Computing and Cloud Computing infrastructures can be efficiently organised to design an Edge Computing-based framework able to tackle cutting-edge issues solved by Artificial Intelligence techniques. Finally, this report presents selected real-world applications of the Edge Computing paradigm across multiple domains affecting our daily life, i.e., healthcare, smart city and grids, industry 4.0 and public safety.

Figure 1: Snapshot illustrating PRACE web infrastructure for market and technology watch deliverables and technical reports

3.1 Edge Computing: An Overview of Framework and Applications

Presently, with the Internet of Things (IoT), we produce lots of data, and this data should be processed and analysed quickly, close to where it is produced. Edge Computing aims to solve these criteria, and it is considered an emerging technology. Cloud Computing and Supercomputers are also able to process and analyse a large volume of data. However, the major bottleneck with these infrastructures is data transfer latency. In particular, real-time applications such as camera and mobile phone data should be processed very quickly; this is where Edge Computing is applicable, because we cannot afford to lose time on data transfer from one location to another. More importantly, data privacy is a primary concern nowadays, and it is better to process the data within the premises where it is produced. Edge Computing is not meant to replace Cloud or Supercomputers; however, it is determined to share the computational burden.

Edge Computing uses embedded architecture devices, which consume less power and are more powerful to process real-time data, such as car cameras and airplane sensors. The Edge Computing architecture hierarchies are IoT, Edge layer, Fog layer, and Cloud layer. The IoT refers to cameras and other sensors that can produce the data; the Edge layer means the data is analysed and processed with advanced embedded architecture. The Fog layer & Cloud layer are handling large volumes of data that cannot be processed at the Edge layer.

Recently, many vendors have launched advanced and energy efficient embedded architectures that can be used at the Edge layer. For example, Nvidia's Jetson AGX Xavier has 512 CUDA cores, 8 CPU cores with 32 GB memory; and Kalray's KONIC200-HP has 160 CPU cores. These machines have tensor cores and, more importantly, are designed to perform the AI/ML computations efficiently. Moreover, these modern embedded architectures support advanced AI/ML numerical libraries to run AI/ML computations.

However, these advanced architectures still depend on the Cloud and Supercomputers in two scenarios: heterogeneous data and distributed computing. Sometimes, at the Edge, IoT devices produce different types of data: it can be video, audio, and text; and it might be hard to solve at the Edge layer, and hence Cloud Computing can be helpful. Often, IoT devices produce large volumes of data; in these situations, part of the data should be processed very quickly and some not. This is where distributed computing is applicable where both Edge Layer and Cloud or Supercomputers share the computation.

There are plenty of situations where Edge Computing applications are vital. They are medical applications, smart cities, industrial applications, and smart grid & public safety. Nevertheless, Edge Computing still faces some problems that prevent its applications from growing further in many areas. Notably, they are naming, programmability, and Edge device management. In particular, since Edge devices are heterogeneous, it is hard to program as an end-user compared to the cloud. For example, on the cloud, the end-users can deploy their application with minimal effort.

The complete technical report can be accessed via [2] on PRACE website or [3] on Zenodo.

3.2 Data Management Services and Storage

HPC storage systems have evolved from relatively simple systems attached to a single cluster to site-wide complex infrastructures supporting migration of data between storage tiers with different performance characteristics and I/O acceleration. Exascale systems and AI workloads will continue this trend by placing even greater demands on speedy access to data.

The report covers topics from the hardware substrate that sits at the base level of the storage infrastructure through file systems at the operating system level up to data management software and protocols that provides services to support research workflows.

One conclusion in the report is about the need for making data findable and accessible, and this requires software support for managing metadata.

Storage infrastructure

Hardware for storage needs to strike a compromise between cost, size and performance. Due to this need complex storage hierarchies have evolved and are now becoming even deeper as very high-performance flash devices using the NVMe protocol are placed closer to CPUs. Traditional hard drives are relegated to a supporting role as large volume storage devices. Flash technologies and storage solutions selected for the first announced Exascale systems are explored as well as hybrid disk arrays and tape libraries.

File systems

In this section some file systems commonly used at HPC sites are examined with regards to functionality (access protocols and tiering support for example) and how they have evolved over time. File systems covered are Lustre, Spectrum Scale, BeeGFS and Ceph.

While some of the file systems support object storage, the focus in this section is on their use as traditional POSIX file systems for compute clusters.

Data management services

To meet future needs, new technologies for data management and tiering are becoming increasingly important and this section explores several available technologies.

Cloud storage system have standardised access methods for object storage, and these can also be used for HPC storage to enable workflows combining traditional clusters and cloud technologies.

Tiering can be used for both handling I/O acceleration for data used in computations, and moving data to systems suited for long-term storage of data that is rarely accessed. Both types are covered with multiple examples of each type.

Handling input data to jobs and their resulting output is increasingly requiring support systems as both jobs and data sets increase in size. Many software suites are available, ranging from very domain specific ones to more general-purpose data management packages. Some systems also support publication of data sets or workflow integration.

The complete technical report can be accessed via [4] on PRACE website or [5] on Zenodo.

3.3 State-of-the-Art and Trends for Computing and Interconnect Network Solutions for HPC and AI

Since 2000, HPC resources have been extremely homogeneous in terms of underlying processors technologies. However, it becomes obvious that new trends tend to bring new microarchitectures for General Purpose Processors (GPPs) and new heterogeneous architectures, combining accelerators with GPPs, to sustain both numerical simulation and AI workflows. The report provides a consolidated view on 2019-2022+ most popular and known technologies both on general purpose processors and accelerators as interconnect capabilities starting by key factors influencing chips performance and their relation to architectural choices up to major trends below:

- The big CPU market players are still Intel and AMD (X86_64) with a strong competition between the two. While the X86_64 microarchitecture is still the most adopted in the HPC market, ARM processors are continuing to expand their market share through Fujitsu (A64FX), Marvell (ThunderX – until their cancelation), Amazon (Graviton), Ampere (Altra) and SiPearl, the company which designs and will sell the EPI Rhea processor, the only European ARM-based processor with HPC features (HBM and SVE capabilities).
- Market focus is to be part both of HPC and AI markets, with the capability to provide both CPU and GPU and to improve the CPU-GPU interconnect performance and CPU/GPU cache memory coherency as well as the global memory bandwidth, either on pure DDR technology or by using HBM. Most of the high-end computing capabilities would, at least, partially, rely on accelerators with CPU technologies hosting either GPUs, accelerators and/or FPGAs. FPGAs might see their use increased in the future for HPC and AI. Similar to competitor technologies, it focuses on high interconnect and memory bandwidth and offers data types suitable for AI workloads.
- There are only a few players capable to power large scale supercomputers on the low-latency interconnect network (> 5000 nodes): Mellanox (IB), Atos (BXI, the only

European Inter-Node Interconnect), HPE Former Cray (Aries) and HPE Cray (Slingshot). Low latency Ethernet is a promising technology that will have to demonstrate its capabilities on the field.

- More advanced intra-node interconnect will allow to build a tighter integration between CPU and GPU (cache memory coherency, limit data movement between CPU and GPU, open to design an accelerated node with DDR-less CPU and memory consumed directly from GPU/accelerators' HBM). It will also help to support the adhesion of MCM design to go beyond the current process manufacturing limits and more powerful chips. Openness of intra-node interconnects will be key for a wide adoption and ensure hardware interoperability as ease software programming.
- While further near-terms developments might include merging CPUs and GPUs onto a single die to build an APU dedicated to HPC (and AI), longer-term investment could be quantum computing with a first approach, to another type of hybrid systems based on current computing technologies and quantum accelerators and/or simulator.
- Future Exascale systems target 50 GFlops/W to sustain a high energy efficiency. This ratio should be achievable in 2023 timeframe, either through heterogeneous architectures based on accelerator computing capabilities combined with ARM processors or through a future well-balanced ARM processor design with enhanced capabilities (AI, edge computing, etc.).

The complete technical report can be accessed via [6] on PRACE website or [7] on Zenodo.

3.4 Quantum Computing – A European Perspective

Quantum Computing (QC) is expected to bring a new revolutionary component to the high-performance computing (HPC) palette. It is expected to have an impact on practically all fields of science, research, development, and innovation that utilise, or *could* utilise computational modelling. When sufficiently mature, quantum computers can tackle problems that due to their size and complexity will forever stay beyond the reach of conventional computing alone.

Incorporating QC into existing supercomputing infrastructure is essential. For real-world problems, quantum computers will never *replace* classical computers, but instead become an integral part of HPC. Europe has a unique opportunity to create world-leading supercomputing infrastructures incorporating quantum technology by capitalising on the established expertise of European HPC centres in conjunction with the European quantum technology ecosystem. Efforts should be pan-European, following the already proven success of the EuroHPC JU for setting up pre-Exascale computing facilities, and the plans for going to Exascale and beyond. This requires dedicated support for quantum hardware and software developments, as well as for education. Coordinated efforts for catalysing early adoption of quantum computing in academia and industry are essential.

As with the traditional supercomputing infrastructure, diversity is key when setting up a distributed European HPC+QC backbone. Endeavours should be concerted, but not overly concentrated. As QC technology is still in its early stages, it is essential to support different strategies for merging HPC and QC in Europe. Two main approaches for implementing HPC+QC that are suitable for the widest selection of algorithms, applications, and use cases exist: (1) the co-located approach, where quantum computers are placed in physical proximity to traditional HPC resources; (2) the more general distributed approach, where the quantum computers and the HPC infrastructure can be separated. Setup of both approaches should commence immediately. The connection to the European efforts to set up a quantum internet, which in the future could connect and pool quantum computers over a distance should be actively maintained.

A specific challenge is to ensure commercial competitiveness on the global scene. In Europe, the industrial quantum ecosystem is largely start-up driven. Sustained, sufficiently long-term EU-level support for the emerging QC industry is needed to ensure that European competence and capacity in the field reaches a critical mass of self-sustainability and sovereignty.

Europe needs to raise its goals for quantum computing sufficiently high to keep up with global developments. Still, we need to remember that quantum computing is in its infancy. To reach the number and quality of qubits needed for truly disruptive quantum computing, the roughly one hundred noisy qubits that we have today have to be scaled up by four to five orders of magnitude. Significant support for basic research into quantum technologies is required. Only then can European competitiveness in the field be ensured. The basic research on quantum technologies performed for decades in partnership between all European countries, whether members of the European Union or not, has laid the seeds for the fruits we reap today. Global collaboration needs to continue, with minimal restriction, so that quantum computing can be harnessed for solving pressing challenges of our society as soon as possible.

The complete technical report can be accessed via [8] on PRACE website or [9] on Zenodo.

3.5 Security in an Evolving European HPC Ecosystem

The goal of the technical report was to analyse challenges and requirements related to security in the context of an evolving European HPC ecosystem, to provide selected strategies on how to address them, and to come up with a set of forward-looking recommendations. A key assumption made in this technical report is that we are in a transition period from a setup, where HPC resources are operated in a rather independent manner, to centres providing a variety of e-infrastructure services, which are not exclusively based on HPC resources and are increasingly part of federated infrastructures. Furthermore, the changing risk assessment due to an increased risk of cyber-attacks as well as the need for a higher level of protection, e.g. in the context of the processing of sensitive data, has to be taken into account.

The technical report documents and discusses a selected set of approaches to improve security in the context of infrastructures that comprise HPC resources. This includes improving security for SSH-based access to HPC systems, making the design of networks within HPC centres more secure, and introduction of monitoring and intrusion detection systems (IDS). Based on a survey circulated to PRACE members it seems that IDS solutions are not yet widely used.

Due to the trend towards federation of services that are in parts based on HPC resources, security issues cannot be addressed at the level of a single data centre anymore, but rather need to be addressed at a federation level. One strategy to enhance security and establish common security levels is to leverage standards. The technical report therefore includes an analysis of different security standards that are relevant in this context.

The technical report formulates the following recommendations:

- Data centres should review their security strategies within the evolving European HPC ecosystem, where an increased number of services are provided beyond providing access to a supercomputer.

- With SSH being currently the most widely used network protocol for connecting to an HPC system, SSH configurations should be hardened for security, e.g. by enforcing DNS hostname checking, disabling password-based authentication, restricting access through white-listing methods, or by using SSH configuration scanners.
- Use of Intrusion Detection Systems should be further explored as relatively few sites seem to use these today.
- Security standards for HPC centres should be adopted at European level and should be leveraged (or even be established) to realise a common security level within a European infrastructure where services start to be federated, which would improve the usability of this infrastructure by users with specific security requirements, e.g. in the context of processing of sensitive data. The C5 catalogue from the German Federal Office for Information Security is a promising starting point, because it prescribes concrete measures.
- The collaboration between HPC centres in Europe should be strengthened to improve the response to security incidents, which in future are even more likely to affect more than one site. Such collaboration would also allow to harmonise security measures to avoid users having to deal with different security restrictions.

The complete technical report can be accessed via [10] on PRACE website or [11] on Zenodo[10].

4 Conclusion

The series of market watch deliverables has been initiated in PRACE-4IP and has been continuously extended since then.

Within PRACE-6IP the scope and focus of the technology watch was refined in collaboration with INFRAG, RIAG and EXDCI-2, to include topics of their interest and avoid overlaps with their work and it was presented in the form of separate technical reports:

- *"Edge Computing: An Overview of Framework and Applications"*, published in December 2020
- *"Data Management Services and Storage"*, published in December 2020
- *"State-of-the-Art and Trends for Computing and Interconnect Network Solutions for HPC and AI"*, published in December 2020
- *"Quantum Computing - A European Perspective"*, published in October 2021
- *"Security in an Evolving European HPC Ecosystem"*, published in November 2021
- *"User Requirements influencing HPC Technologies"*, will be published in January 2022

These reports are published on the PRACE website and Zenodo.

5 Annex 1: Technical reports

Available on-line at www.prace-ri.eu

Partnership for Advanced Computing in Europe

Edge Computing: An Overview of Framework and Applications

Ezhilmathi Krishnasamy[a,*], Sebastien Varrette[a], Michael Mucciardi[b]

[a]University Du Luxembourg, SnT, UL HPC, Luxembourg.
[b]Barcelona Supercomputing Center, Spain.

Abstract

This report gives an overview of the Edge Computing paradigm and its applications. Indeed, with the advent of the Internet of Things (IoT) era, many electronic devices and sensors produce a vast volume of data which should be processed in a timely manner and this novel computing model is nowadays seen as a pertinent answer to this open challenge. This report thus explains why Edge Computing is needed and how the edge architecture is typically structured. It further presents the technologies that help this cutting-edge model to function properly. Since Edge Computing involves a heterogeneous architecture, it requires to adapt to a few technological recommendations for optimal performance. In this context, this report reviews the latest hardware technology trends tied to Edge Computing developments, and points out technical challenges implementing this innovative computing model. In particular, we analyse how High Performance Computing and Cloud Computing infrastructures can be efficiently organised to design an Edge Computing-based framework able to tackle cutting-edge issues solved by Artificial Intelligence techniques. Finally, this report presents selected real-world applications of the Edge Computing paradigm across multiple domains affecting our daily life, i.e. healthcare, smart city and grids, industry 4.0 and public safety.

[*] Corresponding author, e-mail: ezhilmathi.krishnasamy@uni.lu

Contents

1. Introduction

This technical report is part of a series of reports published in the Work Package "HPC Planning and Commissioning" (WP5) of the PRACE-6IP project. The series aims to describe the state-of-the-art and mid-term trends of the technology and market landscape in the context of High Performance Computing (HPC) and Artificial Intelligence (AI), Edge-, Cloud- and Interactive Computing, Big Data and other related technologies. It provides information and guidance useful for decision makers at different levels: PRACE aisbl, PRACE members, EuroHPC sites and the EuroHPC advisory groups "Infrastructure Advisory Group" (INFRAG) and "Research & Innovation Advisory Group" (RIAG) and other European HPC sites. Further reports published so far cover "State-of-the-Art and Trends for Computing and Network Solutions for HPC and AI" [1] and "Data Management Services and Storage Infrastructures" [2]. The series will be continued in 2021 with further selected highly topical subjects.

Edge Computing aims to process the data very close to the source where it is produced. Many electronic devices are currently connected to the Internet of Things (IoT), which will produce a massive volume of data, and it might be even larger with mobile phones in a 5G network. Cisco Global Cloud Index estimated in 2019 that IoT devices will generate around 500 zettabytes of data [3]. Furthermore, data traffic will be approximately 10.4 zettabytes, which is up from 3.4 zettabytes in 2014 [4]. Moreover, by 2020, 50 million streaming IoT devices will be in use [5].

Edge Computing refers to processing the data on the device and very close to the device. This massive volume of data might be hard to handle entirely on the Cloud Computing network. To face this challenge, Edge Computing offers the full computation or part of the computation that can process the data at the Edge network, which is in very close proximity to the data source. It enables low latency, faster response, and more comprehensive data analysis.

Usually, devices connected to the IoT provide the service in healthcare, smart cities, smart grid, transportation, multimedia, and security. In general, those services depend on AI methodologies, which are compute-intensive and use massive data. A few years back, these devices usually sent the data to the cloud or local data centre to process the data. With ongoing development in Edge Computing, the part of the data at the Edge node can be processed, thus minimising the application's overall latency.

The rest of this technical report is organised as follows: Section 2 focuses on why Edge Computing is needed. Section 3 gives a quick overview of the difference between Edge Computing and Cloud Computing. Section 4 explains the technologies and architectures that are available for Edge Computing. Furthermore, virtualisation, resource management, and development for Edge Computing platforms and how these categories define the proper working functionality of Edge Computing are described. Section 5 focuses on Edge Computing's latest architecture trends, giving more detail about Nvidia Jetson, Raspberry PI, Tinker, and Kalray MPPA. Section 6 presents how Edge Computing still depends on supercomputers/data centres or Cloud Computing for data-intensive applications, particularly AI methodologies. Section 7 focuses on the applications of Edge Computing in real-world applications. Four applications are categorised with more examples in each category, such as healthcare, smart city, industrial applications, smart grid, and public safety. Finally, Section 8 presents a few challenges while implementing Edge Computing, specifically about naming and programmability.

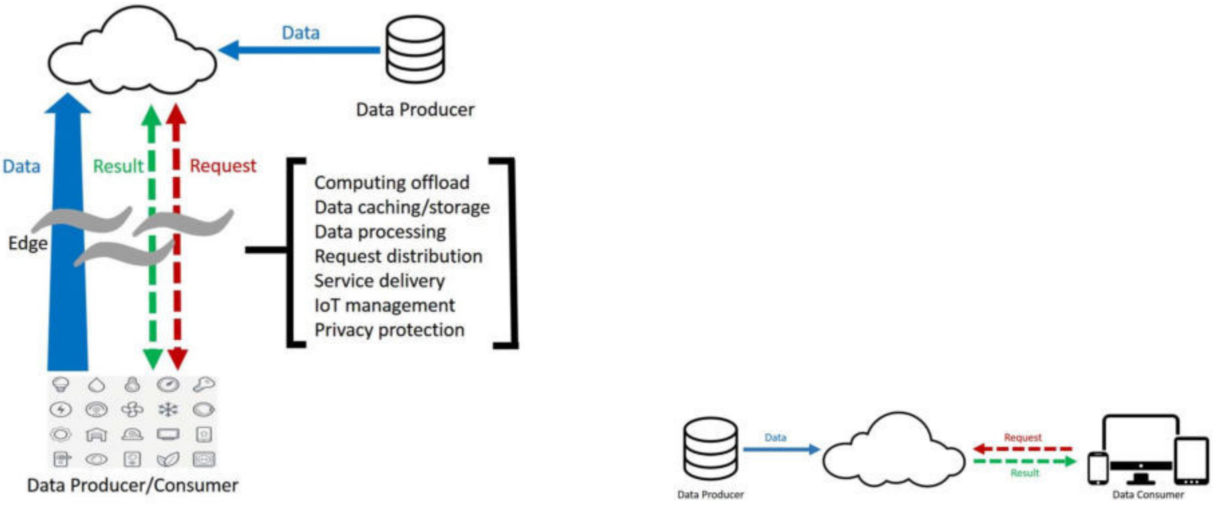
2. Why Edge Computing?

Currently, the entire world is going towards digitalisation, and lots of data is produced in various fields. Moreover, in most cases, this data needs to be processed in a short time to facilitate the present technology (real-time applications). A few years back, cloud technologies have been introduced, gradually reducing the need for small- and medium-scale companies and research institutes to own a computer to do the computations. Nevertheless, the end-users still need to send and receive the data to and from the location where the machine is located. In contrast, Edge Computing is an alternative option for doing computations where the data is located, and is especially suited for real-time applications.

In particular, part of the IoT might require short response time, private data, and Big Data, which could be challenging for the network. However, Cloud Computing cannot handle few of these challenges. Figure 1a and 2 show the paradigm and schematic model of Edge Computing.

Edge Computing is not a direct competition to Cloud Computing or supercomputers, but it is certainly sharing computational burden with cloud technology and supercomputers. If the present trend continues, more robust and energy-efficient small/embedded machines will improve the computations in the future. The following items provide information about why Edge Computing is needed:

- ***Push from the Cloud Services:*** In general, Cloud Computing has proven to be very efficient in terms of computation, but in some situations, there has to be an alternative solution to avoid data transfer bottlenecks. Edge Computing is solving this problem. For example, a Boeing 787 produces 5 Gigabytes (GBs) of data every



(a) Edge Computing Paradigm.

(b) Cloud Computing Paradigm.

Figure 1: Workflow Model of Cloud and Edge Computing Paradigm [7].

minute [6], and transferring this data to a satellite or the ground is not efficient for data processing. Yet another example could be autonomous vehicles, and its information needs to be processed very frequently to steer the vehicle in the right direction. It is not feasible to process these data in the cloud as the bandwidth of the network is a bottleneck. Moreover, Edge Computing consumes less energy compared to cloud technology due to the minimal consumption of embedded devices.

- ***Push from the Internet of Things:*** Presently, electronic devices such as LEDs, surveillance cameras, and air quality sensors are part of the IoT, and they produce and consume a lot of data. In future, there will be even more electronic devices that will be connected to the IoT. It is not feasible to process all the data in the cloud due to the bandwidth and latency. This means some of the data need to be processed at the level of Edge devices. Moreover, privacy is a big concern for cloud solutions, Edge Computing can minimise this concern by restricting the data within the Edge. Figure 1b shows the traditional cloud computing paradigm, where raw data is produced and transferred to the cloud and consumers are sending the request to access the data from the cloud. Basically, this structure is not optimal since a large amount of data needs to be transferred, and in some situations, data privacy is also a concern.
- ***Change from a Data Consumer to a Producer:*** A device at the Edge not only consumes the data from the cloud, but it also produces the data and uploads the data to the cloud. Watching a YouTube video from your mobile phone, using Facebook and Instagram, are examples where Edge users pull the data from the cloud. At the same time, Edge devices produce the data, such as taking pictures or recording videos. When the Edge users try to upload this data to the cloud, it could be a lot of data depending on the resolution. This would occupy even more bandwidth for the uploading. In a situation like this, the resolution can be adjusted at the Edge device before uploading the data to the cloud.

3. Comparison between the Cloud and Edge Computing Paradigms

Edge Computing is a paradigm relying on similar concepts deployed within Cloud models. In Cloud Computing, either in public, private or hybrid types of accessibility models, data processing or computation occurs at the data centre, where it has a substantial computational resource and data needs to be transferred back and forth. The cost model that made this paradigm so popular since the last decade is that the end-users only pay for the resources they used, whether in terms of computing, storage or data transfer capacities. In practice, the following deployment models are traditionally considered within the Cloud Computing paradigm [9]:

SaaS "Software as a Service", this refers to using the existing software or applications from the cloud, for example, using Gmail, Office 365, etc. Here, a cloud provider is controlling software or applications, and end-users use the software.

PaaS "Platform as a Service", this refers to using your software but using the cloud resource as hardware, for exam-

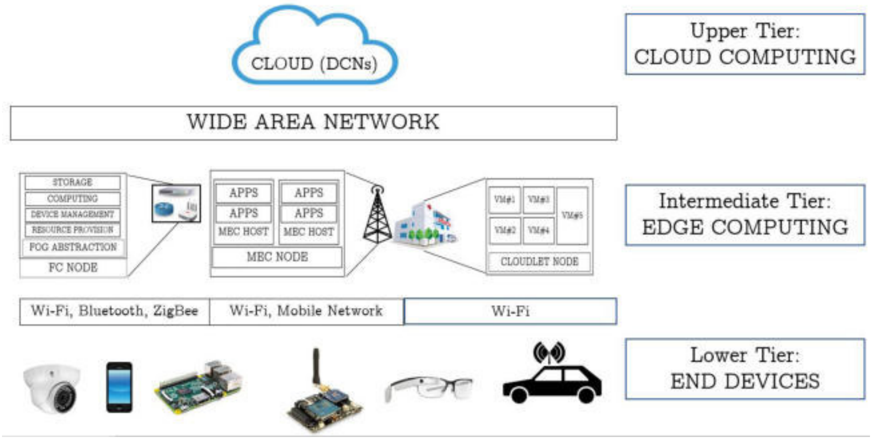


Figure 2: Edge Computing Example [8].

ple, using the cloud's hardware and operating system. Here, a cloud provider offers middleware, development tools, operating systems, hardware and other business tools for end-users or customers.

IaaS "Infrastructure as a Service", refers to providing infrastructure, such as computing, storage, and cloud technology to the end-users. Here, end-users can scale down or scale up their computational platform as they want.

Several derived models were recently proposed, such are HaaS (Hardware as a Service), but their developments are considered out of scope for this report. In all cases, Figure 3 illustrates the available Cloud service models. Furthermore, Cloud technologies lead to several benefits for customers: 1) no need to employ anyone to maintain the computer, 2) no need to update the hardware and 3) minimisation of overall processing costs. At the same time, this model exhibits some drawbacks as well, which are: 1) slow network connectivity and internet traffic which make cloud technologies slower 2) security concerns, and 3) local data which can be stored in foreign countries, not subjected to the same data protection regulation as the GDPR in European countries.

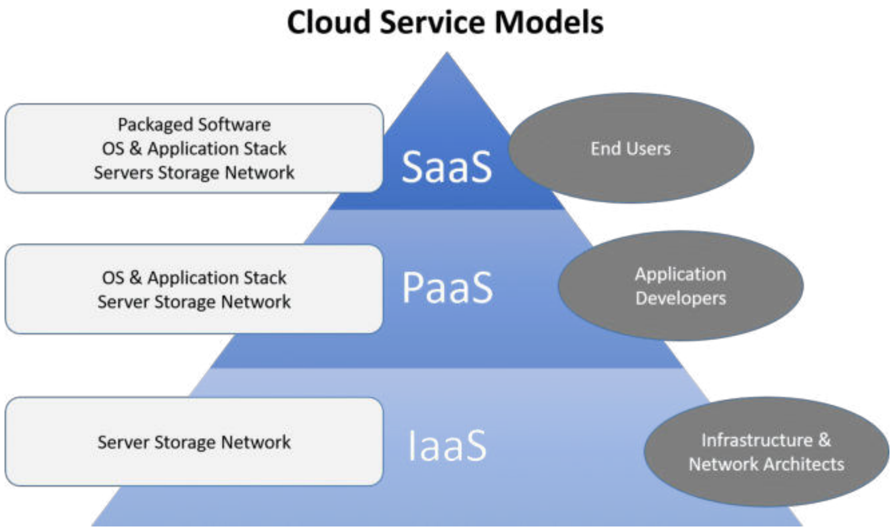


Figure 3: Overview of the main Cloud Computing deployment models [10].

It follows that even though Cloud Computing may be considered faster for data processing or computation, data transfer is a major bottleneck for Big Data analytics workflows. This is particularly relevant for IoT [11, 12, 13] environments, which tend to produce huge volume of data across interdisciplinary disciplines like technology, healthcare, environment, and transportation. In this case, a novel distributed and large-scale computing paradigm is required to effectively treat and analyse such large-scale dataset in a timely manner. That is how the Edge Computing model was introduced, with as its heart the idea to bring data storage and compute power closer to the device or data source where it is mostly needed. More specifically, the Edge Computing paradigm allows computing resources and application services to be distributed along the communication path, via decentralised computing infrastructures organised to treat in a hierarchical fashion the data analytic workflow. The hierarchy coupled with the distribution of computing capabilities aims at solving the bandwidth bottleneck identified for general Cloud architectures.

4. Edge Computing Architecture and Technology

4.1. Edge Computing Architecture

Edge Computing architectures are traditionally composed by several layers playing an essential role in the successful execution of the associated paradigm. Figure 4 and Table 1 show the schematic architecture and characteristics of an Edge Computing environment, which are thus categorised according to the following deployment models:

- ***Cloudlet Computing.*** This refers to computing resources (small cluster) connected via WLAN to the end-users. In general, it can be considered as a "data centre in a box" which provides support (computing and storage) to the end-users over the WLAN network. Cloudlet Computing is based on three layers: the component layer, the node layer, and the cloudlet layer. This is designed to have higher bandwidth, thus lowering the latency for the applications.
- ***Fog Computing***, a decentralised computing resource that can be placed anywhere between the cloud and the end-users. It is based on the so-called Fog Computing Nodes (FCNs) [14]. All of these FCNs are heterogeneous, including switches, routers, and access points. FCNs heterogeneous environment facilitates the devices at different protocol layers and non-IP based technologies to communicate between the FCNs and the end device. These FCNs are hidden for the end-users, thus ensuring security.
- ***Multi-access Edge Computing (MEC)***, which refers to implementing Edge Computing within the Radio Access Network to reduce the latency. Formally known as Mobile Edge Computing, it is an ETSI-defined network architecture located closer to the Radio Network Controller or macro base station. The edge orchestrator organises the MEC, provides network information about load and capacity, and offers information to the end-users about their location and network information.
- ***IoT* (Internet of Things)** contains a large set of devices and sensors that produce a huge volume of data. These also exchange the data through a modern communication network and monitor and control the infrastructure. Typically, end-users at the Edge use the IoT devices and sensors.

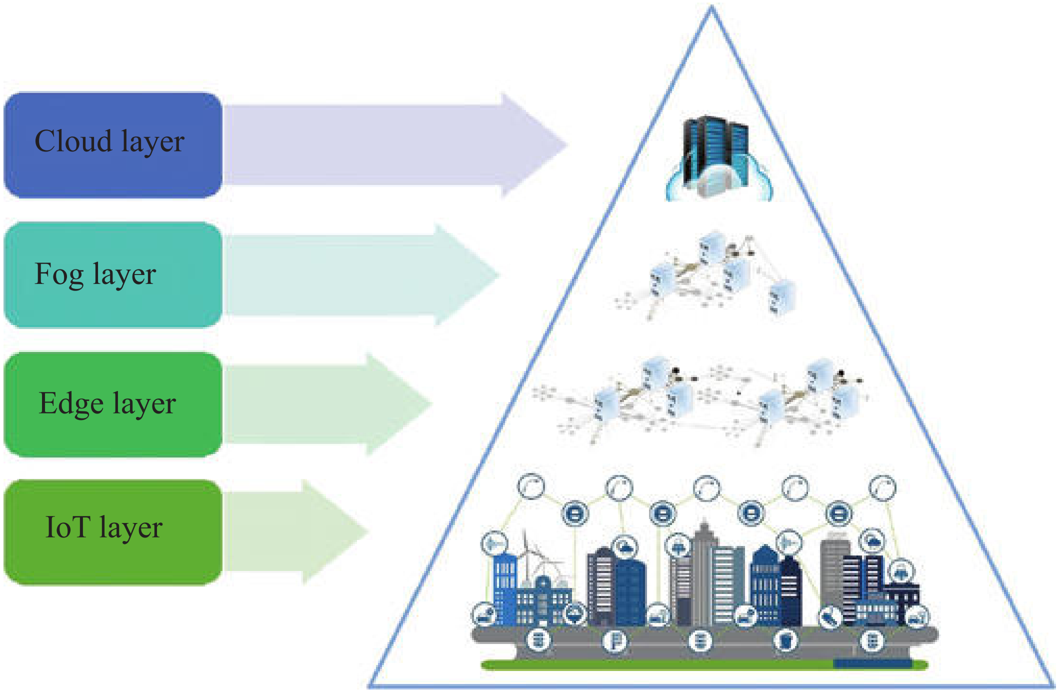


Figure 4: Overview of Edge Computing Architectures [15].

	Edge Computing Architecture Layer			
Characteristics	**IoT**	**Edge**	**Fog**	**Cloud**
Deployment	Distributed	Distributed	Distributed	Centralised
Components	Physical devices	Edge Nodes	Fog Nodes	Virtual resources
Location awareness	Aware	Aware	Aware	Aware
Computational Limits	Limited	Limited	Limited	Unlimited
Storage Limits	Very limited	Limited	Limited	Unlimited
Data	Source	Process	Process	Process
Distance to data source	The source	The nearest	Near	Far
Response time	No response time	The fastest	Fast	Slow
Nodes count	The largest	Very large	Large	Small

Table 1: Main characteristics and functionality within Edge Computing Architectures [15].

In general, Edge Computing involves complex or heterogeneous architecture. It is hard to ultimately make use of this complex architecture for some Edge Computing applications. However, many software platforms help to make Edge Computing work correctly and effectively. The following list provides more description about some of the important software platforms in Edge Computing.

4.2. Virtualisation

Virtualisation refers to an abstraction of an Operating System (OS), computing resources, storage device, and/or network devices. Especially in computing, the term *virtualisation* often implies the reference to the creation of a Virtual Machine (VM) managed by a *hypervisor*, a middleware responsible for providing an abstraction or emulation layer from the hardware. The hardware running the hypervisor is called the host whereas all emulated VMs running inside them are referred to as *guests*. In practice, there exists two types of hypervisors illustrated in Figure 5b:

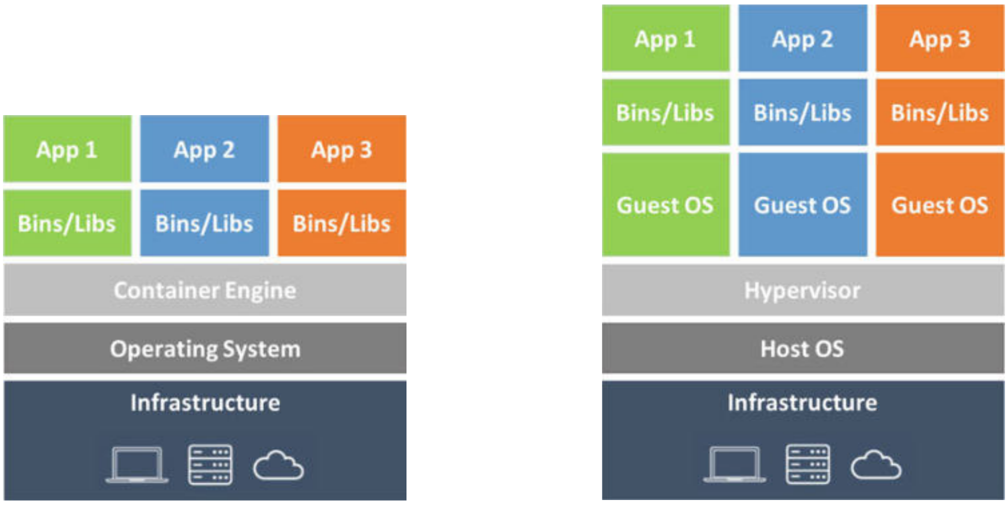
1. *Type-1 (Native) hypervisors* are running directly on hardware (hence often referred to as bare metal hypervisors). Xen [16], VMware ESXI [17] or Hyper-V [18] are examples of such hypervisors.
2. *Type-2 (Hosted) hypervisors* require a host operating system whose capabilities are used in order to perform virtualisation operations and emulations. Examples of such hypervisors includes KVM [19] or VirtualBox [20].

With the advent of Docker [21] in 2013, another virtualisation model became popular: *container*-based virtualisation, which does not emulate an entire computer. Instead, the host operating system is providing most features to the container software in order to isolate processes from other processes and containers as depicted in Figure 5a. It is indeed a built-in feature of the Linux kernel to provide such isolation capabilities to isolate processes. Other operating systems may provide similar mechanisms, such as FreeBSD's jails [22]. This allows to design lightweight images - a code-based file that includes all libraries and dependencies, which makes this technology particularly suitable for the development of Edge Computing architectures. Because there is no full emulation of hardware, software running in containers has to be compliant with the host system's kernel and CPU architecture. Furthermore, since containers are so small, there are usually hundreds of them loosely coupled together, which is why *container orchestration frameworks* such as Red Hat OpenShift [23] and Kubernetes [24] (depicted in the next section) are used to provision and manage them.

Virtualisation is of course considered in other components of large-scale computing infrastructures relevant for Edge Computing. At the level of the network backbone, network virtualisation comes under the form of *Software-Defined Networks* (SDN), a manageable and cost-effective approach to enable dynamic, programmatically efficient network configuration in order to improve network performance and monitoring. This makes this type of architecture suitable for the high-bandwidth and dynamic nature of the applications that run on top of Edge Computing devices. SDN decouples the forwarding process of network packets (or data plane) inherent to the physical network, from the routing and control process. Often composed by one or more controllers, the control plane is generally referred to as "brain of the network", where most of the crucial decisions are made as illustrated in Figure 6. Openflow [25] and OpenvSwitch [26] are a well-known Open-Source implementation frameworks in SDN network virtualisation.

4.3. Resource Management and Edge Orchestration

Resource management is crucial in terms of allocating computing resources, CPU, memory, storage, and network on standard traditional HPC settings. Especially for Edge Computing, the energy resource is very important. For example, a mobile user does not want to recharge his or her smartphone often, and also some sensors should not need



(a) Container Environments [7].

(b) Type-1 (Bare-Metal) or Type-2 (Hosted) Hypervisors [7].

Figure 5: Main types of virtualisation frameworks.

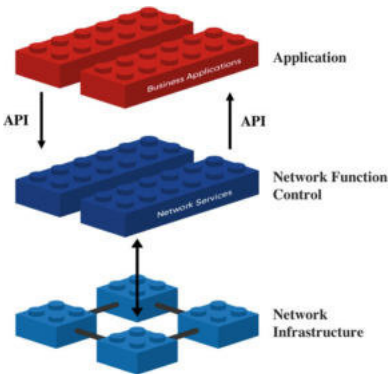


Figure 6: A high-level architecture overview of SDN [7].

to be charged frequently. In particular, for the Edge analytics, developers should know how much power is needed for the specific application and how data could be pulled/pushed from the device. Eventually, this kind of problem falls into the optimisation problem of scheduling and workload placement. In general, there are some important requirements for Edge platforms management, summarised in Table 2.

The orchestration of Edge Computing resources is thus quite challenging and still a work in progress. We can cite several frameworks for which an adaptation to the management of Edge devices is on-going:

- Generic container orchestration middlewares such as OpenShift [23] (Red Hat) or Kubernetes [24] (Google). Figure 7 depicts the generic working model architecture. In particular, Kubernetes features the following characteristics:
    - It is automatically bin-packing the containers based on the containers' resources and other constraints. For example, in Edge Computing, the computing resource of the Edge might host multiple users and applications. It is important to schedule them in an optimised way to use the Edge resources to minimise the energy consumption and achieve lower latency.
    - Kubernetes has a feature of self-healing, which is a significant factor for Edge Computing. For example, when there is a sudden failure or nodes are killed, users do not want to get notified; instead, the service and application should be recovered/migrated to other resources.
    - Another feature of Kubernetes is load balancing, which is based on the IP address and a single DNS name for a set of containers.

Requirement	Description
Scalability	Ability to address a large number of Edge devices of different type and capabilities with appropriate deployment and communicating protocol.
Security	Privacy preserving for security tokens and support for integrity checks within the infrastructure
Heterogeneity	Support for a high degree of heterogeneity within hardware/software
Volatility	Support for volatile availability and mobile hardware/software components
Data Protection	GDPR Compliance, ensure all data is kept locally and on-the-fly encrypted
Infrastructure Performance	Very low latency, lightweight publish-subscribe network protocol as MQTT [27]. High performance containerised resources with fast (Zero-touch) provisioning allowing easy system upgrades
Application Portability	Unified architecture view via MEC compliance enabling Function as a Service (FaaS) capabilities
Data Analytics	Supports for Data Management and Data Analytics Pipeline Engine.

Table 2: General requirement for Edge devices management and orchestration.

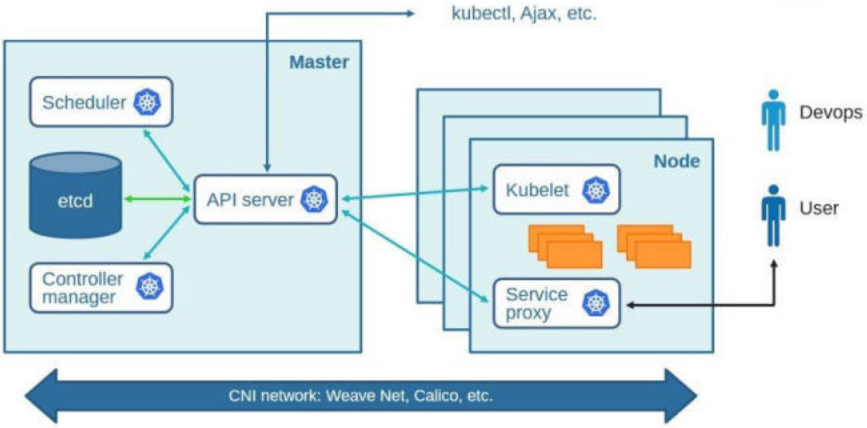


Figure 7: A high-level architecture overview of Kubernetes [7].

Among other initiatives, the KubeEdge [28] project is dedicated to making an open platform, which is built upon Kubernetes and provides fundamental infrastructure support for network, application deployment and metadata synchronization between Cloud and Edge architectures. To enable fast and lightweight communications between all the Edge devices, i.e. from low power single board computers to full-capable servers and HPC compute nodes), the MQTT [27] protocol implemented over the Mosquitto [29] message broker service is used.

- ONAP (Open Network Automation Platform) [30], a comprehensive platform for orchestration, management, and automation of network and Edge Computing services for network operators, cloud providers, and enterprises. More tailored to MEC management and allowing to orchestrate physical and virtual network functions synchronously. The ONAP project provides a unified operating framework for vendor-agnostic, policy-driven service design, implementation, analytics and lifecycle management for large-scale workloads and services. A high-level overview of the ONAP platform architecture is proposed in Figure 8.

4.4. Developing Platform enabling Data Analytics for Edge Computing

Edge Analytics refers to processing the data of the crucial application and service entirely or partially. For example, in healthcare, monitoring the older adults at home using the ECG (ElectroCardioGram) or EEG (ElectroEncephaloGram), it is better to perform the data processing at home, and only if something abnormal happens send out the data to the hospital or the doctor. Thus, the data flow between home and hospital is minimised, which reduces the data failure between two points. Not only the healthcare system can leverage this model, nowadays, but electronic devices which are connected to IoT networks can also process the data and exchange the essential information. For example, such devices can be drones, robots, cameras, and sensors, etc.

As mentioned in Table 2, data analytics capabilities are required within the development tools and platforms, enabling an Edge Computing infrastructure. If the Edge appliances cannot process the data, then it has to send the data to the cloud. In this context, there are several open-source tools available as software platforms supporting the data analytics pipeline engine for Edge Computing.

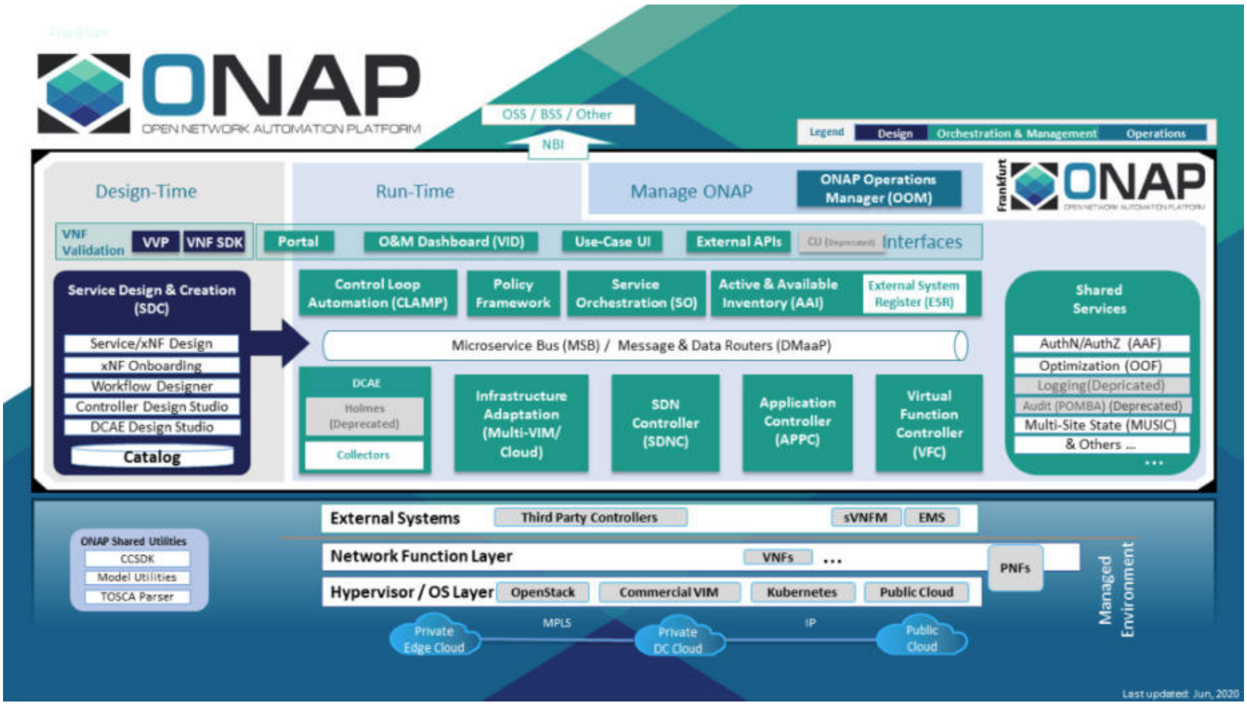


Figure 8: A high-level architecture overview of the Open Network Automation Platform (ONAP) [30].

- TensorFlow [31], a framework that supports Machine Learning (ML)/AI computations. It also supports a wide range of architectures, including, CPUs, GPUs, embedded and mobile systems. TensorFlow supports different programming languages, such as C/C++, Python, and Java. It even supports heterogeneous capability (CPU+GPU) computation.
- OpenCV [32], a software platform that supports computer vision computations. It has more than 250 optimised algorithms that can detect and recognise, i.e. faces, identify objects, classify human actions in videos, etc. OpenCV also provides an API for Java, C/C++, and Python.
- Apache Edgent [33], an analytic tool that runs on the Edge device based on the Apache incubator project. It allows to store fewer data on the Edge device and limits the amount of data to be transmitted to the analytic server. Apache Edgent supports the Java API, and it makes Edge systems to be more autonomous.
- TensorFlow Lite [34], an open-source and designed for on-device inference, used for the AI/ML/ Deep Learning (DL) applications. In particular, it is used for image classification, object detection, and text classification.
- Apache MXNet [35], supporting a distributed computing platform with up to eight programming language bindings. Libraries in MXNet enables use-cases in computer vision and natural language problems. It is also has a rich eco-system that supports the other AI/ML/DL libraries for Edge Computing, i.e. MXFusion [36], Keras-MXNet [37], and InsightFace [38].

5. Latest Trends in Edge Computing

Edge Computing will have to deal with ongoing developments in AI/ML/DL to do computational analysis and data processing. To do such complex arithmetic calculations, powerful embedded devices are required to perform these calculations are needed. There are lots of embedded hardware solutions available nowadays to support these requirements. Here we explain selected embedded architectures that will make Edge Computing more efficient and optimised.

5.1. Nvidia Jetson

Nvidia has introduced the Jetson series of embedded architectures to support ML calculations for embedded applications. Jetson TK1 was first introduced in 2014. Nvidia has released many series in embedded architecture since then. Nvidia also introduced the tensor cores to support the ML calculations for embedded architecture, which can be seen in Table 3. Figure 9 shows the Nvidia Jetson modules of Nano and Xavier NX. The Jetson Volta architecture has a

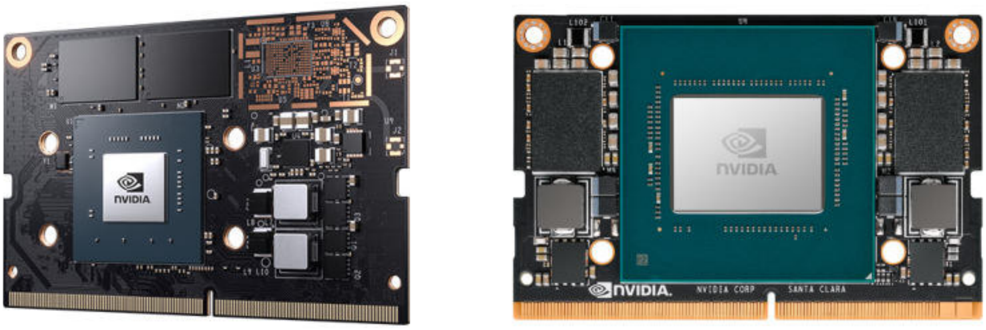


Figure 9: Nvidia Jetson Nano (left); Nvidia Jetson Xavier NX (right) [39].

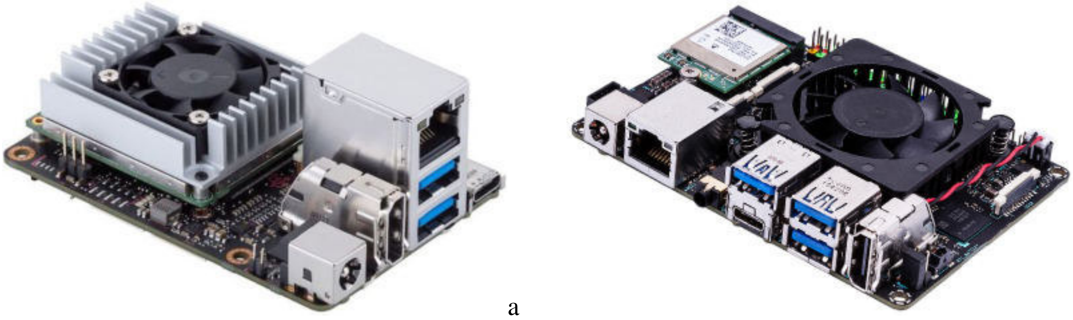
Tensor Processor Unit (TPU), which can do computation on a large volume of data with low precision (which can be low as 8-bit precision). This kind of hardware functionality is quite useful for AI/ML/DL computations.

	Nvidia Jetson			
	Nano	**TX2 Series**	**Xavier NX**	**AGX Xavier**
Architecture	Maxwell	Pascal	Volta	Volta
CUDA cores	128	256	384	512
CPU	Quad-core ARM Cortex-A57 MPCore processor	Dual-Core NVIDIA Denver 2 64-bit CPU Quad-Core ARM Cortex-A57 MPCore	6-core NVIDIA Carmel ARMv8.2 64-bit CPU, 6 MB L2 + 4 MB L3	8-core NVIDIA Carmel ARMv8.2 64-bit CPU, 8 MB L2 + 4 MB L3
Memory	4 GB 64-bit LPDDR4 1600 MHz 25.6 GB/s	8 GB 128-bit LPDDR4 1866 MHz 59.7 GB/s	8 GB 128-bit LPDDR4x 1600 MHz 51.2 GB/s	32 GB 256-Bit LPDDR4x 136.5 GB/s
Storage	16 GB eMMC 5.1	32 GB eMMC 5.1	16 GB eMMC 5.1	32 GB eMMC 5.1
Tensor cores	n/a	n/a	48	64

Table 3: Latest Nvidia Jetson Embedded Architecture Comparison.

5.2. ASUS Tinker Board

ASUS has introduced the ASUS Tinker Board in early 2017. This early embedded architecture can run in 32-bit mode, but the latest ones can run on 64-bit. And they are direct competitors to the Raspberry PI series. Table 4 shows the ASUS Tinker's latest embedded architecture, and Figure 10 shows the architecture model outline. Tinker Edge T has a Google Edge TPU as a coprocessor. The Google Edge TPU is based on an Application-Specific Integrated Circuit (ASIC). This means the TPU can do the specific application, i.e. a digital voice recorder or a high-efficiency Bitcoin miner. On the other hand, Tinker Edge R has an AI-specific accelerator as a coprocessor.



a

Figure 10: ASUS Tinker T (left); ASUS Tinker R (right) [40].

	Asus			
	Tinker Board	**Tinker Board s**	**Tinker Edge T**	**Tinker Edge R**
Architecture	ARMv7-A (32-bit)		ARMv8 (64-bit)	
CPU	Quad core 1.8 GHz ARM Cortex-A17 (up to 2.6 GHz turbo clock speed)		Quad core 1.5 GHz ARM Cortex-A53	Hexa core. 2x Cortex-A72 cores up to 1.8 GHz, 4x Cortex-A53cores @ 1.4 GHz
GPU	600 MHz Mali-T760 MP4 GPU		GC7000 Lite 3D GPU	800 MHz Mali-T860 MP4 GPU
Memory	2 GB dual channel LPDDR3		1 GB LPDR4	4 GB dual channel LPDR4 for system, 2 GB LP DDR3 for NPU
Co-processors	n/a		Google Edge TPU 4 TOPS of performance	NPU 3 TOPS of performance

Table 4: Asus Tinker Board Architecture Comparison.

5.3. Raspberry PI

Raspberry PI is also an emerging embedded architecture in the Edge Computing domain. It has the latest ARM Cortex-A7 CPU and VideoCore GPU. This VideoCore GPU is based on Digital Signal Processing (DSP), which means it can efficiently process multimedia applications with low power consumption. Table 5 and Figure 11 show the latest Raspberry PI architecture and outline.

	Raspberry	
	RPI 3 Model B+	**RPI 4 Model B**
Archictecture	ARMv8-A (64/32-bit)	
CPU	4x Cortex-A53 1.4 GHz	4x Cortex-A72 1.5 GHz
GPU	Broadcom VideoCore IV @ 250 MHz	Broadcom VideoCore VI @ 500 MHz
Storage	8 GB	1, 2, 4 or 8 GB

Table 5: Raspberry PI Architecture Comparison.

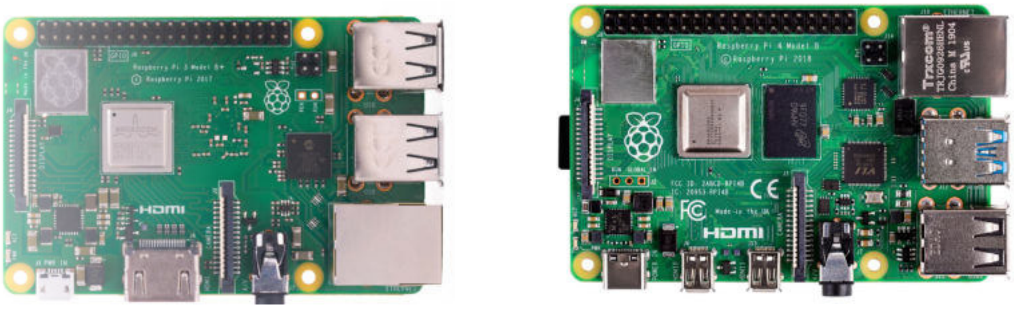


Figure 11: Raspberry 3 B+(left); Raspberry 4 B (right) [41].

5.4. Kalray MPPA

Another impressive embedded architecture is Kalray, which has many CPU cores, unlike other embedded architecture. Kalray named their embedded architecture “Massively Parallel Processor Array” (MPPA). Kalray 3^{rd} generation MPPA architecture is called Coolidge (MPPA3-80 Coolidge), based on FinFET technology with 16 nm size. It has 80 high-performance AI accelerated and fully programmable cores, connected with a 600 GB/s Network-on-Chip and advanced PCIe Gen4 and 200G Ethernet [42]. According to Kalray, the architecture does not have GPU cores comparable with other embedded architectures, yet KONIC200 has a different module, which can support Vision & AI, Storage, SmartNIC and 5G functionality. This can be seen in Figure 12. Table 6 shows the computational performance of AI/ML software. Here, TOPS refers to “Tera Operations Per Second”, which defines the AI computational performance (INT8) on the given architecture. KONIC200 reports good AI performance even though it does not have a GPU, concluding that Kaltay MPPA is probably suitable for Edge Computing.

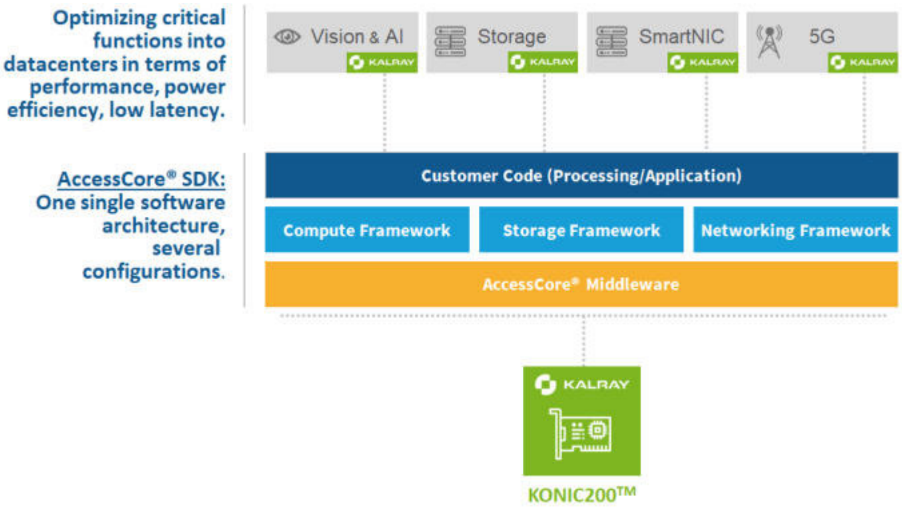

Optimizing critical functions into datacenters in terms of performance, power efficiency, low latency.
Vision & AI
Storage
SmartNIC
5G
KALRAY
AccessCore® SDK: One single software architecture, several configurations.
Customer Code (Processing/Application)
Compute Framework
Storage Framework
Networking Framework
AccessCore® Middleware
KALRAY
KONIC200™

Figure 12: Modular Software in KONIC200 [42].

CNN Model	KONIC200 – FH/LP (1 x MPPA (Coolidge) 80 cores)	KONIC200 - HP (2 x MPPA (Coolidge) 160 Cores)
TOPS	25 TOPS (8-bit)	50 TOPS (8-bit)
GoogLeNet	3025 fps	5445 fps
Faster-RCNN (VGG16)	302 fps	537 fps
Yolo v3	310 fps	564 fps

Table 6: AI and Compute Acceleration in KONIC200 [42].

5.5. Example Applications

Several compute-intensive applications are using the above mentioned advanced embedded architectures for Edge Computing. The following list shows a few relevant test-cases:

- Embedded architecture can process cryptographic functions to fix the privacy and security issues in the smart grid [43].
- A simple test has been done on TK1 at the edge for the MapReduce application. It seems the cluster of TK1 at the Edge shows the same throughput as one single traditional x86/64 Intel server while saving significant energy. Similarly, KMeans [44] also tested in three nodes of TK1 against one Intel server. The throughput is similar to each other, but 68% of energy is saved in TK1 compared to Intel [45].
- In recent years, there has been a lot of ongoing development for embedded architecture software development to use the latest embedded architecture. This enables Edge Computing to do real-time computations very quickly and efficiently. Figure 13 shows the available software in Nvidia Jetson, it can support, i.e. DL and computer vision.
- Kalray supports acceleration of CNN, Computer Vision and Math apps up to 1.1 TFLOPS & 25 TOPS. It is easily programmable in C/C++/Open standards. And accommodates complex data flow in parallel and sequential modes [42].

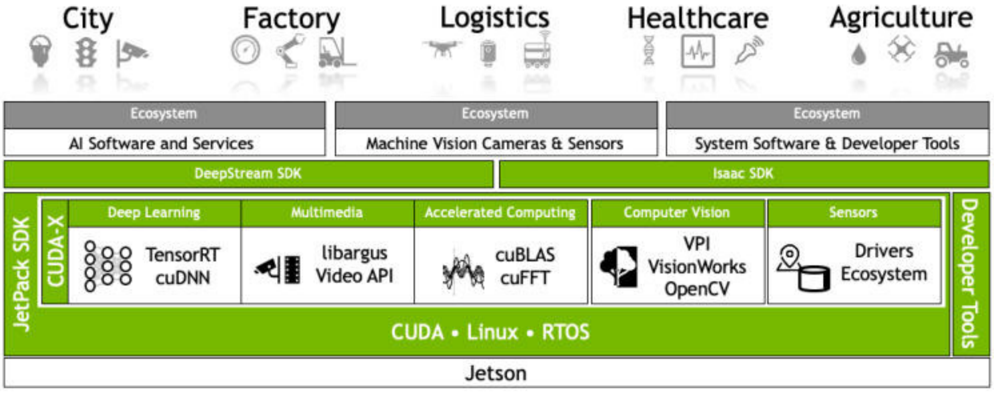


Figure 13: Jetson Software for AI Edge Devices [46].

6. In Reality: Edge vs. Supercomputers/Cloud Computing

As mentioned earlier in Section 5, there is an ongoing development of embedded architectures and Edge Computing based architectures. Moreover, Edge Computing is mainly deployed for AI/ML/DL computations. Nevertheless, almost all of these AI/ML/DL computations are very demanding computationally, based on mathematical and probabilistic modelling. For example, DL has three stages, which are 1) training, 2) evaluation and 3) prediction. In particular, the training part of DL requires significant computational power generally exceeding the capacities of traditional resources from Edge embedded architectures.

In addition, while ML is typically used for language translations, language recognition, autonomous vehicles, computer vision, text generation, and robots, these methodologies require a massive volume of data that needs to be processed at the training stage of DL. Presently, Edge Computing can not handle this massive volume of data, i.e. videos, images, audio files, and text documents) for the training step in DL. Usually, it is best handled by a remote HPC clusters or supercomputers. Once the neural network is trained, it can be deployed at the Edge for prediction, i.e. predicting videos, images, and audio files). In this case, even though Edge Computing is evolving in terms of architecture and software, it still depends on the supercomputer/local cluster (or the equivalent Cloud Computing resources) for the data and computational intensive calculations.

In all cases, the following list depicts a few of the challenges that Edge Computing faces in this configuration:

- *Heterogeneous Data*. Often IoT devices produce different types of data, such as images, text, videos, and sound. Processing mixed data types need a special algorithm and requires more computational power. For example, multimodel deep learning is used for heterogeneous data (to process video and audio). On the other hand, it is not easy to implement this at the Edge device; eventually, this requires a Cloud Computing platform or cluster/supercomputer [47].
- *Distributed Computation*. IoT devices produce a vast volume of data at the Edge. Recently researchers have come up with the concept of an edge-based distributed learning algorithm [48], which is sharing the computation at the Edge and the Cloud, specifically doing less intensive computation at the Edge and high intensive computation at the cloud. This way, the workload is shared between the Edge and the Cloud. Presently, the edge-based distributed learning algorithm is used for fraud detection and market analysis. However, the accuracy and efficiency of the distributed model approach is still an open research challenge [49].

7. Selected Use Cases of Edge Computing enabled Applications

7.1. Medical Applications

- *Overview*: HPC at the Edge for medical imaging merges HPC/AI and medical sensing technology in order to provide precision medicine through the use of real-time advanced monitoring and analysis of a patient's medical data to detect early pathologies while lowering the risk of privacy breaches by keeping the data on site. This granular, yet massive amount of patient data can be analysed at the Edge, transformed, and then only pertinent data is sent to the cloud such as alerts or data stripped of information that could lead to the patient's privacy being compromised. Medical Imaging at the Edge using HPC/AI removes the latency and dependence on Cloud Computing resources, as well as reduces the patient's digital footprint by limiting how many systems have access to data. AI used in medical imaging provides tools that augment the clinician's intelligence in a way where they are able to provide better care at reduced costs [50]. Figure 14 illustrates the digital development in healthcare and how Edge Computing is being used in healthcare.

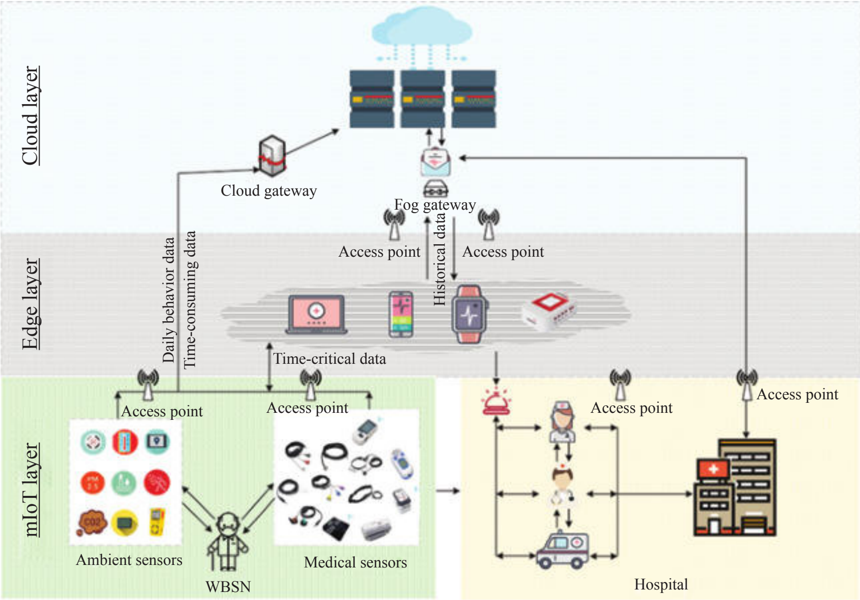


Figure 14: Edge Computing in Healthcare [51].

- *Examples*:

  Case 1: CT/MRI Scanning technology Centre for Clinical Data Science partnering with GE Healthcare, NVIDIA, Nuance, and DASA to build HPC/AI technology that is integrated into CT, MRI, and Workstation machines which utilise AI that is trained on vast data sets of diagnostic medical data [52].

  Case 2: Handheld diabetic retinopathy diagnostic camera Through NVIDIA's virtual accelerator inception program, Taiwanese medical firm MiiS has built a highly portable handheld diagnostic tool that combines NVIDIA Jetson TX2, a high-resolution camera, Edge Computing architecture and GPU-powered AI algorithms to perform instant screening of diabetic retinopathy [53].

- *Security Concerns*: The EU has a strict legal framework in place to ensure consumer protection and to protect personal data and privacy, minimising risks to confidentiality and integrity of data [54]. Moreover, the EU GDPR sets certain rules and regulations for how the data is being processed and handled in EU [55]. In addition, there are specific rules that pertain to sectors such as healthcare that will continue to apply to AI and medical devices: Creation of comprehensive documentation and record keeping of data sets used for training and testing, programming and training methodologies such as the processes and techniques used to build, test, and validate AI systems especially those used to ensure the system is not biased in a way that could lead to prohibited discrimination arising from the usage of AI [56], auditing abilities of how a patient's medical data is being used as well as who is accessing it.

7.2. Smart City

- *Overview*: In the future cities will have sensors that will collect various data, for example, in transportation, medical health, and urban security. Moreover, urbanisation is rapidly increasing. According to the UN, it is estimated that, by 2050, over 6 billion people will be living in the cities [57]. In the future, to have sustainable development in the town, a smart city is an excellent solution. This might help to solve the problems that may arise in food supply, medical care, transportation, culture and entertainment in the cities. These sensors will usually generate a large volume of data, and this data should be processed quickly. Sending these data to the cloud will need faster data movement (latency and data traffic in the network), and privacy. Therefore, these generated data should be processed closer to where it is produced. In general, Edge devices have limited computing and storage, so it is also necessary to integrate multiple computing models. A few cases of Edge Computing used in a smart city are listed below.

- *Examples*:

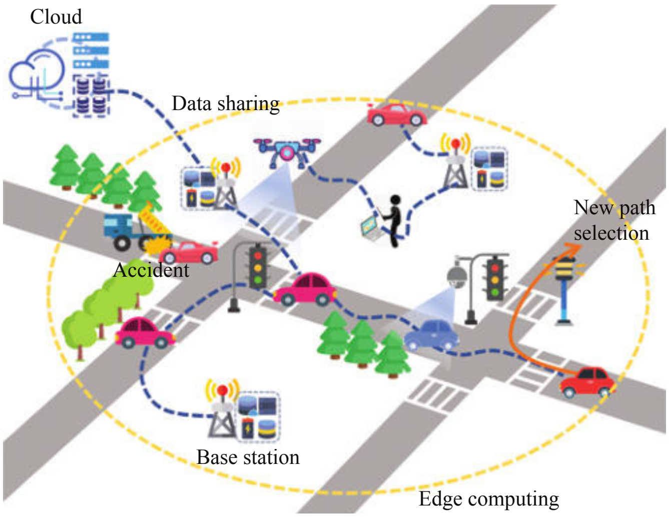


Figure 15: Example of Edge Computing in the Smart City [57].

Case 1: Closed-Circuit Televisions CCTV are nowadays typically installed in almost all private and government premises. These CCTVs will capture the movements of the objects. This will ensure the safety protocol in the given premises. For this reason, the data collected by the CCTVs should be processed quickly. To enable this, CCTVs should be connected to the Edge device through the LAN connection. Edge devices can also use the latest technology called "special image processing chip" to process videos more efficiently [58]. Figure 15 shows an example of a smart city.

Case 2: Smart Home A smart home is controlled by lots of sensors in lighting, kitchen, television, and surveillance, and its technology is also rapidly developing with the help of IoT [59]. Again, to make it more efficient (minimisation of the latency) and effective (privacy of home data), the generated data should be processed quickly where it is generated. Edge Computing will be playing an essential role in facilitating that objective.

- *Security Concerns*: while the user adoption for IoT technologies violating the boundaries of private areas is surprisingly high, as testify for instance by the success of voice assistant like Amazon Echo/Alexa or Google Home, this raises several security concerns for spying intelligence which remains an open challenge.

7.3. Industrial/Manufacturing Applications

- *Overview*: Industry 4.0 combines Edge HPC with AI in industrial automation environments. It aims towards waste reduction, work reduction, and worry reduction in the work space. It is used for connecting machines-to-machines and machines-to-people in a way where on-demand production environments, equipment, and workers can quickly and intelligently react to dynamically changing factory floor/environmental conditions [60]. Certain industrial applications may need to react quickly to real-time changing environmental conditions which may be uncovered in data too voluminous to be sent to the cloud, such as image recognition data that guides a robotic arm to interact with an object on a moving assembly line or creates alerts if dangerous conditions arise. Moving data offsite for analysis may also incur transmission latencies, that exceed the reactions times required for industrial applications, such as being able to shutdown an assembly line if a foreign object interferes with the industrial process. This is all assuming that the industrial site is even able to acquire a high-speed network connection due to geographic constraints.

- *Examples*:

Case 1: Welding Quality Assurance : Rexroth, a Bosch company, has developed a Weld Spot ML analytics software which operates on a smart Edge that is located close to welding controllers. Until now, destructive tests were the only reliable way to test if a spot weld was done according to quality specifications. Weld Spot analytics software contains an AI engine which employs ML algorithms to detect anomalies and provides this information to welding engineers through a UI that displays a variety of data and analyses [61].

Case 2: Worker Ergonomics safety A novel real-time spatio temporal Pyramid Graph Convolutional Network trained on video of warehouse workers performing typical repetitive duties and their associated movements was integrated with a traditional ergonomic risk index to assess the potential of musculoskeletal disorders in the warehouse [62]. This system can send alerts and warnings based on video analysis of actions that are above a certain risk threshold [63].

- *Security Concerns*: Limiting the amount of data sent to the cloud for processing also limits your attack surface from malicious actors. There is a much lower risk of your data being intercepted, tampered with, and stolen when it stays on-premise. Networking automated machinery poses a threat where compromised machines can be manipulated either directly or by tainting real-time and training data that your algorithms are using to control manufacturing processes [64].

7.4. Smart Grid and Public Safety

- *Overview*: Electricity is one of the primary sources for humans to conduct most of the activities in daily life. In recent years, special emphasis has been placed on how electricity is produced and distributed to facilitate better economic, technical, and environmental reports. In particular, how it is generated, distributed, and controlled, and monitored through digital instruments. The smart grid is a term that refers to how the whole electricity production and distribution are controlled by the smart digital instruments (for example, sensors) and embedded systems. Figure 16 shows an example of Edge Computing in the smart grid. Over the past years, surveillance security has been playing an important role in our daily life, for example, ATM centre. Most of the surveillance security is based on the visual feed, where this feed needs to be analysed quickly using AI/ML/DL for better security reasons without taking much time with accuracy. And also, sometimes, there is some high risk of data being manipulated or leaked over the network. The following cases show how Edge Computing will improve or tackle this problem.

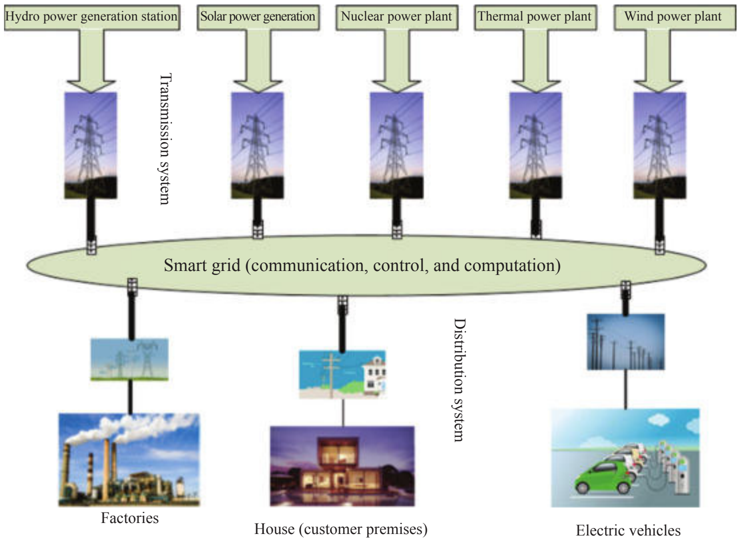


Figure 16: Example of Edge Computing in the Smart Grid [43].

- *Examples*:

  Case 1: Smart Grids Such infrastructures will benefit in numerous ways by adopting Edge Computing. For example, Edge Computing will make electricity to be bi-directional. This methodology allows the customer to be both consumers and producers of energy. This means that it will enable the customers to produce renewable energy, i.e. solar power and biofuels) and sell back their excessive production to other consumers in the accessible market created by the Edge computing-enabled smart grid [43].

  Case 2: Visual Detection During the process of visual detection, there could be a threat that might come from the firmware, direct physical access, and visual layer-based attacks. Sometimes these threats can be identified or not. In order to eliminate these threats, Edge Computing offers many solutions through AI/ML. For example, in face spoofing attacks, one can use the online frame forgery detection technique. Furthermore, in many cases, at the Edge, it would be inefficient to process the high-resolution video frames. There has been an algorithm defined in AI/ML to overcome this problem, for example, the super-pixel-based technique [65].

- *Security Concerns*: the automatic tracking of citizen movements and usage facilitated by Edge Computing architectures raises a serious concern with regards privacy preserving rights, which have to analysed in the context of a degraded security climate due to the increase in terrorists attacks hitting in the last decades EU countries among many others. For sure the ongoing developments within the Edge Computing paradigm could be of great help to protect critical infrastructures as smart grids, or sustain global decisions improving the global security of our continents while preserving our privacy.

8. Present Challenges

Even though Edge Computing is promising, there are still a few more areas that need to be well defined and documented. This will lead to further development in Edge Computing and its uses. The following list presents the areas where Edge Computing faces some challenges.

- **Naming:** Usually, Edge devices run many applications; each architecture has its method and structure. It is more convenient to have a naming scheme for programming, identification of things, and data communication in Edge Computing. But unfortunately, there is no standardised naming mechanism available for Edge Computing. This makes it very difficult for any programmer to understand the various communication and network protocols in Edge Computing [15].
- **Programmability:** In general, the Edge Computing architecture is heterogeneous. This makes the runtime and programming language different from standard architectures. Eventually, programming for Edge Computing becomes difficult. Whereas in the Cloud, the end-users usually can deploy their code written in a specific language. Moreover, Cloud Computing is known for its transparency [15].
- **Edge Device Management:** Managing Edge device is not an easy task, since it involves complex functionality, mainly, scalability, security, heterogeneity, and infrastructure performance. For example, Edge Orchestration should have the functionality of self-healing; this will minimise human intervention. Moreover, mobility is also a challenging topic; often, Edge devices move (for example, vehicles and cell phones). In this scenario, Edge devices shifting or collecting data from different locations, this will make a challenge in processing the data. Furthermore, Table 2 shows more of the critical functionality of the resource management and Edge Orchestration.

9. Summary

This technical report gives an overview of the Edge Computing paradigm and its applications, provides a comparison between Edge and Cloud Computing, and also points out the importance of this novel computing model to sustain the digital developments ongoing within our society. The key characteristics of Edge Computing architectures are discussed, including a brief survey of the orchestration middleware available together with the tools enabling the management, the effective deployment and the integration of data analytics capabilities within this novel distributed computing infrastructure. The latest trends at the heart of hardware developments for Edge Computing platforms are analysed, with concrete examples on the way Artificial Intelligence techniques and associated algorithms are tackled in this context. This technical report further explains why Edge Computing still depends on cloud technology or HPC supercomputers and lists a few critical challenges still opened in Edge Computing implementation. Finally, four categories of real-world applications affecting our daily life are proposed. They only illustrate the concrete benefit and potential impact this novel paradigm can bring to improve our digital society for the coming decades.

References

[1] A. Tekin, A. Tuncer Durak, C. Piechurski, D. Kaliszan, F. Aylin Sungur, F. Robertsen, and P. Gschwandtner. State-of-the-art and trends for computing and network solutions for hpc and ai, prace technical report. *PRACE Technical Report, Dec 2020*, 2020.

[2] A. Johansson, C. Piechurski, D. Pleiter, and K. Wadówka. Data management services and storage infrastructures. *PRACE Technical Report, Dec 2020*, 2020.

[3] Cisco. Cisco Annual Internet Report (2018–2023) White Paper. *White Paper*, 2020.

[4] Cisco. Global Cloud Index: Forecast and Methodology, 2014–2019. *White Paper*, 2014.

[5] Cisco Edge-to-Enterprise IoT Analytics for Electric Utilities Solution Overview. URL https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/big-data/solution-overview-c22-740248.html.

[6] Matthew Finnegan. Boeing 787s to create half a terabyte of data per flight, says virgin atlantic. *Computerworld UK*, 6, 2013.

[7] Jie Cao, Quan Zhang, and Weisong Shi. *Edge Computing: A Primer*. Springer, 2018.

[8] Koustabh Dolui and Soumya Kanti Datta. Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing. In *2017 Global Internet of Things Summit (GIoTS)*, pages 1–6. IEEE, 2017.

[9] Robert B Bohn, John Messina, Fang Liu, Jin Tong, and Jian Mao. Nist cloud computing reference architecture. In *2011 IEEE World Congress on Services*, pages 594–596. IEEE, 2011.

[10] Types of Cloud Computing Structures. URL https://www.uniprint.net/en/7-types-cloud-computing-structures/.

[11] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future generation computer systems*, 29(7):1645–1660, 2013.

[12] Harald Sundmaeker, Patrick Guillemin, Peter Friess, and Sylvie Woelfflé. Vision and challenges for realising the internet of things. *Cluster of European Research Projects on the Internet of Things, European Commision*, 3(3):34–36, 2010.

[13] Kevin Ashton et al. That ”internet of things” thing. *RFID journal*, 22(7):97–114, 2009.

[14] Shreshth Tuli, Redowan Mahmud, Shikhar Tuli, and Rajkumar Buyya. Fogbus: A blockchain-based lightweight framework for edge and fog computing. *Journal of Systems and Software*, 154:22–36, 2019.

[15] Auday Al-Dulaimy, Yogesh Sharma, Michel Gokan Khan, and Javid Taheri. Introduction to edge computing. In *Edge Computing: Models, technologies and applications*, pages 3–25, 2020.

[16] Paul Barham, Boris Dragovic, Keir Fraser, Steven Hand, Tim Harris, Alex Ho, Rolf Neugebauer, Ian Pratt, and Andrew Warfield. Xen and the art of virtualization. *SIGOPS Oper. Syst. Rev.*, 37(5):164–177, October 2003. ISSN 0163-5980. doi: 10.1145/1165389.945462. URL https://doi.org/10.1145/1165389.945462.

[17] Vmware esxi. URL https://www.vmware.com/products/esxi-and-esx.html.

[18] Anthony Velte and Toby Velte. *Microsoft Virtualization with Hyper-V*. McGraw-Hill, Inc., USA, 1 edition, 2009. ISBN 0071614036.

[19] Sun Microsystems Inc. The k virtual machine (kvm). White paper, 1999. URL http://java.sun.com/products/cldc/wp/.

[20] Virtualbox. URL https://www.virtualbox.org/.

[21] Docker. URL https://www.docker.com/.

[22] Matteo Riondato. Freebsd handbook: Jails. URL https://www.freebsd.org/doc/handbook/jails.html.

[23] S. Pousty and K. Miller. *Getting Started with OpenShift: A Guide for Impatient Beginners*. O'Reilly Media, 2014. ISBN 9781491904725. URL https://books.google.fr/books?id=K6aSAwAAQBAJ.

[24] Kubernetes, . URL https://cloud.google.com/kubernetes-engine.

[25] Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker, and Jonathan Turner. Openflow: Enabling innovation in campus networks. *SIGCOMM Comput. Commun. Rev.*, 38(2):69–74, March 2008. ISSN 0146-4833. doi: 10.1145/1355734.1355746. URL `https://doi.org/10.1145/1355734.1355746`.

[26] Open vswitch (ovs). URL `https://www.openvswitch.org/`.

[27] Message queuing telemetry transport (mqtt), iso/iec 20922. URL `https://mqtt.org/`.

[28] Kubeedge, . URL `https://kubeedge.io/`.

[29] Eclipse mosquitto: An open source mqtt broker. URL `https://mosquitto.org/`.

[30] Open network automation platform (onap). URL `https://www.onap.org/`.

[31] Tensor flow, . URL `https://www.tensorflow.org/`.

[32] Opencv. URL `https://opencv.org/`.

[33] Edgent incubator, . URL `http://edgent.incubator.apache.org/docs/edgent-getting-started.html`.

[34] Tensorflow lite, . URL `https://www.tensorflow.org/lite/models`.

[35] Apache mxnet, . URL `https://mxnet.apache.org/versions/1.7.0/`.

[36] Mxfusion. URL `https://mxfusion.readthedocs.io/en/master/index.html`.

[37] Keras-mxnet. URL `https://github.com/awslabs/keras-apache-mxnet`.

[38] Insightface. URL `https://github.com/deepinsight/insightface`.

[39] Autonomous machines. URL `https://developer.nvidia.com/embedded-computing`.

[40] Asus tinker board series. URL `https://tinker-board.asus.com/product-series.html`.

[41] Raspberry pi products. URL `https://www.raspberrypi.org/products/`.

[42] Programmable accelerator cards for data centers. URL `https://www.kalrayinc.com/download/konic-80-200/`.

[43] Alem Fitwi, Zekun Yang, Yu Chen, and Xuheng Lin. Smart grids enabled by edge computing. In *Edge Computing: Models, technologies and applications*, pages 381–408, 2020.

[44] Kmeans. URL `https://github.com/NVIDIA/kmeans`.

[45] Dumitrel Loghin, Lavanya Ramapantulu, Oana Barbu, and Yong Meng Teo. A time–energy performance analysis of mapreduce on heterogeneous systems with gpus. *Performance Evaluation*, 91:255–269, 2015.

[46] Jetson software. URL `https://developer.nvidia.com/embedded/develop/softwarE`.

[47] MG Murshed, Christopher Murphy, Daqing Hou, Nazar Khan, Ganesh Ananthanarayanan, and Faraz Hussain. Machine learning at the network edge: A survey. *arXiv preprint arXiv:1908.00080*, 2019.

[48] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 2020.

[49] Diego Peteiro-Barral and Bertha Guijarro-Berdiñas. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence*, 2(1):1–11, 2013.

[50] C Van Himbeeck. *Edge AI In Health Applications With A Focus On Hearing*. 2020.

[51] Yucen Nan, Wei Li, Shuiguang Deng, and Albert Y. Zomaya. Smart healthcare systems enabled by edge computing. In *Edge Computing: Models, technologies and applications*, pages 337–356, 2020.

[52] Clinical data science. URL `ttps://www.ccds.io/partnerships/`.

[53] Nvidia blog. URL `https://blogs.nvidia.com/blog/2020/05/24/medimaging-integrated-solution-edge-ai/`.

[54] Epf presentaion, . URL `https://www.eu-patient.eu/globalassets/policy/data-protection/data-protection-guide-for-patients-organisations.pdf`.

[55] Bob Duncan. Can eu general data protection regulation compliance be achieved when using cloud computing? In *Cloud Computing 2018: The Ninth International Conference on Cloud Computing, GRIDs, and Virtualization*, volume 35. IARIA, 2018.

[56] EU commission white paper, . URL `https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf`.

[57] Wuhui Chen, Zhen Zhang, and Baichuan Liu. Smart cities enabled by edge computing. In *Edge Computing: Models, technologies and applications*, pages 315–337, 2020.

[58] Amira Hadj Fredj and Jihene Malek. Real time ultrasound image denoising using nvidia cuda. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 136–140. IEEE, 2016.

[59] Biljana L Risteska Stojkoska and Kire V Trivodaliev. A review of internet of things for smart home: Challenges and solutions. *Journal of Cleaner Production*, 140:1454–1464, 2017.

[60] Jay Lee, Hossein Davari, Jaskaran Singh, and Vibhor Pandhare. Industrial artificial intelligence for industry 4.0-based manufacturing systems. *Manufacturing letters*, 18:20–23, 2018.

[61] rexroth a bosch company. URL `https://www.boschrexroth.com/en/us/products/product-groups/welding-technology/weld-spot-analytics/index`.

[62] Matteo Rubagotti, Tasbolat Taunyazov, Bukeikhan Omarali, and Almas Shintemirov. Semi-autonomous robot teleoperation with obstacle avoidance via model predictive control. *IEEE Robotics and Automation Letters*, 4 (3):2746–2753, 2019.

[63] Washington edu news. URL `https://www.washington.edu/news/2019/08/19/ergonomics-machine-learning/`.

[64] Policy department EU parliament. URL `https://www.europarl.europa.eu/RegData/etudes/STUD/2016/570007/IPOL_STU(2016)570007_EN.pdf`.

[65] Deeraj Nagothu, Ronghua Xu, Seyed Yahya Nikouei, Xuan Zhao, and Yu Chen. Smart surveillance for public safety enabled by edge computing. In *Edge Computing: Models, technologies and applications*, pages 409–433, 2020.

Acknowledgements

This work was financially supported by the PRACE project funded in part by the EU's Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 823767.

Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

Data Management Services and Storage

A. Johansson[a*a], C. Piechurski[b*b], D. Pleiter[c*c], K. Wadówka[d*d]

[a]Swedish National Infrastructure for Computing, [b]GENCI, [c]Forschungszentrum Jülich GmbH,
[d]Poznań Supercomputing and Networking Center

Abstract

HPC storage systems have evolved from fairly simple systems attached to a single cluster to site-wide complex infrastructures supporting migration of data between tiers of different performance characteristics and I/O acceleration. Exascale systems and AI workloads will continue this trend by placing even greater demands on speedy access to data. This report summarises the history of this development and examines some of the technologies that are building blocks of near-future storage systems, both hardware and the software required to manage the large amounts of data.

[a] andjo@nsc.liu.se
[b] christelle.piechurski@genci.fr
[c] d.pleiter@fz-juelich.de
[d] kwadowka@man.poznan.pl

Table of contents

1. Introduction

This technical report is part of a series of reports published in the Work Package "HPC Planning and Commissioning" (WP5) of the PRACE-6IP project. The series aims to describe the state-of-the-art and mid-term trends of the technology and market landscape in the context of HPC and AI, edge-, cloud- and interactive computing, Big Data and other related technologies. It provides information and guidance useful for decision makers at different levels: PRACE aisbl, PRACE members, EuroHPC sites and the EuroHPC advisory groups "Infrastructure Advisory Group" (INFRAG) and "Research & Innovation Advisory Group" (RIAG) and other European HPC sites. Further reports published so far cover "State-of-the-Art and Trends for Computing and Network Solutions for HPC and AI" [1] and "Edge Computing: An Overview of Framework and Applications" [2]. The series will be continued in 2021 with further selected highly topical subjects.

HPC storage systems have evolved from fairly simple systems attached to a single cluster to site-wide complex infrastructures supporting migration of data between storage tiers with different performance characteristics and I/O acceleration. Exascale systems and AI workloads will continue this trend by placing even greater demands on speedy access to data.

Topics covered here are related to storage infrastructures and management of data stored on these infrastructures. Hardware and software are discussed due to the need for making cost trade-offs that influence both. Well managed data workflows are a concern if data is expected to be available over 20+ years and thus far beyond the lifetime of a single compute resource with attached storage system. Much data is only used during the computation and there the concerns are mainly high bandwidth and low latency, but results and data sets required to reproduce results require more long-term infrastructure.

2. Storage Infrastructure

This section explores a number of storage system options from a systems perspective, examining hardware interfaces and protocols.

2.1. Evolution of Data Storage Systems

The amount of data created in our society is growing rapidly, and even experts struggle to predict the growth rate. In response, there is an insatiable need for more advanced high-performance storage systems to store such amounts of data appropriately taking both cost and performance factors into account.

When discussing storage systems, the concept of storage tiers is often used. Different tiers have different trade-offs between volume cost and performance since the access pattern for most data varies during the lifetime of the dataset. In this report the following tiers will be used.

Level 1	Short-term data, also known as scratch or work storage, often referred to as "warm"
Level 2	Medium-term data, also known as project storage
Level 3	Long-term data, also known as archival storage, less frequently accessed data that is often referred to as "cold"

Ten years ago, the largest HPC data storage systems contained only a few petabytes (around 10 PB) of disk space based on traditional magnetic hard disk drives (HDD) for short-term data supported by an archive or Hierarchical Storage Management (HSM) system, both relying mainly on tape libraries to store cold data. Today, the equivalent systems are supporting an order of magnitude (around 20 times) of the capacity for short-term system storage, with additional storage tier levels to store data in an efficient and cost-optimised way depending on data access or criticality. See Figure 1 for an overview of the current storage hierarchy.

Generally, this ratio can also be applied to any common industrial and academic research computing systems, which can support up to several petabytes of storage today. Some research laboratories are able to create petabytes of data from their own scientific instruments. As an example, the large-scale scientific radio telescopes that form the Square Kilometre Array (SKA), built to explore the Universe, will generate Exabytes of scientific data per year. The raw SKA data will be filtered and post-processed using HPC resources, but the project still expects to archive 600 Petabytes of data per year [3]. Also, the volume of generated computational data expands dramatically with increasing computing capabilities.

Flash memory technology has evolved to become part of mainstream high-performance storage devices. In the HPC context these are non-volatile disk drive devices, not to be confused with slow memory cards or small embedded memories for storing system configuration data. This technology is the basis for current low latency and high bandwidth devices such as a Solid-State Drive (SSD).

Today's storage needs will be further increased by the upcoming Exascale systems. The performance capabilities of these systems will also increase, entailing new storage requirements in term of capacity, throughput and I/O operations per second (IOPS) relative to the data profile of the workload. While bandwidth and storage capacity are still the dominant performance indicators when choosing a storage system, IOPS are now an important criterion to consider when analysing HPC storage needs. The data handling requirements of Exascale, Big Data and Artificial Intelligence systems are very high for both IOPS and volume. Level 1 storage systems based on SSDs are relatively small in capacity for cost reasons but capable of handling high throughput of IOPS intensive data. Protocol limitations on how the storage devices are interfaced to systems drove the creation of the Non-Volatile Memory Express (NVMe) protocol, a new communication standard between the processor and data storage. NVMe is a communication interface and driver specification that defines a set of commands and a set of functions for PCIe-based SSDs to increase IOPS performance and interoperability across a wide range of corporate and client systems. Due to its low latency and direct PCIe connection allowing tighter coupling, it can handle specific workloads more efficiently than large shared storage systems that support all types of workloads but in a more general manner.

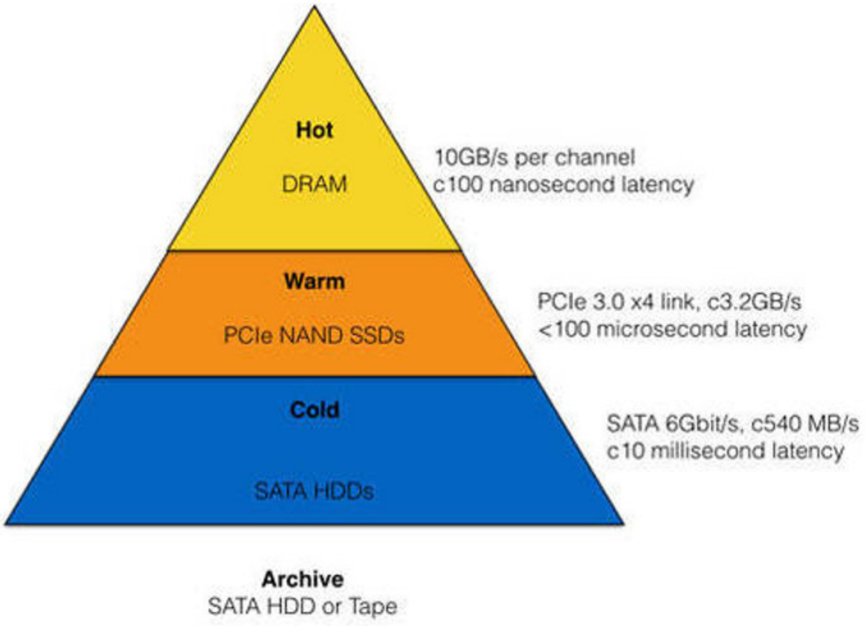


Figure 1: Storage and Memory Hierarchy today [4]

Compared to high-performance storage solutions for warm data, cold data is usually stored in less expensive storage environments with lower IOPS and bandwidth. Tape systems have been and are still a popular storage medium for cold data. Linear Tape-Open (LTO) was originally developed in the late 1990s as a low-cost vendor neutral storage option while some vendors (ex. IBM with their Jaguar tape drives) also market proprietary formats. See Section 2.4.2 for a more in-depth look at tape technologies.

Systems based entirely on flash memories are ideal for warm data where the trade-off favours IOPS and bandwidth above capacity. Large volume storage and data retention policies create an environment more in favour of low-cost capacity for cold data. Operational costs are also lower if data can be stored on powered down storage when the latency of bringing it online is acceptable. For security reasons the use of totally offline media may also be mandated.

Until recently, the cold storage system was mostly provided by tape libraries which guarantee low power consumption, large capacity and average read/write speeds while incurring longer waiting times for data. Volume storage systems based on HDDs with SATA interfaces are starting to compete with tape libraries for cold storage due to them both becoming capable of spinning down drives to reduce power consumption and still retain their random-access nature. The time it takes to spin up a drive is much shorter than the mount and spool time for a tape cartridge, so for data that is intermittently used rather than archived this can justify the higher cost of HDD storage.

2.2. High-Performance Storage

In the Exascale computing era, data processing capabilities are becoming a key factor. Standard approaches used by traditional I/O solutions are increasingly becoming a bottleneck. While new technologies such as data caching can help solve performance issues at a higher cost per terabyte, it is important to consider implementing another layer of hierarchial storage management into HPC architectures using ultra-high-speed storage technologies to support the current and futures challenges. This section examines a few of these technologies.

2.2.1. 3D-NAND and V-NAND Memory

Typical NAND memory chips (SLC, MLC, TLC) are built in such a way that all cells that store data are in one plane (type (a) in Figure 2). To increase the capacity of the modules, the cells must be placed more densely. This can be done by reducing the space between them, but it cannot be done indefinitely. At some point the density of the cells will be so high that the electric charges stored in them will leak between them. The result will be data corruption or irretrievable loss. The solution to this problem is 3D-NAND and V-NAND modules where memory cells are stacked in layers. This technique not only allows to increase the capacity of the media, but also has a positive effect on their efficiency and is not associated with higher production costs. For this reason, 3D-NAND can be considered a major breakthrough for flash technologies. With the new layered style of memory more and more data can be stored within the same physical area.

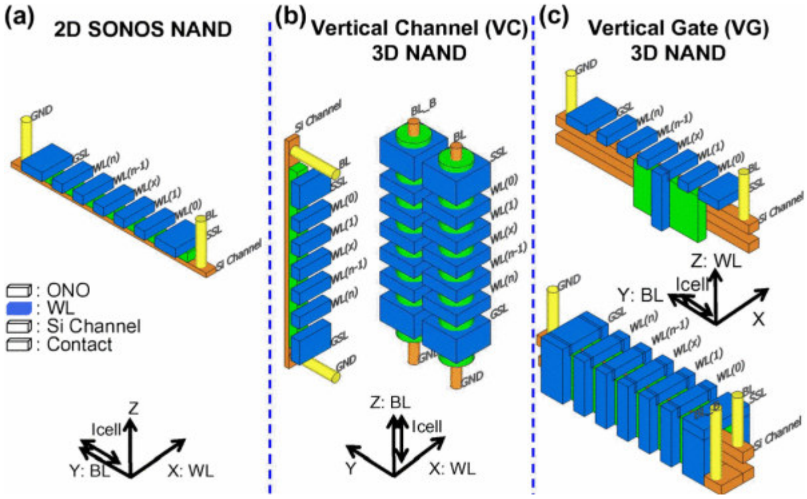


Figure 2: Schematic structure of some NAND flash types [5]

2.2.2. Intel Optane Memory

Intel Optane memory is not typical disk drive or DRAM computer memory (Figure 3), but a proprietary Intel standard. This is not a technology used for conventional storage. Instead, the M.2 form factor Optane module is a large cache bridge between the volatile DRAM and non-volatile storage, capable of storing a larger amount of data than traditional DRAM (but in a persistent way) and enabling faster data transfer between memory, storage and processor. Given proper OS support this additional layer speeds up most end user operations, using caching software that stores relevant data on an Optane drive for almost instant recall. Intel also uses the Optane name for smaller high-endurance low-latency SSDs with U.2 form factor.

The idea of using a small amount of super-fast flash memory to increase the performance of a basic memory disk is nothing new. In fact, Optane is essentially the next generation of Intel's Smart Response Technology (SRT) that can use low-capacity, expensive SSDs (compared to HDDs) to cache data for slower, conventional high-capacity hard drives. The difference is that Optane uses memory produced and sold by Intel in conjunction with special hardware and software components on compatible motherboards. Optane memory works with all kinds of RAM modules, storage drives and graphics cards that match a compatible motherboard.

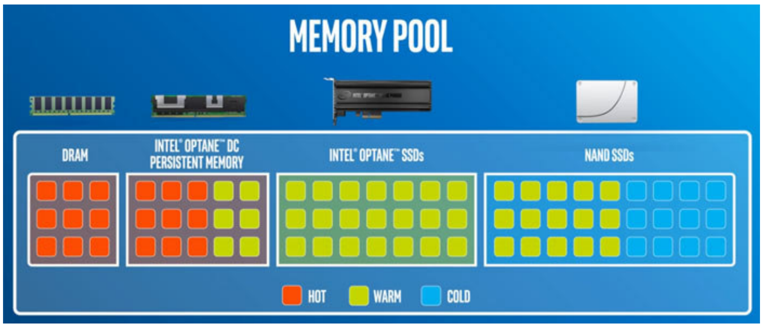


Figure 3: Optane Memory Pool Example [6]

2.2.3. NVMe Memory

NVMe memory eliminates much of the software overhead between applications and storage and is optimised for interfacing flash storage solutions to host systems, thus significantly reducing latency and increasing system and application performance. This is a vendor neutral standard for connecting storage to host systems, and while for example Optane described in Section 2.2.2 uses NVMe for its physical layer it uses proprietary software in addition to NVMe.

The NVMe specification has been designed for flash memory. I/O tasks performed with NVMe drivers have lower latency and more IOPS relative to older storage models using standards such as AHCI (Advanced Host Controller Interface) used with SATA SSD drives. Since the specification was designed specifically for flash storage, NVMe is becoming the new industry standard for data centre systems that require massive bandwidth and short access times. It is available in several form factors, with PCI-e and M.2 being the most popular. The M.2 form factor used by many NVMe devices allows high capacity in a small storage enclosure, such as a cache device, and are ideal for systems where the physical size of the device is a limiting factor. A disadvantage of NVMe is the use of flash memory which is more expensive than HDDs with respect to capacity, lacking legacy support that prevents it from upgrading older storage systems, and the generally lower capacity of flash drives compared to traditional disk drives.

NVMe memory is very well suited for IOPS intensive applications with very heavy workloads that require ultra-low latency.

2.2.4. Solid State Drives

While most storage systems based on HDD support large transactions (large file & large block transfer) well, an increasing number of those systems are facing I/O performance bottlenecks as they are less capable of absorbing low latency transactions in the same way. So, mixing both type of transactions, throughput and IOPS oriented is somewhat disturbing storage system performances. Therefore, operators are considering using SSD based storage system mainly for IOPS driven workloads on a low storage capacity, also being cost driven (min. 20 times higher per PB built in a high-performance oriented way) to increase overall storage system efficiency and reliability and reduce overall maintenance costs. On the other hand, it is unclear how operators can control the lifetime on an SSD based on DWPD (Drive Writes Per Day) set by device manufacturers. SSDs are manufactured in such a way that they can easily be implemented as replacements or additions to hard disks (HDDs) equipped with rotating magnetic plates. They are available in a variety of form factors, including standard 3.5- and 2.5-inch drive sizes, and support various communication protocols / interfaces. Devices can be directly attached using Serial ATA (SATA), Serial Attached SCSI (SAS) and recently PCIe (NVMe standard) to enable data transfer to and from server processors.

2.3. Storage Systems Solutions for the First Announced HPC Exascale Systems

Exascale storage requirements are no longer driven by traditional workloads with large streaming writes like checkpoint/restart but is increasingly driven by complex I/O patterns from new types of applications. High-performance data analytics workloads are generating vast quantities of random reads and writes. Artificial Intelligence workloads are reading far more than traditional high-performance computing workloads. Data streaming from instruments into an HPC cluster require better quality of service to avoid data loss. Data access time is now becoming as critical as write bandwidth. New storage semantics are required to query, analyse, filter, and transform datasets. A single storage platform in which next generation workflows combine HPC, Big Data, and AI to exchange data and communicate is essential.

The three planned DOE Exascale storage systems will rely on the Lustre file system. On top of Lustre, additional layers can be found including one based on the Intel Distributed Asynchronous Object Storage (DAOS) product which is optimised for non-volatile memory (NVM) technologies such as Optane persistent memory and Optane SSDs. DAOS is the foundation of the DOE Exascale storage stack and is an open-source software stack for scale out object storage that provides high bandwidth, low latency, and high I/O operations per second (IOPS) storage containers to HPC applications. It enables next generation data centric workflows that combine simulation, data analytics, and AI.

The DAOS instances of the Aurora system at Argonne National Lab, scheduled for delivery in 2022 and targeted as an Exascale system, is planned to feature a bandwidth of more than 25 TB/s and a capacity of 230 PB [7].

The EuroHPC pre-Exascale system LUMI will be using Lustre for both the all-flash warm data storage (7 PB) and the capacity storage (80 PB). In addition to this more traditional cluster storage LUMI will use Ceph storage for a data management service (30 PB). All these components will be connected to the cluster interconnect.

2.4. Large Capacity Storage

2.4.1. Hard Disk Drives and Hybrid Systems

While the popularity of SSDs has been growing for several years, the near future of the disk array market probably belongs to hybrid systems that support both HDDs and SSDs. While SSDs are characterised by higher performance and shorter access time than HDDs, their weakness is their lower capacity and corresponding higher price per TB compared to HDDs. For this reason, solid state drives are still not suitable for mid-price class storage solutions.

Automatic tiering is the most popular technology used in hybrid arrays. The device stores the most-used data on the fastest tier (ex. SSD) and migrates data to the slowest media tier (ex. HDDs) based on policy rules (age, access frequency, etc.) defined in the storage system. This method enables performance improvements on a global storage system level, where for example data accessed frequently are placed on fast storage to potentially benefit multiple hosts.

One example of a hybrid system that can be used as a building block for larger storage systems is the DDN SFA18KX hybrid disk array. It offers up to 3.2 million IOPS and 90GB/sec from a single 4U appliance. It uses NVMe devices and spinning drives. This high level of density makes the SFA18KX suited for data centres with limited space or any high-performance environment that aspires to expand capacity without adding the complexity of many appliances to manage and the cost of powering and cooling a large number of controllers.

Another interesting solution combining disk technologies is the NetApp E5700 series hybrid flash array where SAS attached HDDs and SSDs can be combined. It was designed specifically for heavy duty environments, including those for analysing large data sets, the E5700 provides over 1 million persistent IOPS and response times in microseconds. Bandwidth oriented loads can reach up to 21 GB/s. The E5700 is a 2U (24 drives) or 4U (60 drives) array supporting multiple high-speed host interfaces, including 32 Gb/s Fibre Channel, 25 Gb iSCSI, 100 Gb InfiniBand, 12 Gb SAS, 100 Gb NVMe via InfiniBand and 100 Gb NVMe via RoCE (RDMA – Remote Direct Memory Access – over Converged Ethernet).

2.4.2. Tape Libraries

Historically, tape drives were used for local system backups with a system administrator changing tapes in drives directly attached to servers. High speed networks, large disk drives and cloud storage have made the directly attached drives for small scale backups a niche market. In most cases tape usage has been concentrated on large automated tape libraries used by many systems for backups and archive storage. This consolidation of storage and fewer drive sold have had consequences both for the business and the technical side.

Economies of scale mean that tape technology becomes a winner takes it all market with the need to amortise R&D costs over a low number of units. The StorageTek T10000 format was the main contender to IBM 3592 in the high-end tape market but could not compete on production volume, and development of the "E" format was cancelled a few years back. Thus, tape technology has become a mostly single vendor market on the drive side with IBM producing drive heads for both LTO and 3592 (aka Jaguar) drives. LTO media is produced by Sony and Fujifilm which in 2019 settled their lawsuit which had essentially halted production of LTO-8 media and slowed down the adaption of LTO-8 technology. In late 2020 LTO-9 is being introduced into the market, so LTO-8 may end up as one of the less widely deployed LTO generations.

IBM classifies the 3592 drives as "enterprise" technology, and new features are usually first introduced there. Main differences to LTO were historically the storage capacity, bandwidth and seek/access times. In recent generations the bandwidth gap has narrowed, and the pricing for tape media means that the price per TB is similar. Latency is the remaining large difference, with 3592 supporting recommended access order (RAO) and higher resolution directories for high-speed seeking. See Table 1 below for a summary.

Tape Media	**Uncompressed Capacity**	**Uncompressed Bandwidth**	**RAO Support**	**High Resolution Directory Size**
3592-JE	20 TB	400 MB/s	Yes	64
LTO-8	12 TB	360 MB/s	No	2
LTO-9	18 TB	400 MB/s	No	2

Table 1 Tape technology feature comparision

Mechanical constraints limit the possible performance of LTO drives due to pressure from system vendors to support the use case of tape drives inside a server chassis. This limits the physical form factor of the drive to what is known as half-height in the tape world. Combining this size with the fixed size of the tape cartridge leaves little room for the intricate mechanics needed for moving the tape with both high speed and precision. In the future this will probably require the LTO format to have divergent performance tiers for half- and full-height drives, for LTO-8 the difference is 300 vs 360 MB/s. Tape libraries usually use full-height drives since they are less constrained by space, but the lower price of half-height drives makes them an option for some libraries where price is more important than performance.

Competition remains on the tape library side with IBM, Oracle and Spectra Logic providing a range of library models targeting even the largest sites while Quantum are more focused on the low/mid-sized sites. All vendors are offering LTO technology in the libraries, with IBM and Spectra Logic also providing 3592 drives. Some notable differences between libraries are the different methods used for storing tapes inside the library. Optimising for mount latency leads to having libraries where all tapes are directly accessible and have short paths to drives (example StorageTek SL8500) and optimising for density leads to depth stacking of cartridges (for example, IBM TS4500). Trying to strike a balance is the drawer approach (for example Spectra Logic T950) where cartridges are placed in containers that can be pulled out and a single cartridge taken, or the entire container moved to/from the I/O station.

The ransomware attacks in the last few years have caused a resurgence of interest in tape technology for normal enterprise backups due to the possibilities of both keeping them separate from other systems and also physically removing the tapes for vault storage. HPC sites usually have tapes for archival storage and lower cost tiers in storage systems and keep all tapes accessible online. For most of these use cases the mount latency is not critical, and seek latency is more important.

Tape storage is viewed as low performance but is mostly high latency storage due to its sequential nature, not low bandwidth. When reading and writing data, tape drives prefer a steady stream of data to keep up with the movement of the tape and will do speed matching within a range but cannot go arbitrarily low. An LTO-8 drive, for example, has a lower limit of 112 MB/s when streaming. Technology projections [8] are becoming an increasingly important issue in the future. To be able to feed the tape libraries during writes they need to be matched with high-speed disk storage, so coupling flash and tape tiers directly will be attractive.

3. File Systems

Most HPC installations rely on distributed/parallel file systems, with Lustre and Spectrum Scale (formerly GPFS) being the most common. Thanks to their built-in scalability, they are able to manage huge amounts of data, support high bandwidth and provide high metadata performance. The increasingly demanding requirements of HPC systems are a good test case for new storage technologies. These file systems can perform hundreds of thousands of operations on metadata per second and stream multiple TBs of data per second. To meet the ever-increasing demands, cluster file systems are constantly evolving towards more universal, stable, and useful solutions. Current user expectations include high-performance access to small files, increased levels of security (encryption for example), support for data replication mechanisms and data retrieval via automatic tiering. Some of these are features that enterprise systems (NetApp FAS, EMC, etc.) have provided earlier without having the same level of performance.

In this section we will examine the evolution of the cluster file systems Lustre, Spectrum Scale and BeeGFS. All these have recently been developed with new features. Lustre is now implementing Distributed Namespace (DNE), Erasure Coding, Data on Metadata (DOM) and Persistent Client Cache (PCC). Spectrum Scale is now implementing Native Declustered RAID and BeeGFS has Storage-On Demand.

3.1. Lustre

Lustre is open-source software whose development is supported and coordinated by the non-profit EOFS and OpenSFS organisations. Developed since 1999, it is used as a file system by many computing environments in the world. For more than 15 years, it has been used by at least half of the top 10 largest supercomputers and is known for supporting the largest high-performance computing clusters in the world, with tens of thousands of client systems, petabytes of storage deployed, and hundreds of gigabytes per second of I/O bandwidth. The central component of the Lustre architecture is the Lustre file system, which is supported on the Linux operating system and ensures compatibility with the POSIX standard.

Until recently, the Lustre file system performance has been optimised for large files. This results in many Remote Procedure Call (RPC) round trips to the Object Storage Targets (OSTs), which reduces small file performance. Therefore, a new functionality has been implemented to allow the placement of small files on Meta Data Targets (MDT) (Figure 4) so that these additional RPCs can be eliminated, and performance improved correspondingly. Used in conjunction with the Distributed Namespace (DNE), this will preserve efficiency without sacrificing horizontal scaling. Users or system administrators can set a layout policy that places small files on MDT. Files that grow beyond this size will use Progressive File Layouts to extend larger files onto OST objects and leave the small part of the file (defined by user or system admin) on the MDT.

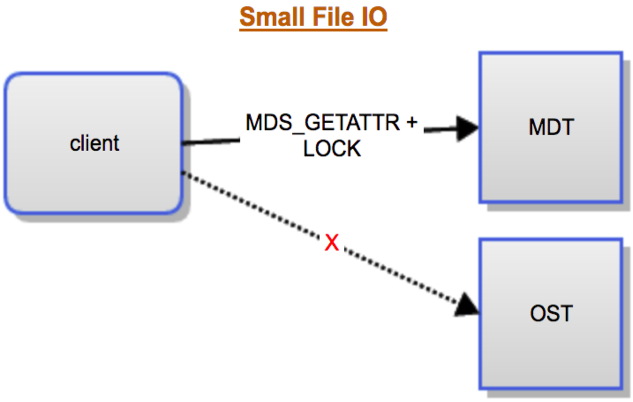


Figure 4: Small File IO (http://wiki.lustre.org/)

Persistent Client Cache (PCC) is another new feature that was implemented in Lustre version 2.13. Data is cached locally on SSDs or NVMe drives at the Lustre client side. These caches are not part of the global Lustre namespace, instead each client uses its own SSD as a purely local cache. Cached data is managed using a local file system with I/O for cached files being fulfilled locally while other I/O is directed to the shared OSTs. PCC uses the previously existing HSM (Hierarchical Storage Management) support in Lustre for data synchronisation. It uses the HSM copy tool to move files from the local cache to the Lustre OSTs, acting as a HSM backend with a unique archive number. If another Lustre client accesses data cached in this manner it will trigger data synchronisation. Clients using PCC going offline are handled by making the data temporarily inaccessible for other clients. When the PCC client is online again the copy tool restarts, and the data is accessible again.

One feature on the Lustre 2.14 roadmap is client-side data encryption. This will increase security to a higher level, where a leak of data sent to the server is no longer a threat. Assumptions about this ability are as follows:

- encrypting file contents
- encrypting file name
- using the master key to encrypt data
- file data is no longer available after deleting the key
- ability to change the key without re-encrypting the files
- denying access to encrypted data after deleting the master key from the client's memory.

3.2. Spectrum Scale (formerly GPFS)

Another very well-known product on the clustered file systems market is IBM General Parallel File System (IBM GPFS) renamed IBM Spectrum Scale. The filesystem layout spreads data between multiple servers simultaneously, thus creating a global namespace. It supports both large scale HPC environments as well small-scale HPC systems. The word "parallel" in the former product name indicates the main feature of Spectrum Scale (Figure 5), namely the mechanism of data partitioning and their simultaneous distribution on many disks / disk arrays. This feature allows faster reading and writing of data. Additional mechanisms exist in Spectrum Scale to increase the reliability and performance of the entire system, such as automated management functions, high availability of resources, replication and mirroring.

Spectrum Scale is a clustered file system. This means that it provides simultaneous access to one file system from multiple nodes. All nodes can be connected to the SAN or network data storage systems. This enables high-performance access to the same resources through multiple access nodes simultaneously.

The Spectrum Scale file system provides the following mechanisms:

- Increasing the total throughput of a file system by spreading reads and writes across multiple disk resources.
- Simultaneous access to multiple processes or applications on all cluster nodes using standard file system calls.
- Load balancing by evenly distributing data across all drives. This increases the total system capacity and eliminates the bottlenecks in the transfer.
- Each physical disk device intended for use in the Spectrum Scale cluster should be defined as an NSD (Network Shared Disk). This allows you to create an additional layer of logs for input/output operations. This directly translates into increased system efficiency.
- Support for very large file sizes.
- Parallel, simultaneous reads and writes from multiple IBM Spectrum Scale cluster nodes.
- An extensive system of managing distributed tokens (locks) on files. Token management distribution reduces facility maintenance latency.
- Writing data to multiple disks via various disk controllers. Large files in IBM Spectrum Scale are split into blocks of equal size, and successive blocks are placed on different disks.
- Acceleration of reading by pre-fetching data into buffers.
- Native Declustered Raid: IBM Spectrum Scale RAID implements a sophisticated data and spare space disk layout scheme that allows for arbitrarily sized disk arrays while also reducing the overhead to clients when recovering from disk failures. To accomplish this, IBM Spectrum Scale RAID uniformly spreads or *declusters* user data, redundancy information, and spare space across all the disks of a declustered array.

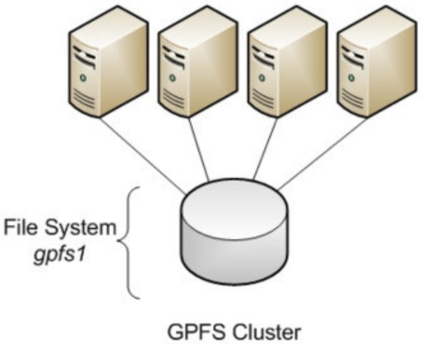


Figure 5: Simultaneous Access by Multiple Clients to GPFS Resources (IBM)

3.3. BeeGFS

The high-performance file system market is dominated by players like Lustre and Spectrum Scale. In recent years the European developed file system BeeGFS has emerged as a competitor with a growing number of sites implementing it. This file system was originally developed by Fraunhofer as an internal file system named FhGFS, but development was spun out by forming the company ThinkParQ.

BeeGFS boasts many features that are useful for users of high-performance file systems. It is software defined storage based on the POSIX file system standard. It means that applications can easily and efficiently use BeeGFS resources. System clients communicate with the cluster via a TCP/IP network or a high-performance Infiniband network.

The main features of a BeeGFS cluster are:

- Data is spread across multiple servers and increasing the number of servers and disks in the system translates directly into the capacity and performance of the file system represented as a single namespace.
- The BeeGFS network protocol is independent of the hardware platform. Hosts of different platforms can be mixed within the same file system instance.
- Management, metadata and storage services do not have direct access to the disks. Instead, they store data in any local POSIX file system (Ext4, XFS or ZFS). This gives you the flexibility to choose a basic file system that gives you maximum performance in the context of the hardware used.
- BeeGFS uses all available RAM in the server (which is not needed by other processes) automatically for buffering data. This gives a huge performance gain when handling small I/O requests and then aggregating them into larger blocks before saving to disk.
- BeeGFS has a feature that allows flash drives to become directly accessible to users. Users can request BeeGFS (via the beegfs-ctl command line tool) to transfer the current project to high-performance flash drives (e.g. NVMe) (Figure 6).
- BeeGFS supports all networks based on the TCP/IP protocol and the native InfiniBand or Omni-Path protocols. Servers and clients can handle requests from/to different networks at the same time.

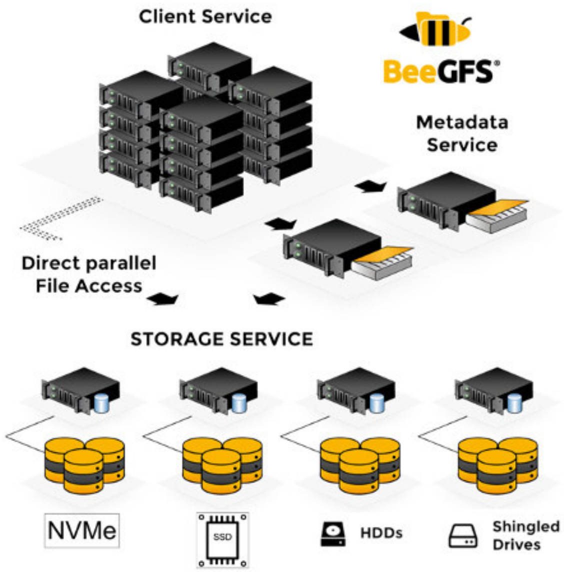


Figure 6: BeeGFS The Parallel Cluster File System [9]

3.4. Ceph

The Ceph storage system was introduced as a prototype in 2006 [10] and its file system client code been a part of the Linux kernel since 2010. Funding for the open-source project is provided by the Ceph Foundation, which in itself is hosted by the Linux Foundation. Ceph is designed as a distributed object storage cluster for commodity hardware with object, block and file system storage services layered on top (see Figure 7 for an overview of the architecture).

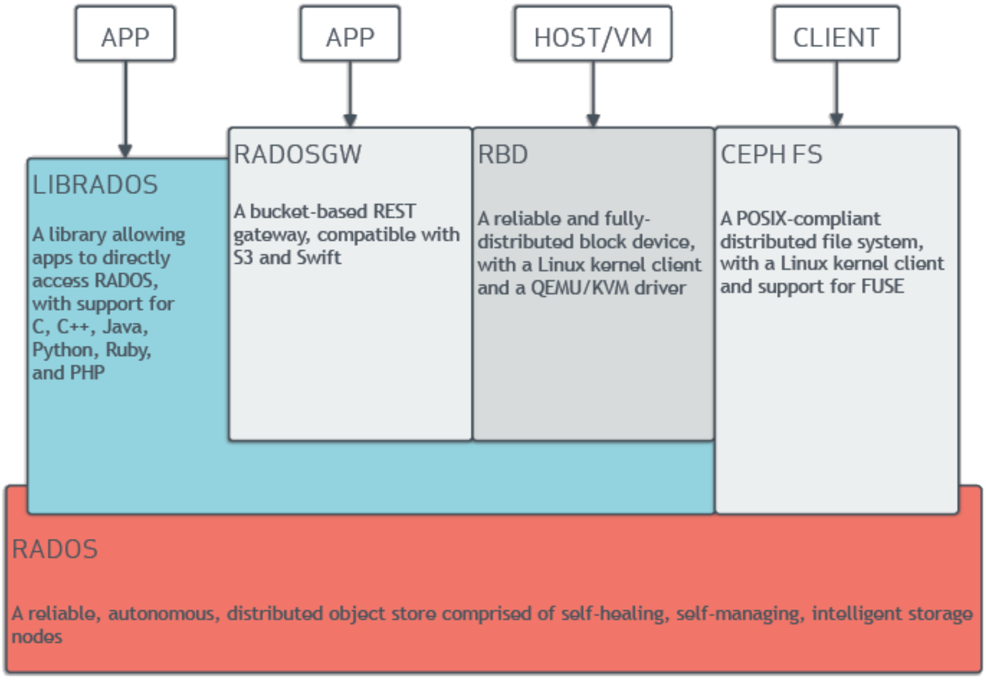

APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver
CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

Figure 7: Ceph architecture [11]

Ceph is not designed to provide storage from single server, to take advantage of its features multiple Object Storage Devices (OSD) are needed to support for example replication and erasure coding. Most of the early deployments of Ceph have been in cloud environments where S3 or Swift access protocols and block storage are needed for the virtual machines. As a file system Ceph is not yet a mainstream HPC choice but is starting to be used at sites that handle both traditional HPC clusters and cloud environments. One notable example is CERN where it is used as storage for their Openstack cloud. Upcoming deployments include the LUMI EuroHPC systems which also support a container cloud platform based on OpenShift and Kubernetes.

3.5. File System Feature Comparison

Table 2 compares the feature sets of some file systems popular in the HPC environment with regards to automatic data movement within the storage hierarchy.

Features mentioned below should be interpreted as

- **S3 Provider**: The file system can also act as object storage and provides an interface that is compatible with the S3 protocol popularised by AWS
- **Tier to Drives**: Data is migrated between different drive technologies transparently to optimise latency and cost trade-offs, with an example being SSD to/from HDD movement
- **Tier to Tape**: Data is migrated to a tape library, but still visible in the file system namespace and transparently migrated back if accessed
- **Tier to Cloud**: Data is migrated to object storage, either on premises or using an external provider

	S3 Provider	**Tier to Drives**	**Tier to Tape**	**Tier to Cloud**
BeeGFS	No	Yes	No	No
Ceph	Yes	Yes, cache	No	No
Lustre	No	Yes	Yes	Yes, unofficial
Spectrum Scale	Yes	Yes	Yes	Yes

Table 2: File System Features for Cloud and Tiering

4. Data Management Services

Today, parallel file systems, possibly in combination with a Hierarchical Storage Management (HSM) system, are still the most important approach to provision storage resources for HPC systems. To meet future needs, new technologies for data management and tiering are becoming increasingly important. For the future we expect not only growing needs in terms of storage capacity and performance, but also needs related to collaborative data management, realisation of workflows extending a single HPC data centre to support multiple groups sharing data as well as the provisioning of data according to the FAIR data principles [12] are becoming key drivers.

In this section we report on select developments on the market, based on conference presentations and the authors' personal experience, which are relevant in this context. In Section 4.1 we analyse primarily the evolution of cloud storage technologies which, among others, can help to facilitate data sharing. Large-scale cloud providers are pushing for a new approach to architecting storage. In Section 4.2 we report on the relatively new concept of data lakes. Unlike cloud storage architectures, HPC storage architectures have long relied on tiered architectures comprising tiers optimised for capacity and others optimised for performance. In Section 4.3 we summarise the status of more traditional solutions for managing tiered storage architectures, while in the following Section 4.4 we focus on emerging solutions for I/O acceleration architectures and technologies. Finally, in Section 4.5 we consider the existing and emerging data management frameworks that are attracting interest in the context of HPC infrastructures.

4.1. Cloud Storage Interfaces

Object store architectures are receiving an increasing interest in the context of HPC mainly as a possible option for addressing scalability issues related to POSIX compliant parallel file systems when going to Exascale. Object store technologies that are used for cloud infrastructures are, however, also of interest for HPC infrastructures as they are designed for geographically distributed storage infrastructures that are much more openly accessible than parallel file systems. Thus, they help to meet the need of data sharing and realising workflows that extend beyond a single data centre. This is an important aspect for realising the recently formulated vision of "Transcontinuum Extreme-Scale Infrastructures" [13].

The focus here is not on object store architectures but rather their interfaces, i.e. on S3 and Swift. S3 is an API introduced and controlled by Amazon [14], while Swift is an object store technology and API developed by the OpenStack community [15]. What they both have in common is that they are web-based interfaces implementing a REST API that supports a small set of operations like *get object* or *put object*.

The use of cloud storage interfaces to facilitate external access to HPC data centres is still at an early stage. The Fenix project has recently announced that they will use Swift to facilitate access to federated storage resources [16]. There are several commercial solutions on the market that will support this development. There are two approaches to provision access to storage through the aforementioned APIs: one can either use (1) native object store solutions like OpenStack Swift or (2) provision these interfaces on top of other storage solutions, e.g. parallel file system solutions.

The latter approach is realised by IBM's Cluster Export Services (CES) for Spectrum Scale [17]. A CES node acts as a protocol node providing non-Spectrum Scale clients with access to data managed by Spectrum Scale. This approach has multiple benefits in the context of HPC data centres. Parallel file systems are technologies that are well integrated and supported by these data centres. Additionally, various commercial solution providers, which are active in the HPC market, provide the necessary support for such a configuration. Currently, similar solutions are not yet available for Lustre.

Commercial support for deployment of native objects stores is improving. Atos announced its new BullSequana Xstor [18] with support of Ceph, which provides both an S3 and a Swift compatible API through its Rados gateway. The new object store architecture Mero, which is partially developed within the EU-funded Sage projects [19], supports different storage interfaces (like S3), based on a component called Lingua Franca. This component implements different meta-data formats and interfaces. A number of smaller suppliers started to provide proprietary object store solutions including S3 interface, including the French company OpenIO, which is positioning its solution also in the HPC market and show-cased their product at SC'19 [20]. Yet other suppliers bundle open-source software stacks like OpenStack Swift for enabling commercial offerings, e.g. SwiftStack [21], a company recently acquired by NVIDIA. The latter solution is notably positioned as a solution that allows to extend storage beyond the data centre towards the Edge Computing.

While commercially supported solutions are receiving an increasing interest by HPC data centres, there are also solutions developed by research organisations in the context of grid computing [22] moving towards support of web-based interfaces. One example is dCache that has been developed for high energy physics storage application and supports grid protocols, network file system access as well as WebDAV [23] using protocol "doors", similar to the CES nodes for Spectrum Scale mentioned above. It is designed to be distributed among sites and supports tertiary storage systems for tiering to tape, for example. Developed for grid applications, it has also been used in federated national storage.

4.2. Data Lake and ad-hoc Storage Systems

Driven by commercial cloud providers like Amazon, Azure and Google, the new concept of *data lakes* is gaining momentum and is expected to be adopted also in the context of HPC infrastructures. A data lake can logically be seen as a centralised repository that might be realised on distributed storage resources, which allows for the storage of structured and unstructured data at any scale. The data is imported from different sources and provisioned in its original format. The aim is to make data available for processing and analytics pipelines soon after it becomes available.

A data lake can typically be expected to be a storage tier for cold data objects that can be accessed with limited performance and in formats that are not suitable for further processing. To make data access suitable for high-performance computing and high-performance data analytics specialisation and locality must be improved [24]. This can be achieved through, for instance, dynamic provisioning of storage [25] [26]. One solution that realises this is BeeGFS on Demand (BeeOND) that is commercially supported through ThinkParQ [27].

4.3. Tiered Storage Management

Tiering is not new in HPC environments and solutions, which may be considered "classical", continue to be further enhanced. The classical tiered storage approach in HPC environments is extending the file system to support multiple tiers. Cluster file systems such as Spectrum Scale and Lustre support this natively or through addons that can be extensively configured to select which files are migrated between tiers. This creates the appearance of all files being locally online while they may, for example, be stored in a faraway tape library with high access latency. Tiered storage management solutions may be largely invisible to the user.

The following part highlights different products for tiered storage management which are all actively developed and can be expected to continue being relevant for HPC infrastructures.

IBM Spectrum Scale is a parallel file system used at many HPC sites. Tiering support is built in for disk and cloud tiers with space management add-ons supporting tape. Spectrum Scale Information Lifecycle Management (ILM) is a set of tools that allows to define placement and migration policies (see [28] for a recent overview).

Lustre HSM support was added in release 2.5 of Lustre. The design is based on a coordinator and agents that are responsible for moving data between the Lustre and HSM worlds. Migration requests can be user-triggered or initiated by a policy engine like Robinhood, which was developed at CEA and is the most commonly used addon for Lustre HSM. Lustre HSM continues to be actively developed (see [29] for a recent update).

HPE Data Management Framework (DMF) is a software-defined framework for managing multiple storage tiers [30]. It can connect high-performance file systems, e.g. Spectrum Scale or Lustre, and a back-end data store which could, for example, be based on tape or an object store with off-site data replication enabled. In the most recent version DMF7 support for extensible metadata was added, which enables new data management capabilities, e.g. handling of data sets, and better integration with HPC job schedulers.

HPSS (High Performance Storage System) is an HSM system built mainly by IBM and US DOE lab [31]. It has been developed for a long time and with support for tape usage. It is optimised for I/O bandwidth by supporting parallel I/O through software striping, e.g. through RAIT (Redundant Array of Independent Tapes), which allows for striping data on tape. The Spectrum Scale can also use HPSS as a space management backend.

Versity Storage Manager (VSM) is a software platform that automates the process of storing and retrieving archival data [32]. Versity's product VSM2 comprises an open-source archiving file system with a POSIX interface called ScoutFS [33] and the proprietary Scout Archive Manager ScoutAM. A design target of ScoutFS, which makes it particularly interesting for the future, is advanced indexing capabilities to allow for quick discovery of inode attribute and file content changes. Version 1 of the product was based on SAM-QFS and offers a migration path for installations using SAM-QFS/Oracle HSM.

4.4. I/O Acceleration Solutions

While in the past typical HPC data centres realised storage infrastructures based on an online tier using HDDs and an offline tier using tapes, the increasing need for performance requires adding another shared storage tier or facilitating use of node-local storage. This storage is based on fast non-volatile memory technologies to realise high-performance both in terms of bandwidth and throughput of I/O operations. While the storage is persistent, the data is expected to remain there for short periods of time and is typically staged from or migrated to a slower but much larger storage system. The corresponding solutions can therefore be considered to be I/O accelerators.

In the following part we provide different I/O accelerator solutions that are being used for HPC infrastructures and are expected to continue being relevant in the future. Section 2.3 has further coverage on this subject.

IME (Infinite Memory Engine) from DDN is meanwhile used at various leading supercomputing centres [34]. It is designed as an intermediate storage layer between an HPC system and an external storage system and is implemented by servers with a larger number of NVMe SSDs. Data stored in IME can be accessed either through the IME native interface or via a POSIX client. It uses the namespace of the backing file system, which could be Spectrum Scale or Lustre.

DAOS is a software-defined object store solution optimised for distributed non-volatile memory [35]. It was developed mainly by Intel and has been open-sourced. Applications can access datasets stored in DAOS either directly through the native DAOS API or by using I/O libraries (e.g. POSIX emulation, MPI-IO, HDF5) or frameworks (e.g., Spark, TensorFlow). DAOS will be used for the upcoming Aurora system at ANL, which is planned to be the first US Exascale system. DAOS is supported by multiple HPC system vendors including HPE and Lenovo.

Excelero NVMesh is a software-defined storage solution that allows to dynamically create a block storage volume on top of distributed NVMe SSDs. It can be implemented using any of the high-speed network technologies commonly used in HPC systems. Different storage solutions can be deployed dynamically on top of block store volumes. STFC in the UK uses, for instance, BeeGFS on top of NVMesh [36]. Excelero is an SME in the US, which offers its solution also through HPC systems vendors like Lenovo.

Atos Smart Data Management Suite comprises two solutions for I/O acceleration [37]. At hardware level they are both realised by servers that host a set of high-performance SSDs and are integrated in the HPC system's high-performance network. When using the Smart Burst Buffer solution (SBB), the SSDs are used as an intermediate cache transparent to the user. It relies on I/O calls interception through the scheme implemented in the Bull IO Instrumentation library. With the Smart Bunch of Flash (SBF) applications can be enabled to explicitly request a static allocation of NVMe storage.

4.5. Data Management Solutions

In this section we consider different solutions that provide user interfaces and tools for managing data. As of today, none of these solutions can claim a wide uptake in HPC infrastructures. However, with the growing importance of collaboratively managing data, the increasing need for enabling data analytics on structured and unstructured data as well as the support of the FAIR principles for data access, these solutions are expected to become more relevant.

iRODS (Integrated Rule-Oriented Data System) is an open-source data grid middleware. It is based on an abstraction for data management processes and policies. It provides users with a uniform interface to heterogeneous storage systems (both POSIX and non-POSIX) [38]. It allows federating a distributed storage infrastructure under a unified namespace. It also includes, to give a few examples, a workflow engine where rules that trigger actions can be added when defined conditions apply and the possibility to define microservices that run inside the iRODS system. A key focus for iRODS in the future is improved support of metadata for managing data [39]. One example of a large European project using iRODS is EUDAT. iRODS is developed by a consortium of private and publicly funded organisations. Limited commercial support is available through a partner program.

Rucio is a framework for scientific data management developed in the high-energy physics community [40]. The impetus for the original work was the ATLAS experiment at CERN and its storage requirements. It was designed to integrate easily with other already existing components and to provide high level integration. Workflow and physical storage are handled by other systems, but for example rules on how many replicas a dataset should have and where to find them are in Rucio. Several access protocols (including WebDAV and S3) as well as different types of authentications (including username and password, SSH-RSA public key exchange) are supported. Therefore, Rucio could be a candidate for HPC infrastructures. Rucio is developed by a scientific community and no commercial support is available.

Starfish is a solution for managing data in the context of very large-scale storage systems possibly based on multiple file systems (including Spectrum Scale or Lustre), object stores or tape libraries [41]. It is designed to scale to billions of files or objects. Starfish allows users and applications to assign tags and key-value pairs to files and directories to classify content, drive batch processes, and enforce policies. Starfish can be used to migrate files between multiple storage devices, synchronise and replicate data between locations or different file systems, and archive stale or old data manually or automatically based on metadata attributes. Starfish Storage is a US company founded in 2013 and positions its product also in the HPC market.

Nodeum is a storage management framework that can work on top of multiple data stores based on different storage types, including POSIX file systems or object stores with S3 compliant interfaces [42]. A global view is implemented through a virtual file system layer. Nodeum is a small company in Belgium, which starts to explore the use of its product in the context of HPC infrastructures.

Except for more general-purpose solutions, domain specific solutions continue to play a critical role for realising important HPC workflows. Two specific examples are listed below.

MARS is the Meteorological Archival and Retrieval System developed at ECMWF (European Centre for Medium-Range Weather Forecasts). It stores GRIB and NetCDF files [43]. While the stored files can be retrieved as-is the main usage is to query MARS for certain parameters over time ranges, where the output files are synthesised from data in a stored file.

The **Earth System Grid Federation** has developed a publication system for climate research data, partly supported by the Horizon 2020 IS-ENES projects [44]. Data is shared by research groups across the world with QA processes before publication and checks for not publishing data sets with errors. Storage and global search indexes are distributed among federation sites.

5. Trends

With the exponential growth of data, distributed/parallel storage systems have become not only an essential part, but also one of the bottlenecks of large-scale supercomputing centres. High latency data access, poor scalability, difficulty managing large datasets, and lack of query capabilities are just a few examples of common hurdles. Traditional storage systems have been designed for HDD media and for POSIX I/O. These storage systems represent a key performance bottleneck, and they cannot evolve to support new data models and next generation workflows. A strong trend is observed leading towards very high-performance media, based on NVMe solutions. By designing new hardware interfaces and creating new software solutions such as I/O accelerators higher performance than previously can be achieved.

Storage requirements in terms of both capacity and performance will continue to increase, and the storage stack will be expanded to include more levels in the hierarchy as Exascale systems appear. Data is becoming more important in itself and not only an adjunct to the computation. Moving data around is becoming more costly and creating multiple copies for different access methods does not scale. Storage systems are starting to support multiple access methods (such as file system I/O and S3 protocol) to the same data.

The importance of long-term handling of data will be greater in the future with the increased move towards making data more publicly available. FAIR data principles increase the importance of handling data in a structured way during its entire lifetime. Finding data and making it reusable requires extensive meta data, and to access the data publicly documented protocols are needed. In Section 4 we have looked at a number of data management technologies that provide basic storage, handling of meta data and multiple access protocols.

For the foreseeable future, storage systems based on HDD technology and/or tape libraries will provide space for storing data with suitable cost/latency trade-offs.

6. Conclusion

Here the general conclusions for the areas investigated in this report are summarised.

Storage infrastructure:

1. Data infrastructures needs to strike a balance between capacity and performance, with the right balance depending on their tiering level. The growing necessity for efficient operation with large datasets leads to purely flash-based solutions for warm data due to latency and bandwidth considerations.
2. The latest and fastest technologies, starting from NVMe disks, through cluster file systems such as DAOS, will become the basis for building new, ultra-fast Exascale systems.
3. Both traditional HDD based storage systems and tape libraries remain competitive for large volume storage, less IOPS intensive use cases and storage of cold data.

Data management and access:

4. Data must be both findable and accessible, and software support for managing meta data is required. Some scientific workflows may benefit greatly from domain specific solutions, but unless resources for maintaining such tools are provided, a more general solution is recommended.
5. HPC systems traditionally use parallel file systems, which can be extended with I/O accelerators for faster access during computations and with tiering support for automatically moving data to lower cost media. Enabling access to this data through object storage interfaces allows more software workflows to use the data directly.

References

[1] A. Tekin, A. T. Durak, C. Piechurski, D. Kaliszan, F. A. Sungur, F. Robertsen and P. Gschwandtner, "State-of-the-Art and Trends for Computing and Network Solutions for HPC and AI, PRACE Technical Report," 2020.

[2] E. Krishnasamy, S. Varrette and M. Mucciardi, "Edge Computing: An Overview of Framework and Applications," PRACE Technical Report, 2020.

[3] Square Kilometer Array, "Software and Computing," [Online]. Available: https://www.skatelescope.org/software-and-computing/. [Accessed 16 11 2020].

[4] The Register, [Online]. Available: https://www.theregister.com/2015/11/03/intels_allflash_data_center/.

[5] P.-Y. Du, H.-T. Lue, Y.-H. Shih, K.-Y. Hsieh and C.-Y. Lu, "Overview of 3D NAND Flash and progress of split-page 3D vertical gate (3DVG) NAND architecture," in *12th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, Guilin, China, 2014.

[6] Intel, "Reimagining the Data Center Memory and Storage Hierarchy," [Online]. Available: https://newsroom.intel.com/editorials/re-architecting-data-center-memory-storage-hierarchy/. [Accessed 11 2020].

[7] "DAOS For Applications," 6 Feb 2020. [Online]. Available: https://ecpannualmeeting.com/assets/overview/sessions/DAOS_ECP.pdf.

[8] "INSIC Technology Roadmap 2019," [Online]. Available: http://www.insic.org/wp-content/uploads/2019/07/INSIC-Technology-Roadmap-2019.pdf.

[9] Thinkparq, "BeeGFS," [Online]. Available: https://www.beegfs.io/c/. [Accessed 16 11 2020].

[10] S. Weil, S. Brandt, E. Miller, D. Long and C. Maltzahn, "CRUSH: Controlled, scalable, decentralized placement of replicated data," Tampa, FL, 2006.

[11] Ceph Project, "Ceph Architecture," [Online]. Available: https://docs.ceph.com/en/latest/architecture/. [Accessed 16 11 2020].

[12] "FAIR Principles," [Online]. Available: https://www.go-fair.org/fair-principles/.

[13] M. Malms, "ETH4HPC's SRA 4. Strategic Research Agenda for High-Performance Computing in Europe," 2020.

[14] Amazon, "Amazon Simple Storage Service. Developer Guide".

[15] "Swift," [Online]. Available: https://docs.openstack.org/swift/latest.

[16] "Fenix," [Online]. Available: https://www.fenix-ri.eu.

[17] D. Hildebrand, "A Deployment Guide for IBM Spectrum Scale Unified File and Object Storage," 2017.

[18] "XSTOR," [Online]. Available: https://atos.net/en/solutions/high-performance-computing-hpc/bullsequana-xstor.

[19] S. Narasimhamurthy, "The SAGE project: a storage centric approach for exascale computing," in *Proceedings of the 15th ACM International Conference on Computing Frontiers 2018*, 2018.

[20] "OpenIO," [Online]. Available: https://www.openio.io.

[21] “Swiftstack,” [Online]. Available: https://www.swiftstack.com.

[22] I. Foster and C. Kesselman, The Grid 2, Morgan Kaufmann, 2003.

[23] “dCache,” [Online]. Available: https://www.dcache.org.

[24] P. Carns, “BYOFS: The opportunities and dangers of specialisation in the age of exascale data storage”.

[25] F. Tessier, “Dynamically Provisioning Cray DataWarp Storage,” p. arXiv:1911.12162, 2019.

[26] A. Brinkmann, “Ad Hoc File Systems for High-Performance Computing,” vol. 35, no. 1, 2020.

[27] “BEEOND,” [Online]. Available: https://thinkparq.com/products/beeond/.

[28] N. Haustein, “IBM Spectrum Scale Information Lifecycle Management,” London, 2019.

[29] B. Evans, “HSM, Data Movement, Tiering and More”.

[30] HPE, “HPE Data Management Framework 7,” November 2019. [Online]. Available: https://support.hpe.com/hpesc/public/docDisplay?docLocale=en_US&docId=a00056652enw.

[31] “HPSS,” [Online]. Available: http://www.hpss-collaboration.org.

[32] “VERSITY,” [Online]. Available: https://www.versity.com.

[33] “ScoutFS,” [Online]. Available: https://www.scoutfs.org.

[34] DDN, “IME Datasheet,” 19 08 2020. [Online]. Available: https://www.ddn.com/download/ime-datasheet/.

[35] “DAOS,” [Online]. Available: http://daos.io.

[36] “NVMeshSTFC,” [Online]. Available: https://www.excelero.com/wp-content/uploads/2019/11/GPU-Servers-for-Machine-Learning-and-AI.pdf.

[37] “SDMS,” [Online]. Available: https://atos.net/wp-content/uploads/2019/06/Smart-Data-Management-Suite.pdf.

[38] “iRODS,” [Online]. Available: https://irods.org.

[39] T. Russell, “Beyond Discoverability: Metadata to drive your data management,” 2020.

[40] M. Barisits, “Rucio: Scientific Data Management,” *Computing and Software for Big Science,* p. 3:11, 2019.

[41] “Starfish,” [Online]. Available: https://starfishstorage.com.

[42] “Nodeum,” [Online]. Available: https://www.nodeum.io.

[43] “MARS,” [Online]. Available: https://confluence.ecmwf.int/display/UDOC/MARS+user+documentation.

[44] “ESGF,” [Online]. Available: https://esgf.llnl.gov.

List of acronyms

AHCI	Advanced Host Controller Interface
AI	Artificial Intelligence
ANL	Argonne National Laboratories
BeeOND	BeeGFS on Demand
CES	Cluster Export Services
DAOS	Distributed Asynchronous Object Storage
DMF	Data Management Framework
DNE	Distributed Namespace Environment
DoE	Department of Energy
DWDP	Disk Write Per Day
ECMWF	European Centre for Medium-Range Weather Forecasts
EOFS	European Open File System
EU	European Union
GDPR	General Data Protection Regulation
HBM	High Bandwidth Memory
HDD	Hard Disk Drive
HDR	High Data Rate
HPSS	High Performance Storage System
HSM	Hierarchial Storage Management
ILM	Information Lifecycle Management
INFRAG	Infrastructure Advisory Group
IOPS	I/O operations per second
iRODS	Integrated Rule-Oriented Data System
LTO	Linear Tape Open
MARS	Meteorological Archival and Retrieval System
MDT	Metadata Target
MLC	Multi Level Cell
NVMe	Non-Volatile Memory Express
OS	Operating System
OST	Object Storage Target
OU	Organisational Unit
PCC	Persistent Client Cache
PRACE	Partnership for Advanced Computing in Europe
QoS	Quality of Service
RAIT	Redundant Array of Independent Tapes
RDMA	Remote Direct Memory Access
REST	REpresentational State Transfer
RIAG	Research and Innovation Advisory Group
RoCE	RDMA over Converged Ethernet
RPC	Remote Procedure Call
SAS	Serial Attached SCSI
SATA	Serial ATA
SKA	Square Kilometer Array
SLC	Single Level Cell
SMT	Simultaneous Multithreading
SoC	System on Chip
SR-IOV	Single Root Input/Output Virtualization
SRA	Strategic Research Agenda
SRT	Intel Smart Response
TCO	Total Cost of Ownership
TCP/IP	Transmission Control Protocol/Internet Protocol
TLC	Triple Level Cell
UI	User Interface
VM	Virtual Machine
VSM	Versity Storage Manager

Acknowledgements

This work was financially supported by the PRACE project funded in part by the EU's Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 823767.

Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

State-of-the-Art and Trends for Computing and Interconnect Network Solutions for HPC and AI

A. Tekin [a*1], A.Tuncer Durak [a*2], C. Piechurski [b*3], D. Kaliszan [c*4],
F. Aylin Sungur [a*5], F. Robertsén [d*6], P. Gschwandtner [e*7]

[a]National Center for High Performance Computing – UHEM, [b]GENCI, [c]Poznań Supercomputing and Networking, [d]CSC,
[e]Universität Innsbruck – UIBK

Abstract

Since 2000, High Performance Computing (HPC) resources have been extremely homogeneous in terms of underlying processors technologies. However, it becomes obvious, looking at the last TOP500, that new trends tend to bring new microarchitectures for General Purpose Processors (GPPs) and new heterogeneous architectures, combining accelerators with GPP, to sustain both numerical simulation and Artificial Intelligence (AI) workflows. The present report provides a consolidated view on the current and mid-term technologies (2019-2022+) for two important components of an HPC/AI system: computing (general purpose processor and accelerators) and interconnect capabilities and provides an outlook on future trends in terms of mid-term projections about what users may expect in the coming years.

[1] adem.tekin@be.itu.edu.tr
[2] a.tuncer.durak@uhem.itu.edu.tr
[3] christelle.piechurski@genci.fr
[4] damian.kaliszan@man.poznan.pl
[5] aylin.sungur@itu.edu.tr
[6] fredrik.robertsen@csc.fi
[7] philipp.gschwandtner@uibk.ac.at

Table of contents

1. Introduction

This technical report is part of a series of reports published in the Work Package "HPC Planning and Commissioning" (WP5) of the PRACE-6IP project. The series aims to describe the state-of-the-art and mid-term trends of the technology and market landscape in the context of HPC and AI, edge-, cloud- and interactive computing, Big Data and other related technologies. It provides information and guidance useful for decision makers at different levels: PRACE aisbl, PRACE members, EuroHPC sites and the EuroHPC advisory groups "Infrastructure Advisory Group" (INFRAG) and "Research & Innovation Advisory Group" (RIAG) and other European HPC sites. Users should refer to this series of reports as an overall view of HPC technologies and expect some of the solutions described to be available to them soon. The present report covers "State-of-the-Art and Trends for Computing and Network Solutions for HPC and AI". Further reports published so far are covering "Data Management Services and Storage Infrastructures" [1] and "Edge Computing: An Overview of Framework and Applications" [2]. The series will be continued in 2021 with further selected highly topical subjects.

Since 2000, High Performance Computing (HPC) resources have been extremely homogeneous in terms of underlying processors technologies being mostly based on clusters of nodes equipped with microprocessors. However, it becomes obvious, looking at the last TOP500 (June 2020) [3], that new trends tend to bring new microarchitectures for General Purpose Processors (GPP) and new heterogeneous architectures combining accelerators/GPUs with GPP to sustain both numerical simulation and Artificial Intelligence (AI) workflows.

While the GPP market was mostly led by Intel and its X86_64 processor family for more than 15 years and the GPU market was mainly sustained by NVIDIA until recently, there is a lot of existing companies and newcomers proposing new chips capable to satisfy application computing needs while been extremely efficient in terms of GFlops/Watt. With a large amount of information available on these technologies from various sources, the present report provides an overall and consolidated view on the current and mid-term technologies (2019-2022+) available for two important components of an HPC/AI system: computing (GPP and accelerators) and interconnect technologies. This report does not claim to be an exhaustive view of what is available today though covering the most popular and know current and future technologies.

Computing technologies are introduced first (Section 2) through key factors to consider for the analysis of processor performance and their relation to architectural choices. Section 2 aims to familiarise the reader with processor performance aspects, highlighting the most important problems and the proposed solutions. Besides covering competing technologies, it mentions trends and failed attempts in the past, define the state-of-the-art, and conclude with general projections into the future considering theoretical constraints and established trends. This section also, sparingly, draws on comparisons of coprocessor technologies.
Sections 3 and 4 discuss the current and near-future computing technology products for general purpose processors and accelerators/GPUs/FPGAs (Field Programmable Arrays). They both include technical specifications not discussed on purpose in Section 2.
The last section (Section 5) focuses on interconnects at two important levels: the first considers the high-speed and low-latency interconnects used to run massive MPI computations and the second focuses on local interconnects of computing components needed to improve data movement and ensure cache coherency within a single node.

Finally, building on top of the understanding of theoretical concepts and commercially available - or soon to be available - products, the conclusion section provides an outlook on future trends and summarises mid-term projections (3 to 5 years) about what users can expect to be available in the near future.

Readers may notice that some technologies are covered in more details than others. This is mainly due to the following reasons:

(1) Time elapsed since release on the market (the longer, the more information is available); as an example, x86 has been widely adopted by the market since the 2000s until today,
(2) Adoption by the market: while IBM Power and Arm technologies have been on the market for a while, both technologies are not so widespread nowadays; just looking at the June 2020 Top500, there are 10 supercomputers in total that are based on Power processors while 469 (93,8%) supercomputers are powered by x86 and only 4 are Arm-based.
(3) Size of the market (the larger, the more information is available) as information on widespread technologies is easier to find that information on niche/emergent technologies, for which technical details are generally less accessible.

2. Key Factors in Processor Performance

2.1. Manufacturing Process

The discussion about processor technologies should start with the manufacturing process, as it governs every aspect of processor capability. The basic element of processor design is the transistor, with the connection layout of transistors forming the architecture. Moore's Law [4] offers an observation on the industrial progress rate rather than a physical law and often is misquoted to claim that "processing power doubles every N months". Actually, this law refers to the number of transistors in a processor doubling every 18 months. The number of transistors in a processor is defined by the size of its die and the possible density of transistors. While Moore's Law was a good match for the actual development for the past decades, it has no more reflected the reality for the last 2 or 3 years. This observation is based on transistor density's limitation, closely linked to manufacturing process technology. The die area of processors has remained largely constant, as it is limited by the communication times across the die, in addition to manufacturing constraints (larger dies mean lower yield). The transistor density, in turn, is limited by the minimum feature size at which a transistor can operate reliably. Previous projections predicted a limit of manufacturing processes at 5nm, where they were expected to suffer from quantum tunnelling effects resulting in substantial power drain and heating effects which would both influence high-speed and reliability of results. However, a 1 nanometre wide transistor was shown to be possible in the past and Taiwan Semiconductor Manufacturing Company (TSMC) has already announced that 3-nanometre chips will come on the market in 2021. Furthermore, the junction – the actual gate that is switched on or off to create transistor functionality – could be as small as a single atom, but the gate contacts will remain relatively large on such designs. The possible breakthroughs that may overcome the density barrier are the use of photonics instead of electronic information transfer and utilisation of 3D (or pseudo-3D) stacked designs. The former is theoretical and the latter finds a limited use in memory design at the time of this technology report.

Currently, the most widely deployed processors in HPC, e.g. from Intel, are manufactured at 14nm level (Intel Cascade Lake), with a 10nm manufacturing process announced in Q4 2020 for Intel Ice Lake while other foundries or manufactures like TSMC and Samsung already offer 5-7 nanometre designs. However, it should be noted that the labels for commercial products, including marketing labels such as 7nm+ or the 10nm "refresh" iterations from Intel should not be taken to depict literal transistor sizes. Designs labelled as 14nm, for example, were "substantially" larger than 14 nanometres and Intel's 10nm is comparable to other foundries' 7nm, while Intel's 7nm is comparable to other 5nm processes. On this subject, Godfrey Cheng from TSMC, is quoted as follows: "I think we need to look at a different descriptor for the nodes beyond what we currently see today", here the term "nodes" refers to the transistors, currently described (inaccurately) by their size.

2.2. CPU Frequency

A dominant factor in processing power is the operational frequency. It strongly impacts the number of operations per second. However, the technological limits, most importantly power and temperature budgets, have also led to a stall in the increase in frequency. The 3.6+ GHz levels, common in the era of Intel's Pentium 4, have been replaced by multicore designs operating at modest 2.4 GHz, making up the reduced speed at which instructions complete by retiring more instructions per cycle. Another recent trend is the utilisation of "boost levels", where parts of the processor are able to operate at higher frequencies as long as the temperature and power remain in the limits of the processor. These "boost levels", coupled with the recent advancements in manufacturing processes, may signal the return to frequency levels of the previous era.

2.3. Instruction Sets

Besides the increased core counts to be discussed below, the main boost in recent processor performance comes from complex new instructions. These allow higher level operations to be expressed in denser instruction series (most of the time reducing the stress on the memory bandwidth).

A simple, yet powerful, example is the fused-multiply-and-add (FMA) instruction, which executes the $a = b * c + d$ operation in a single step. Most modern architectures support this instruction in some form (e.g. FMA3 updating the target in place, or FMA4 writing the result to a new location), while the design decisions have evolved over time (i.e. Intel preferred FMA3 in AVX2, while AMD initially used FMA4 but now allows both). NVIDIA took this instruction a step further in its GPU architecture, executing FMA over 4x4x4 tensors in a single step. In the future we might expect CPU architectures to also incorporate such tensor operations as pressure from AI and

Machine Learning (ML) users increases. Currently, as a result of these demands, Intel and some others (e.g. Arm and IBM) have included some additional ISA instructions for these specific workloads, such as Vector Neural Network Instructions (VNNI) as part of AVX-512 for Intel. However, it should be noted that these instructions and the tensor operations provided by NVIDIA, operate on half precision floating point numbers for intermediate values as double precision implementation through Floating Point 64bits (FP64) Tensor Core. This precision level, sufficient for common AI applications, is also deemed acceptable for Molecular Dynamics (MD) simulations in some cases, covering another large portion of the HPC portfolio. The usage of reduced precision for numerical simulation is also an active area of research in terms of numerical methods. The half precision is formalised in the BFLOAT16 format, which instruction sets from Intel and Arm both supports. AMD’s support for BFLOAT16 is currently limited to its GPU products, but future CPUs from AMD may also follow this trend.

2.4. Vector Length

The operations on lower precision floating point numbers also influence the vector instructions. The optimal length of vectors in vector operations is a subject of an ongoing debate. The length of vector operands in such instructions has been steadily increasing in the x86 architecture for the past few years, eventually reaching 512 bits width in AVX-512 extensions implemented by Intel. These 512 bits wide instructions debuted with the release of Intel's Xeon Phi architecture and were carried over to the Xeon architecture in the Skylake series. However, Intel's implementation came with the caveat that the operating frequency had to be reduced during the execution of these instructions which led to an AVX frequency much below the nominal frequency, allowing the processor to run under its specific TDP power. Regarding extensions to the x86 instruction sets, AMD has typically supported Intel’s new instructions for compatibility, and subsequently AVX2 (Intel's 256 bits wide vector extensions) were implemented in AMD’s Rome processors. AVX-512 has been an exception until now and since it did not reach the expected level of performances for most of the applications, it is not sure to be adopted in a future AMD design. Furthermore, Intel might also focus on 256 bits wide instructions, leaving 512 bits instructions for compatibility and special cases. It should be noted that other architectures, besides x86, have competing (yet not directly comparable) solutions, with SVE (Scalable Vector Extension) from Arm supporting a variable width ranging from 128 to 2048 bits as opposed to x86's fixed vector length.

The issue with introducing complex higher-level instructions, besides introducing complexity taking up valuable real estate on the die, is that such complex operations need a large number of cycles to complete. During the heated debate between Complex Instruction Set Computer (CISC) and Reduced Instruction Set Computer (RISC) design philosophies, the RISC proponents argued against the complex instructions for this very reason. Modern CPU architectures are CISC architectures, featuring a very diverse set of highly specialised functions. However, these CPUs handle these instructions by decoding them into smaller, RISC-like sub-operations and dispatching them into appropriate Execution Units (EUs). This allows software to be represented in a terse chunk of code in memory, reducing the cost in memory accesses. At the same time, it allows finer control over execution order inside the processor, thus increasing efficiency by hiding latency via pipeline exploitations or even multiplexing access to resources by Simultaneous Multithreading (SMT). Therefore, additional criteria for comparing architectures and their several iterations are pipeline depth and number of threads in SMT. The discontinued Netburst architecture from Intel took the pipeline depth to an extreme level by featuring 31 stages. The discontinuation of this architecture should indicate the caveats in implementing very long pipelines, and current architectures follow a more modest approach (e. g., 14-19 in Intel’s Ice Lake and 19 in AMD’s EPYC).

2.5. Memory Bandwidth

Memory bandwidth has been one of the key factors for CPUs to perform efficiently, both benefiting to memory bound applications as large-scale use cases to feed the CPU cycles as fast as possible and keep up with processors increased computing power. For the last few years, processor technologies have been designed with higher memory bandwidth: first, through a higher number of channels per processor; while in 2011, an X86 processor like the Intel Sandy Bridge had 4 memory channels, now X86 (Intel/AMD) and Arm processors have, in 2020, typically between 6 and 8 memory channels, allowing a theoretical global memory bandwidth performance improvement of +50% and +100%. The second criterion to take into account is the frequency of the DDR memories which has drastically improved over time. While in 2011, an X86 processor like the Intel Sandy Bridge supported DDR3 running at 1600 MT/s, an X86 processor and an Arm processor are, in 2020, supporting DDR4 running between 2933 MT/s and 3200 MT/s. Taking into account both the increase in the number of memory channels and the DDR technology improvement, the global memory bandwidth per processor for an X86 processor improved by a factor of 4, while the memory bandwidth per core remains nearly the same due to density growth on current available chips. A new trend is driving the market to High-Bandwidth memory (HBM) both within CPU and accelerators, providing high throughput memory access for applications (>= 1TB/s vs maximum 205 GB/s for the current AMD and Marvell ThunderX processors available on the market – minimum 5 times the memory bandwidth of DDR depending on the HBM type) and a more balanced byte-per-flop ratio than then one supported by DDR-only processor technology. However, HBM provides a maximum capacity of 32GB (HBM2) today requiring transfers between DDR and HBM in case the data does not fit the HBM size.

2.6. Simultaneous Multithreading (SMT)

In SMT, multiple instructions are issued at each clock cycle, possibly belonging to different threads; thus increasing the utilisation of the various CPU resources and making it possible to reduce the effect of memory latency. SMT is convenient since modern multiple-issue CPUs have a number of functional units that cannot be kept busy with instructions from a single thread. By applying dynamic scheduling and register renaming, multiple threads can be run concurrently. Regarding SMT, it should be noted that the number of hardware threads in x86 is extremely low (i.e. 2), compared to other architectures: POWER from IBM supports up to 8 threads per core (SMT8) and various Arm-based processor implementations feature up to 4 threads per core (SMT4). It is possible to explain this discrepancy by multiplexing features such as register windows being absent in x86. GPU architectures prefer a simpler, straightforward approach to instruction complexity, spending die area real estate in multiplication of EUs to increase parallelism instead of featuring complex instructions. It is possible to implement this approach in CPUs by building semi-independent vector operation sections in future models.

2.7. Processor Packaging

Another, largely orthogonal step towards achieving an increase in computational power is putting multiple individual cores onto a single die. Since the stall in frequency increase has begun to present itself, building multicore CPUs has been the preferred solution to continue increasing the processing power. However, the increase in core counts was limited by the need for communication and synchronisation between the cores and the difficulty in increasing the number of transistors on a single monolithic-die with a manufacturing process reaching nanometer level. This need arises largely from the constraint of maintaining a consistent shared memory to prevent race conditions. One approach to this subject is making the discrepancy and memory ownership explicit by presenting a Non-uniform Memory Architecture (NUMA), similar to the pre-existing multi-socket arrangements or chiplets. This is achieved through the MCM (Multi-Chip Module) concept. MCM is an electronic assembly of multiple chips or semiconductor dies, also called chiplets, that are integrated usually onto a unifying substrate, so that it can be treated as a larger integrated circuit. The chiplets are then connected through an intra-chiplet interconnect, as for example the Infinity Fabric (IF) interconnect for AMD zen2 and further AMD generation processors. While MCM has been early adopted by companies like IBM or AMD to increase core density (rather than clock speed) on a processor, Intel has decided so far, to remain with its monolithic chip architecture for general purpose processors (except for their Cascade Lake-AP processor) despite all the complexity of the manufacturing process faced at 14 and 10nm levels. MCM presents large advantages: It has helped AMD both to enter the market earlier than its main competition and reduce the price of their processor. MCM now underwent a broader adoption by the market and is a manufacturing process well mastered by foundries like TSMC. Some drawback at application level: As the processor presents several NUMA domains, it requires a strong knowledge of the processor micro-architecture to support suitable task placement.

2.8. Heterogeneous Dies

Planting heterogeneous cores that are specialised in different application areas in a single die is not a new approach. The Cell Broadband Engine Architecture from IBM, usually called Cell in short, has combined a two-way SMT general purpose PowerPC core with eight Synergistic Processing Elements (SPEs) specialised in vector operations on a single die. Unfortunately, despite being able to provide high performance, it has been a victim of its radical approach and has never become popular, suffering a fate similar to Itanium from Intel.

The general philosophy of combining heterogeneous compute elements, however, is not abandoned. In fact, it has been observed multiple times in the computing industry that the rise of co-processors has been inevitably followed by integrating them into the central processor, resulting in cyclical trends. The rise of Integrated Graphics Processing Units (iGPUs) could be given as a general example, and the fate of Xeon Phi architecture represents an interesting twist in the HPC world. The implementation of heterogeneous dies takes the form of big.LITTLE in the Arm architecture, combining low frequency energy efficient cores with high performance ones, but while reducing power consumption, the utilisation is limited in the HPC area. A different approach is exemplified in the efforts of AMD, where vector operation focused GPU cores are being moved into the unified address space and die space of the central processor.

In terms of the aforementioned observation of cyclical trends, the industry is at the early stages of the integration phase, where discrete co-processor products from NVIDIA dominate the market, but the demands for unified memory address space and simpler programming models put pressure on the vendors. As an extreme example, the ominously named 'The Machine' from HPE has proposed a universal address space operated on by various specialised processors, connected not by a single continuous die, but a high-speed interconnect based on photonics. The future of this ambitious project, however, is unclear: widespread adoption is unlikely, based on the fate of such radical departures from the traditional model in the past.

3. General Purpose Computing Capabilities

3.1. X86_64 Processors

3.1.1. Intel X86_64

After years of delays, Intel's 10nm designs have finally seen the light of day in the Ice Lake series of the 10th generation Intel Core processors for desktops. However, even after this long delay, the new designs' yields and clock speeds have been generally unimpressive, resulting in the products of the ageing 14nm+ (and further iterations denoted by additional + symbols in their names) to continue being offered alongside with the newer, 10nm process-based ones for desktop and server platforms. This has also been the case for the HPC market for almost 4 years now, starting with the 5th Intel core processor generation code named Broadwell released in 2016 and ending with the Intel Copper Lake-SP processor for the HPC segment market that will be delivered at the earliest at the end of Q2 2020 for key customers in specific configurations (Cedar Island Platform only). Lack of satisfactory progress in this area has also been admitted by Intel with its CFO, George Davis, recognising that the 10nm process has not been as profitable as its long exploited 22nm and 14nm processes were [5].

The first 10nm CPU for HPC workloads, code named Ice Lake-SP (Whitley Platform) [6], should be available by the end of 2020. Its main purpose should be driving the path to 10nm mass production with the expected new Intel product called Sapphire Rapids [7] (together with a new LGA 4677 socket) that should be deployed in 2 phases: the first version might be based on Sapphire Rapids without HBM (High Bandwidth Memory), supporting up to 8 memory channels DDR5 and PCI-e gen5 as NVDIMM memory. The second step may add several important features such as the capability to interface with the new Intel GPU Called Intel Xe HPC "Ponte Vecchio" (PVC) (see the GPU section below for more details). The latter ensures a unified memory architecture between the Sapphire Rapids CPU and Intel PVCs GPU through the Xe links based on the CXL (Compute Express Link) standard.

In terms of the manufacturing process, the currently released plans from Intel state that 7nm might be available in 2021 and that Intel's 5nm should be released in 2023. When putting these improvements into perspective, Intel mentioned that its 10nm process is comparable to competing products labelled as 7nm in TSMC, also that their 7nm process is roughly equivalent to TSMC's 5nm process, with Intel's 5nm being similar to TSMC's 3nm [8]. As for frequency, the high-end products from Intel reach 3.0 GHz, but 2.5 to 2.7 GHz models are expected to be the common choice. As with the current trend, the focus is more on the boost levels, achieving 4.0 GHz for a single core. At least for the lifetime of the 10nm process, these frequency levels should not change radically. However, some of the improvements to the processor performance may come in the form of instructions that do require lower operating frequencies (as AVX-512 does today). In addition to extensions to AVX-512, widespread adoption of BFLOAT16 (starting within the Intel Copper Lake processor which should be delivered to large key AI customers) and Galois Field New Instructions make up the major changes to the instruction set in new products. A radical addition, however, is the inclusion of Gaussian Neural Accelerator v1.0 (GNA) in client version of Ice Lake. However, GNA is a low power, inference focused component, and it is unclear if it will be included in the server version.

3.1.2. AMD X86_64

The current AMD EPYC 7002 Series Processors (Rome) is the second generation of the EPYC x86 platform. Its implementation relies on an MCM (Multi-Chip Module) implementation and on the Zen2 core architecture built in a 7nm process technology to support up to 64 compute cores on a single CPU, enhanced IO capabilities through 128 lanes of PCI-e Gen4 I/O and 8 DDR4 memory channels (with up to 2DIMMs per channel) running at up to 3200MT/s, boosting memory bandwidth up to 204.8 GB/s peak. The chiplet of a MCM in AMD language is called

CCD (Compute Core Die), with one CCD supporting up to 8 compute cores. The AMD Rome processor supports up to 280W TDP (Thermal Design Power) per socket and up to 3.4 GHz Turbo boost CPU clock speed.

The upcoming AMD Zen3 core is expected to enter mass production in Q3/Q4 2020 and might rely on TSMC's advanced 7nm processor, N7+. While the Zen2 cores are the main components of AMD Rome processor, the Zen3 cores will be the main components of the AMD Milan processor [9]. The main differences between Zen2 and Zen3 implementation should be the clock speed and micro-cache level architecture implementation on a CCD as the memory. The former means that at equivalent core counts the Zen3 core should be capable of operating at higher frequencies targeting increased per core performance. The latter means reducing the number of NUMA domains inside a single CPU, having 8 CCD Zen3 cores sharing now 32 MB L3 cache on a single CCD (one core being capable of addressing 32 MB memory) while previously one single core was capable of addressing a maximum of 16 MB L3 cache. While the maximum memory bandwidth was sustained with 16 Zen2 cores on the Rome processors, the optimal number of Zen3 cores to achieve the best memory bandwidth might then be 8 cores. In addition, the Zen3 cores may have the capability to run Secure Encrypted Virtualisation-Encrypted State (SEV-ES) without any specific software changes. This feature would allow to encrypt all CPU register contents when a VM (Virtual Machine) stops running and prevent the leakage of information in CPU registers to components like the hypervisor. It can even detect malicious modifications to a CPU register state.

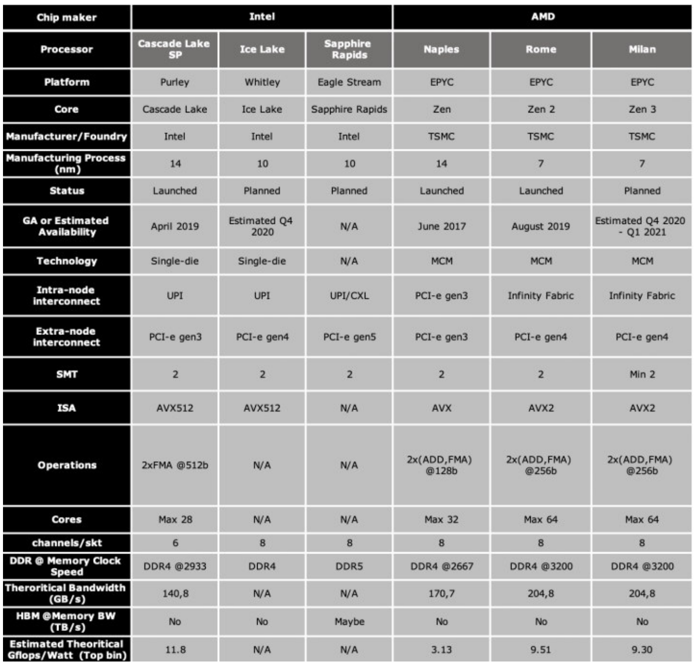
One key point to note is that, while the CPU frequency for Naples was far from the ones seen on Intel Skylake processor, the AMD Rome (and the future AMD Milan) clock speeds are now comparable to Intel's products, being slightly lower, 2.6 GHz for the highest core count top level variant. What makes the AMD Rome/Milan competitive with Intel is the density on the AMD processors, reaching up to 64 cores.

However, there will also be low-core count SKUs (Stock Keeping Units) (24, 16 and 8 cores variants) with higher frequencies. As for the instruction set, the main difference to Intel is that AMD Rome/Milan only support AVX-256 instead of AVX-512. However, the lack of wide vectors is made up by the fact of AMD having its own high-performance GPU line, and the plans for integrating them onto the same die. Furthermore, the higher core count also results in more vector processing units being available even without integrated co-processors. There are no disclosed plans from AMD to include a tensor processor, similar to Intel's GNA or VNNI feature, simply due to the fact that AMD now has its own high-performance GPU line (MIXX line).

The next generation of AMD general purpose processor should be based on Zen4 cores and should form the Genoa processor as announced [10]. This new microarchitecture is expected to be based on a 5nm process technology and might incorporate new features as DDR5, PCI-e gen5, HBM support and cache coherency (between CPU and GPU). It is already well-known to be the CPU that will power one of the three Exascale machines announced by the US DoE (Department of Energy), El Capitan (LLNL) in 2022. The processor on the Frontier machine (another of these three exascale machines) should be a custom Milan SKU, a Milan ++ probably with some of the Genoa capabilities. This Genoa processor should again enhance the strong I/O capability of the AMD processor providing again more IO capabilities and higher memory bandwidth which should benefit memory bound applications. While it is known that this new AMD processor might introduce a big step in the AMD roadmap, there is little public information available on the Genoa processor now [11].

3.1.3. Comparison of main technical characteristics of X86_64 processors

The Table 1 summarises the X86_64 processors main technical characteristics.



Chip maker	Intel			AMD		
Processor	Cascade Lake SP	Ice Lake	Sapphire Rapids	Naples	Rome	Milan
Platform	Purley	Whitley	Eagle Stream	EPYC	EPYC	EPYC
Core	Cascade Lake	Ice Lake	Sapphire Rapids	Zen	Zen 2	Zen 3
Manufacturer/Foundry	Intel	Intel	Intel	TSMC	TSMC	TSMC
Manufacturing Process (nm)	14	10	10	14	7	7
Status	Launched	Planned	Planned	Launched	Launched	Planned
GA or Estimated Availability	April 2019	Estimated Q4 2020	N/A	June 2017	August 2019	Estimated Q4 2020 - Q1 2021
Technology	Single-die	Single-die	N/A	MCM	MCM	MCM
Intra-node interconnect	UPI	UPI	UPI/CXL	PCI-e gen3	Infinity Fabric	Infinity Fabric
Extra-node interconnect	PCI-e gen3	PCI-e gen4	PCI-e gen5	PCI-e gen3	PCI-e gen4	PCI-e gen4
SMT	2	2	2	2	2	Min 2
ISA	AVX512	AVX512	N/A	AVX	AVX2	AVX2
Operations	2xFMA @512b	N/A	N/A	2x(ADD,FMA) @128b	2x(ADD,FMA) @256b	2x(ADD,FMA) @256b
Cores	Max 28	N/A	N/A	Max 32	Max 64	Max 64
channels/skt	6	8	8	8	8	8
DDR @ Memory Clock Speed	DDR4 @2933	DDR4	DDR5	DDR4 @2667	DDR4 @3200	DDR4 @3200
Theroritical Bandwidth (GB/s)	140,8	N/A	N/A	170,7	204,8	204,8
HBM @Memory BW (TB/s)	No	No	Maybe	No	No	No
Estimated Theoritical Gflops/Watt (Top bin)	11.8	N/A	N/A	3.13	9.51	9.30

Table 1: X86_64 Intel and AMD processors main technical characteristics

Note: N/A means the information is not available.

3.2. Arm Processors

3.2.1. *EPI (European Processor Initiative)*

The European Processor Initiative (EPI) [12] is in charge of designing and implementing a new family of low-power European processors designed to be used in extreme scale computing and high-performance Big Data applications as in the automotive industry.
EuroHPC [13] has an objective to power 2 European Exascale machines in 2023-2025, with at least one of them built with a European processor technology, hopefully a result of the EPI. In addition, EuroHPC also plans the acquisition of pre-Exascale systems (2021-2024/2025) and support for the first hybrid HPC/Quantum computing infrastructure in Europe.

The EPI product family will mainly consist of two computing products: an HPC general purpose processor and an accelerator. The first-generation of the general-purpose processor family named Rhea will rely on Arm's Zeus architecture general purpose cores (Arm v8.3/v8.4; up to 72 cores [14]) and on highly energy-efficient accelerator tiles based on RISC-V (EPAC – an open-source hardware instruction set architecture), Multi-Purpose Processing Array (MPPA), embedded FPGA (eFPGA) and cryptography hardware engine. First Rhea chips are expected to be built in TSMC's N7+ technology aiming at the highest processing capabilities and energy efficiency. The EPI Common Platform (CP) is in early development and may include the global architecture specification (hardware and software), common design methodology, and global approach for power management and security in the future. The Rhea chip will support Arm SVE 256 bits (Dual Precision, Single Precision, BFLOAT16), HBM2e, DDR memories and PCI-e gen5 as HSL (High Speed Links), which would support the interconnection of two Rhea dies or one Rhea die with an HPC accelerator like Titan Gen 1 (based on RISC-V instead of Arm). The Zeus cores and the memory subsystems (built on top of HBM, DDR and Last Level of Cache) will be connected through a Memory-coherent-on-chip network. The CP in the Rhea family of processors will be organised around a 2D-mesh Network-on-Chip (NoC) connecting computing tiles based on general purpose Arm cores with previously mentioned accelerator tiles. With this CP approach, EPI should provide an environment that can seamlessly integrate any computing tile. The right balance of computing resources matching the application needs will be defined through the carefully designed ratio of the accelerator and general-purpose tiles. The Rhea chip will support PCI-e and CCIX to interconnect and accelerators.
The second general purpose chip family is named Cronos (Rhea+) and should be based on the Arm Poseidon IP possibly with enhanced capabilities like Compute Express Link (CXL) built-in to ensure cache memory coherency between the CPU and the accelerator.

The Rhea chip and its next generation are designed and commercialised by SiPearl (Silicon Pearl). SiPearl is a European company that is using the results of the European Processor Initiative (EPI) project.

3.2.2. Marvell ThunderX

The current Marvell ThunderX processor on the market is the well-known ThunderX2 processor which has been available since 2018. The ThunderX2 is the second generation of Marvell 64-bit Armv8-A processors based on the 16nm process technology. It is also the first processor which has demonstrated the capability to compete with Intel and AMD. It is available with up to 32 custom Armv8.1 cores running at up to 2.5 GHz and supports Simultaneous Multi-threading (SMT) with 4 threads per core, so twice the number of threads compared to x86 processors. The ThunderX2 die is built on top of a monolithic implementation like all Intel processor generations up to Cascade Lake, in contrast to AMD with its Multi-Chip implementation (MCM). Each processor supports up to 8 memory channels or 8 memory controllers with up to 2 DPC (DIMM Per Channel), with DDR4 DIMMs running at up to 2667 MT/s (1DPC only). The processor has strong I/O capabilities with up to 56 PCI-e gen3 lanes and 14 PCI-e controllers along with integrated I/O and SATAv3 (Serial ATA) ports. Each ThunderX2 core has a dedicated L1 cache (32 KB instruction and data cache) and a dedicated 256 KB L2 cache. The L3 cache is 32 MB and is distributed among the 32 cores. In terms of computation, the Marvell ThunderX2 supports 128 bits NEON instructions (Arm ISA) and up to 2 FMA EUs, which means that each core is capable of executing 2 FMA instructions using a 128 bits vector during a single cycle. This has led to one core being capable of running 8 DP floating operations per second. The Marvell ThunderX2 socket is available as single or dual-sockets server with CCPI2 (Cavium Cache Coherent Interconnect) providing full cache coherency.

The following part of this section, regarding Marvell ThunderX3 processors, was written before the cancellation of ThunderX3 processors by Marvell and presents what was initially planned by Marvell before this cancellation. The authors of this report have decided to maintain the information regarding ThunderX3 processors for several reasons: (1) In case Marvell decides to sell their design to another chip maker, information will be known by users (2) To provide information about what could be achievable at the horizon of 2021.The next ThunderX line processors should have been ThunderX3+ and ThunderX3 [15] [16], based on TSMC's 7nm lithography, again with monolithic chips rather than chiplets. Both CPUs should target different markets: ThunderX3+ and ThunderX3 should have focus on cloud and High-Performance Computing workloads, respectively, due to their internal properties. The ThunderX3 and X3+ were supposed to be based on ArmV8.3+ (+ means that it includes selected features of the ArmV8.4 ISA). The ThunderX3 was planned to be built on top of a single die, with up to 60 Armv8.3+ cores and 8 DDR4 memory channels running at 3200 MT/s supporting up to 2 DPC while the ThunderX3+ is planned to be built on top of 2 dies, each with 48 cores for a total of 96 cores, with also up to 8 aggregated channels (4 per die) DDR4 at 3200 MT/s, leading to the same global memory bandwidth on a SoC but a lower memory bandwidth per core, though giving penalty to this for memory bound applications. On the ThunderX3+, the 2 dies are interconnected through the new CCIP3 (Cavium Cache Coherent Interconnect) over PCI-e gen4. Each processor was designed with 64 PCI-e gen4 lanes over 16 PCI controllers and a 90 MB L3 shared cache while L1 and L2 caches remain single to each core.

Like their predecessor ThunderX2, ThunderX3/X3+ were expected to support SMT4, leading to 384 and 240 threads on ThunderX3+ and ThunderX3, respectively. These processors should have supported NEON (SIMD - Single Instruction Multiply Data) instruction sets with 4 FMA per cycle combined to 128 bits EUs and 16 operations per cycle. The native/base clock speed for ThunderX3+ should rather be around 2 to 2.2 GHz, while this would be increased by 2 bins (200 MHz) for ThunderX3, reaching 2,1 TFlops+ (minimum) peak performance for the HPC version of ThunderX3, for a TDP reaching 200W (minimum). The TDP was expected to depend on the core clock speed provided on the CPU, since while the clock speed will be higher, the TDP might also increase. The design should have come in both 1 and 2-socket configurations, and the inter-socket communication CCPI 3rd generation.

At this point in time, it is not clear if Marvell will pursue their ThunderX line.

3.2.3. Fujitsu A64FX

The Fujitsu A64FX is an Arm V8.2 64bits (FP64) processor designed to handle a mixed processing load, including both traditional numerical simulation HPC and Artificial Intelligence, with a clear target to provide an extremely high energy-efficient performance (performance/watt) and a very good efficiency for a large spectrum of applications. Built on top of a 7nm TSMC process technology, its design has been HPC optimised being the first general purpose processor supporting 32 GB HBM2 (around 1 TB/s aggregate - 4x 256 GB/s) and native hardware SVE, while considering various AI instruction set extensions, such as supporting half precision (FP16) and INT16/INT8 data types. The A64FX has 48 compute cores and 4 additional assistant cores to process the OS and I/O. Like AMD, Fujitsu has chosen to build its processor based on MCMs. These modules are called CMG (Core

Memory Group) in the Fujitsu design. The compute and assistant cores are split into 4 CMGs, each with 12 compute cores and 1 assistant core sharing one 8MiB L2 cache (16-way) through a cross-bar connection and accessing 8GB HBM2 through a dedicated memory controller (maximum 256 GB/s between L2 and HBM2). In addition, each core has its own 64 KiB L1 cache and supports 512-bit wide SIMD SVE implementation 2x FMAs, leading to around 2.7 TFlops/s DP on a single A64FX processor. The 4 CMGs are connected by a coherent NoC capable of supporting Fujitsu's proprietary Tofu interconnect and standard PCI-e gen3 [17] [18] [19].

The Fujitsu A64FX is provided as a single socket platform only, while most of its competitors have chosen to provide single- and dual-socket platforms. The processor powers Fugaku, the first Exascale class machine in Japan and worldwide in 2020 timeframe. The machine has been built on top of more than 150,000 (158,976) compute nodes in more than 400 racks. The nodes are connected by a Tofu-D network running at 60Pbps, reaching 537 PF+ FP64 peak performance with access to more than 163 PB/s theoretical memory bandwidth. Data is stored on a hierarchical storage system with 3 levels: the 1st layer relies on high throughput NVMe (Non-Volatile Memory Express) GFS cache/scratch file systems (>30 PB), the 2nd layer is a high capacity Lustre-based global file system (150 PBs) based on traditional magnetic disks, and the last layer is currently planned to be an archive system stored off-site on a cloud storage. The cooling infrastructure relies on DLC to reach a 1.1 PUE. This system required around 30 MW during benchmarking as reported in June 2020 TOP500 list. Fugaku's installation started in December 2019 and was completed by mid-May 2020. The system should be fully operational and open to the user community in 2021. A key point was the capability of Fujitsu to power a Fugaku like prototype for SC19, which was ranked number 1 in the November 2019 Green500, with a machine based on the General-Purpose Processors only, while most of the other systems at the top of the Green500 list mainly rely on GPUs. The prototype was powered with 768 A64FX CPUs supporting the Arm SVE instructions for the first time in the world. This performance measurement demonstrated that Fugaku technologies have the world's highest energy efficiency, achieving 1.9995 PFlops sustained performance against 2.3593 PFlops as peak performance, and 16.876 GFlops/W (Gigaflops per watt).

In addition, early February 2020, Fujitsu announced that it would supply the Nagoya University Information Technology Center with the first commercial supercomputer powered by the Arm-based A64FX technology. The new system will have 2,304 Fujitsu PRIMEHPC FX1000 nodes, offering a theoretical peak performance of 7.782 PFlops and a total aggregated memory capacity of 72 terabytes. In the meantime, other customers have acquired Fujitsu A64FX systems mostly as test beds for now, e.g. the Isambard 2 system from University of Bristol and the Wombat cluster at Oak Ridge National Laboratory. The 4th Petascale EuroHPC supercomputer, the Deucalion machine at Minho Advanced Computing Centre (Portugal) should be equipped at least with a large partition relying on Fujitsu A64FX processor.

Two other Arm processors, the Graviton from Amazon and the Altra from Ampere, are described in the next sub-sections, even though these 2 processors are more dedicated to compete with AMD and Intel x86 processors for the data centre market rather the HPC market. These 2 platforms are based on Arm Neoverse (ArmV8.1/2) microarchitecture, which is the same Arm platform than the Fujitsu A64FX processor.

3.2.4. Amazon Graviton

As the dominant cloud service provider, Amazon has a keen interest in cost-efficient cloud services and started working on an Arm-based SoC in 2015. Amazon recently announced its second-generation Graviton processor based on Arm's new Neoverse N1 microarchitecture implemented on top of Arm v8.2 and a CMN-600 mesh interconnect. This second generation offers a monolithic die of 64 cores running at 2.5 GHz along with 64 KB of L1 and 1 MB L2 private data caches and a shared L3 cache of 32 MB; it is manufactured in TSMC's 7nm process. Clearly, Amazon also targets accelerated platforms given that the Graviton provides 64 PCI-e 4.0 lanes compared to, for example, the 16 PCI-e 3.0 lanes of the A64FX. Further characteristics of the chip show that it is designed for rather compute-intensive workloads (or co-scheduled compute- and memory-intensive workloads), with 8-16 cores already able to max out the available peak memory bandwidth (which is also the case for the AMD processor). Also, Machine Learning workloads are in focus with explicit support for INT8 and FP16 data types in order to accelerate AI inference workload. Given its single-NUMA design, it is also optimised for low core-to-core latencies across the entire chip compared to more conventional architectures of comparable core numbers that rely on multi-die designs (AMD) or multi-socket designs (Intel), both of which show a significant increase in latency for inter-NUMA communication. This is further illustrated by high scalability results, for example for most SPEC 2017 benchmarks that are not memory-bound. Nevertheless, also sequential performance is competitive with x86-based systems, as the second-generation Graviton shows a large boost over the first generation and preliminary benchmarks reach between 83% and 110% of sequential compute performance compared to systems employing the Intel Xeon Platinum 8259 (Intel Skylake Lake, the microarchitecture previous to Intel Cascade

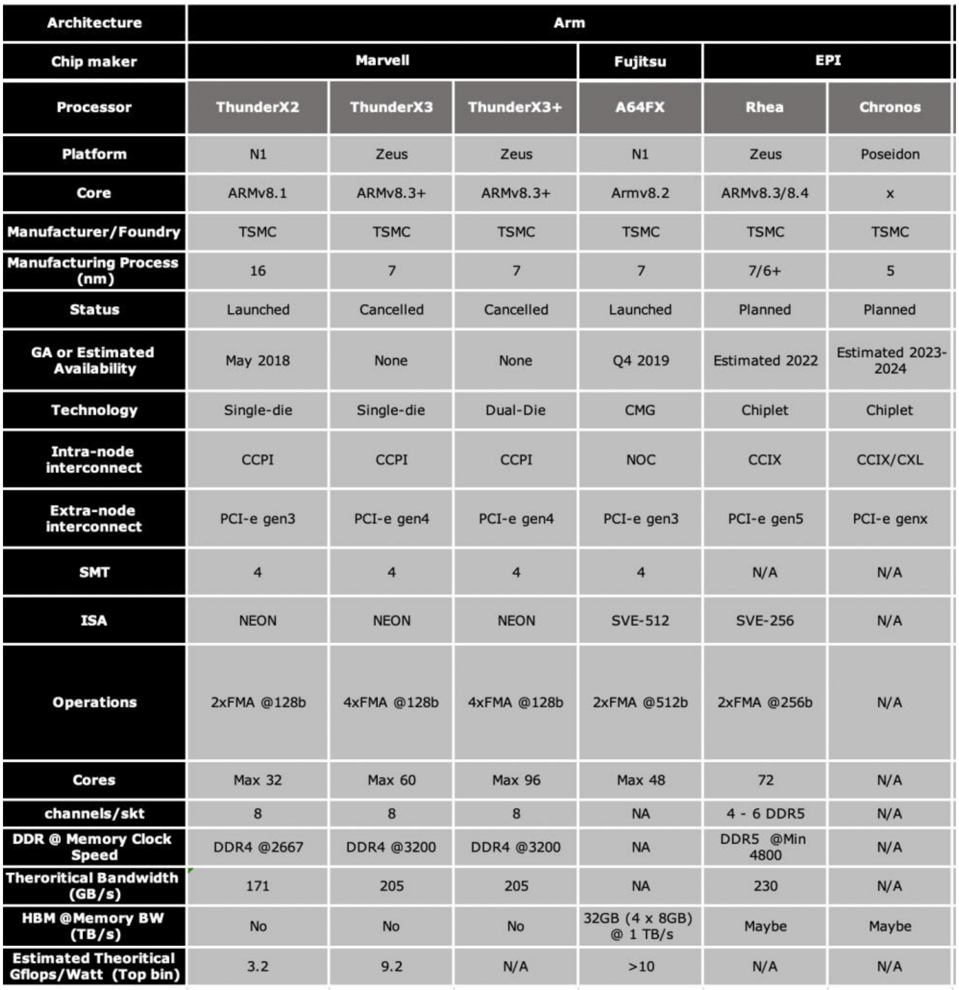
Lake) or AMD EPYC 7571 (AMD Naples) processors, while doubling or tripling the achievable single-core memory bandwidth performance of these Intel and AMD processors.
Both GCC 9.2.0 as well LLVM 9 currently offer support for the Neoverse N1 microarchitecture and hence facilitate a fast adoption rate of software ecosystems for the Graviton. Given Amazon's integration of Graviton in their new M6g instance types, current benchmark results and a "40% better performance per dollar than its competition" claim, this chip might introduce a massive change to Amazon's traditionally x86-heavy cloud solutions and hence the entire data centre market [20].

3.2.5. Ampere Altra

Another Arm CPU architecture is built by Ampere and aims at challenging Intel and AMD for the data centre market. Ampere's new Altra CPU is a TSMC-manufactured 7nm chip with up to 80 cores clocked at a maximum of 3 GHz and a TDP of 210 Watts. Similar to Amazon's second-generation Graviton it implements the Arm v8.2 architecture (along with some additional features not yet part of this standard), supplies each core with dedicated 64 KB of L1 and 1 MB of L2 cache and allocates a total of 32 MB shared L3 cache for all cores (yielding slightly less L3-per-core than the Graviton and falling well below Arm's recommendation of 1 MB per core). It also features 8 DDR4-3200 memory controllers with the same peak memory bandwidth (204.8 GB/s) than AMD Zen2/Zen3 processor. In contrast to the Graviton, it offers a much larger contingent PCI-e 4.0 lanes per CPU (128) and enables dual-socket setups. The chip reserves 32 of these lanes for inter-socket communication, leaving up to 192 lanes and hence a maximum throughput 378 GB/s for accelerator communication in a fully stacked node. This exceeds the 160 lanes offered by the current leader in this field, AMD, in dual-socket configurations (128 lanes per CPU with at least 48 dedicated to inter-socket configuration). While reliable, independent benchmark data is not yet available, Ampere claims a performance matching that of contemporary AMD systems (AMD Rome EPYC 7742) and outperforming Intel (Cascade Lake SP Xeon Platinum 8280) by a factor of 2. The chip targets applications in data analytics, AI, databases, storage, Edge Computing and cloud-native applications. Ampere plans to continue this line of processors with the next model Mystique following in 2021, using the same socket as Altra, and Siryn following in 2022 at a planned 5nm process [21] [22] [23] [24].

3.2.6. Summary of Arm processors' main technical characteristics (dedicated to HPC)

The Table 2 summarises the Arm HPC processors main technical characteristics.



Architecture	Arm					
Chip maker	Marvell			Fujitsu	EPI	
Processor	ThunderX2	ThunderX3	ThunderX3+	A64FX	Rhea	Chronos
Platform	N1	Zeus	Zeus	N1	Zeus	Poseidon
Core	ARMv8.1	ARMv8.3+	ARMv8.3+	Armv8.2	ARMv8.3/8.4	x
Manufacturer/Foundry	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
Manufacturing Process (nm)	16	7	7	7	7/6+	5
Status	Launched	Cancelled	Cancelled	Launched	Planned	Planned
GA or Estimated Availability	May 2018	None	None	Q4 2019	Estimated 2022	Estimated 2023-2024
Technology	Single-die	Single-die	Dual-Die	CMG	Chiplet	Chiplet
Intra-node interconnect	CCPI	CCPI	CCPI	NOC	CCIX	CCIX/CXL
Extra-node interconnect	PCI-e gen3	PCI-e gen4	PCI-e gen4	PCI-e gen3	PCI-e gen5	PCI-e genx
SMT	4	4	4	4	N/A	N/A
ISA	NEON	NEON	NEON	SVE-512	SVE-256	N/A
Operations	2xFMA @128b	4xFMA @128b	4xFMA @128b	2xFMA @512b	2xFMA @256b	N/A
Cores	Max 32	Max 60	Max 96	Max 48	72	N/A
channels/skt	8	8	8	NA	4 - 6 DDR5	N/A
DDR @ Memory Clock Speed	DDR4 @2667	DDR4 @3200	DDR4 @3200	NA	DDR5 @Min 4800	N/A
Theroritical Bandwidth (GB/s)	171	205	205	NA	230	N/A
HBM @Memory BW (TB/s)	No	No	No	32GB (4 x 8GB) @ 1 TB/s	Maybe	Maybe
Estimated Theoritical Gflops/Watt (Top bin)	3.2	9.2	N/A	>10	N/A	N/A

Table 2: Main technical characteristics of Arm processors dedicated to HPC

Note: N/A means the information is not available.

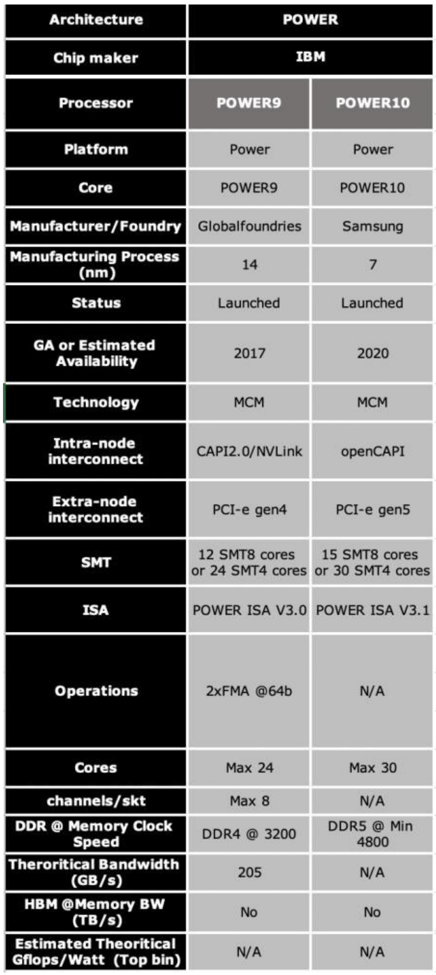
3.3. POWER Processors

Despite being a contender for the highest slots in TOP500 in recent years, the amount of publicly available information about the current plans for the POWER architecture is scarce at the moment. The June 2020 TOP500 has ranked Summit and Sierra, 2 supercomputers based on POWER9 architecture, at the second and third place. At the third place on that list with 94.64 PFlops, Sierra (introduced in the June 2018 edition of the TOP500) boasts a very similar performance level compared to the fourth contender, Sunway TaihuLight (introduced in the June 2016 edition of the TOP500) 93.01 PFlops, despite having almost an order of magnitude less cores (1.5 million vs 10.6 million) and requiring almost half the power (7.4 MW vs 15.4 MW). However, both of these contenders have been surpassed by Fugaku on the most recently published June 2020 TOP500 list.
The bulk of the processing power in these systems comes from the accelerators, namely, NVIDIA GPUs. In fact, most of the marketing material from IBM has focused on the architectural advantages of the POWER system as a whole, instead of focusing on the raw processing power of the POWER CPU alone. In particular, the early adoption of the NVLink communication protocol from NVIDIA has given the architecture a significant advantage over competitors when combined with NVIDIA GPGPUs.

Another area, where IBM had a leading edge over competitors, was the manufacturing process, which did not pan out as expected. In 2015, IBM announced that they were able to produce transistors operational at the 7nm level using silicon-germanium gates, but declined to give a product delivery date at the time. However, in 2018, Globalfoundries announced that they would be ceasing 7nm Extreme-Ultraviolet Lithography (EUV), due to the lack of demand. This led to uncertainty in both AMD's and IBM's products and, since then, AMD has decided to work with TSMC for their Zen2 line. In late 2018, after considering all three major producers (i.e., TSMC, Samsung and, interestingly, their rival Intel), IBM opted to partner with Samsung Foundry for using their 7nm

EUV process. POWER10 is available in 2 versions: 15 SMT8 cores or 30 SMT4 cores per processor while Power 9 was either 24 SMT4 cores or 12 SMT8. It supports PCI-e Gen5, wider sustained memory bandwidth (800+ GB/s as opposed to 650 GB/s in POWER9), double I/O signalling speed (50 GT/s, as opposed to 25 GT/s in POWER9) and a new microarchitecture, in addition to the 7nm manufacturing process (down from 14nm in POWER9). As for POWER11, even fewer details are available to the public, but William Starke, IBM's POWER10 architect has reiterated their preference for the chiplet design for the best utilisation of the manufacturing process in future products, in a recent interview [25]. It is to be noted also that, while support for NVLink on-chip was part of POWER8 and POWER9 architecture, it is no more the case on POWER10 with PCI-e Gen5 providing the suitable bandwidth to feed GPUs.
In the meantime, IBM has also released a new iteration of their current, 14nm based POWER9 line, featuring the new Open Memory Interface (OMI) for decoupling the memory interface from the core CPU design, in order to exploit the advances in memory technology without waiting for the release of their next generation architecture.

The Table 3 summarises the POWER processors main technical characteristics.



Architecture	POWER	
Chip maker	IBM	
Processor	POWER9	POWER10
Platform	Power	Power
Core	POWER9	POWER10
Manufacturer/Foundry	Globalfoundries	Samsung
Manufacturing Process (nm)	14	7
Status	Launched	Launched
GA or Estimated Availability	2017	2020
Technology	MCM	MCM
Intra-node interconnect	CAPI2.0/NVLink	openCAPI
Extra-node interconnect	PCI-e gen4	PCI-e gen5
SMT	12 SMT8 cores or 24 SMT4 cores	15 SMT8 cores or 30 SMT4 cores
ISA	POWER ISA V3.0	POWER ISA V3.1
Operations	2xFMA @64b	N/A
Cores	Max 24	Max 30
channels/skt	Max 8	N/A
DDR @ Memory Clock Speed	DDR4 @ 3200	DDR5 @ Min 4800
Theroritical Bandwidth (GB/s)	205	N/A
HBM @Memory BW (TB/s)	No	No
Estimated Theoritical Gflops/Watt (Top bin)	N/A	N/A

Table 3: Power processors' main technical characteristics

Note: N/A means the information is not available.

3.4. Other Processor Technologies

China plans to build several Exascale systems using their own manufactured CPUs and GPUs. The first one is NRCPC, a CPU-only machine equipped with ShenWei 26010 (SW26010) processors which is the one used in Sunway TaihuLight (Rank 4 in June 2020 TOP500) [26]. The SW26010 contains 260 cores which produce nearly 3.06 TFlops of 64 bits floating point peak performance per CPU. In that respect, with an expected number of dual-sockets nodes larger than 100,000 in their Exascale system, NRCPC should reach a peak performance over 0.6 EFlops. However, it is most probable that the CPU for the future Tianhe-3 Exascale system will be the next Sunway

CPUs which should deliver a peak performance above 10 TFlops. If this scenario comes into reality, NRCPC can reach an aggregate peak performance above 2 EFlops.
The second system in China is the Shuguang Exascale machine relying on two Hygon x86 CPUs and two DCUs (26). While Hygon's CPU is licensed from AMD's first-generation Zen architecture, DCU is a domestic accelerator produced by Hygon delivering 15 TFlops.

4. GPU, Accelerator and FPGA

While NVIDIA has led the GPU market for the HPC world over the last 10 years, new players like AMD and Intel are entering the game. However, while AMD is still at an early stage to deliver their MI GPUs to the HPC market to support both HPC and AI workloads, Intel is working on a 2021 timeframe to launch the Intel Xe Ponte Vecchio GPU. Overall, it is evident that efforts to include more accelerator performance into HPC nodes at large scale continue to be intensified with specialised units for AI covering not only training but also inference, ray tracing and other use cases to be included in newer generations, enabling their use for new applications and supporting convergence of all workloads. Also, the traditional gap between separate memory spaces and device architectures will decrease thanks to new hardware implementations as well as software solutions (e.g. Intel OneAPI), shielding the user from diverging host and device code.

4.1. GPUs

4.1.1. NVIDIA GPUs

In the past few years, the history of the fastest supercomputers worldwide has shown a steady increase of accelerated systems, the majority being equipped with GPUs. Today, 34% of the 50 fastest systems (TOP500[3], June 2020) are GPU-powered by NVIDIA. NVIDIA has a strong history of GPU accelerators in the context of HPC, with only 2% among those accelerated systems using non-NVIDIA accelerator hardware in 2020. With Piz Daint and Marconi, Europe is a prominent marketplace: Piz Daint (Swiss National Supercomputing Centre) is equipped with 5,704 Tesla NVIDIA P100 nodes providing a theoretical peak performance of 27 PFlops, mainly dedicated to numerical simulation while Marconi (CINECA) is built on top of 980 V100 nodes, each node with 4 GPUs Volta100. With the French national supercomputer Jean Zay built to answer to the French AI plan for humanity [27], the machine is built with up to 28 PFlops (one scalar partition and one hybrid partition with 2696 GPU NVIDIA V100) dedicated to both HPC and AI with the capability to run HPC/AI combined simultaneously for science. Following the Pascal (P100) and Volta (V100) generation, the new generation of NVIDIA GPUs released is the Ampere GPU A100 announced by Jensen Huang, NVIDIA CEO, on 14 May 2020. The Ampere A100 is built on TSMC's 7nm process and is both delivered in an SXM form factor (400W TDP) and as a PCI-e card (250W TDP). While the FP64 performance of A100 compared to V100 only increases from 7.8 TFlops to 9.7 TFlops (+25% performance improvement per chip) and FP32 similarly by the same ratio from 15.7 TFlops to 19.5 TFlops, the most important added value for numerical simulations is the memory bandwidth improvement (+75% compared to V100) with a higher HBM2 capacity (40GB) and a higher number of NVLINK3 links allowing to double the global performance capability of A100 to 600 GB/s theoretically. NVLink3 has a data rate of 50 Gbit/s per signal pair, nearly doubling the 25.78 Gbits/s rate compared to V100. A single A100 NVLink provides 25 GB/s bandwidth in each direction, using only half the number of signal pairs per link compared to V100. The total number of links is increased to 12 in A100, vs. 6 in V100, yielding 600 GB/s total bandwidth vs. 300 GB/s for V100.
The A100 will support 6912 FP32 cores per GPU (vs 5120 on V100) and 432 tensor cores per GPU (vs 640 on V100).
The other big jump is for the AI workloads that can leverage instructions using the BFLOAT16 format with performance improving by 2.5x. Furthermore, there are new instructions that enable the use of tensor cores using INT8/4 and TF32 (TensorFloat-32), FP64 and FP32 data. While Volta 100 was mainly focussing on training, the A100, with the support of multiple high precision floating-point data formats as well as the lower precision formats commonly used for inference will be a unique answer to training and inference.
Another important aspect of the A100 for sustainability is the capability of supporting Multi-Instance GPU (MIG) allowing the A100 Tensor Core GPU to be securely partitioned into as many as seven separate GPU instances for CUDA applications, providing multiple users with separate GPU resources to accelerate their applications. This new feature will help optimise resource utilisation knowing that not all the applications are taking advantage of a single GPU while providing a defined QoS (Quality of Service) and isolation between different clients, such as VMs, containers, and processes. Due to its implementation, it ensures that one client cannot impact the work or scheduling of other clients, in addition to providing enhanced security and allowing GPU utilisation guarantees

for each workload. Effectively, each instance's processors have separate and isolated paths through the entire memory system. The on-chip crossbar ports, L2 cache banks, memory controllers, and DRAM address busses are all assigned uniquely to an individual instance. This ensures that an individual user's workload can run with predictable throughput and latency, with the same L2 cache allocation and DRAM bandwidth, even if other tasks are thrashing their own caches or saturating their DRAM interfaces.

In addition, as A100 supports PCI-e gen4 with SR-IOV (Single Root Input/Output Virtualisation), allowing to share and virtualise a single PCI-e connection for multiple processes and/or virtual machines to support a better QoS for all over services (I/O, etc.) [28].

In addition, NVIDIA has announced a new software stack including new GPU-acceleration capabilities coming to Apache Spark 3.0. The GPU acceleration functionality is based on the open source RAPIDS suite of software libraries, built on CUDA-X AI. The acceleration technology, named the RAPIDS Accelerator for Apache Spark, was collaboratively developed by NVIDIA and Databricks. It will allow developers to take their Spark code and, without modification, run it on GPUs instead of CPUs. This makes for far faster ML model training times, especially if the hardware is based on the new Ampere-generation GPU due to its characteristics.

4.1.2. AMD GPUs

Although less visible in the HPC market, AMD is taking a position in the landscape with its planned CDNA GPU architecture at an efficient 7nm fabrication process [29]. Optimised for ML and HPC, AMD envisions these architectures to pioneer the road to Exascale by specifically focusing on the CPU-GPU interconnect. This general trend also adopted by other vendors is further detailed in Section 5.2. With the recent acquisition of Mellanox by NVIDIA showing continued interest in interconnects, also higher bandwidth connections such as AMDs Infinity Fabric will manifest themselves in future large-scale HPC systems, offering around 100 GB/s of full-duplex bandwidth for fast data movement among CPUs and GPUs. Coupled with AMDs aggressive roadmap for X3D packaging, this is expected to lead to more tightly integrated intra-node components, partially mitigating the current relative cost of moving data as computational power increases and limiting the responsibility of programmer and software stack to provide efficient software. Furthermore, specialised hardware units such as ray tracing units have also been confirmed, showing AMDs ambition to continue to compete with NVIDIA in that regard. AMD's successful development is evident in part also due to recently awarded supercomputer contracts, namely Frontier at a planned 1.5 EFlops (ORNL) and El Capitan at 2 EFlops (LLNL). Both systems are planned with AMD CPUs and GPUs, and will be one of the first benchmarks of closely coupled CPU-GPU technologies. The awarding of these contracts shows the commitment of part of the HPC community to AMDs technologies for the next couple of years, with new generations of devices to be released approximately once per year [30]

The Radeon Instinct MI50 compute card, available now, is and designed to deliver high levels of performance for deep learning, high performance computing (HPC), cloud computing, and rendering systems. The MI50 is designed with deep learning operations (3.3 TFlops FP32; 26.5 TFlops FP16; 53.0 TOPS INT8) and double precision performance (6.6 TFlops FP64) with access to up to 32GB HBM2 (ECC) memory delivering 1 TB/s theoretical memory bandwidth. In addition, the Infinity Fabric Link (AMD technology) can be used to directly connect GPU to GPU with 184 GB/s peer-to-peer GPU communication speeds, GPU/CPU communication being run on PCI-e gen3 and 4 with up to 64 GB/s between CPU and GPU. While AMD is coming back in the GPU world, one of the key points is maturity of the software stack with its ROCm (Radeon Open Compute) open ecosystem. The current ROCm3.0 (2019) is more focused on ML which includes MIOpen libraries supporting frameworks like TensorFlow PyTorch and Caffe 2. On the HPC side, AMD is working on programming models like OpenMP which is still not supported in ROCm3.0 though it should be in the next generation ROCm software stack currently under development to have Frontier running optimally in 2021/2022. Another important feature was the AMD capability of providing developers tools on ROCm to help translate the CUDA code automatically into codes capable of running on AMD GPUs. For this reason, HIP (Heterogeneous-Computing Interface for Portability) was created. It is a C++ Runtime API that allows developers to create portable applications for AMD and NVIDIA GPUs from a single source code, removing the separation into different host code and kernel code languages.

The next generation MI Radeon Instinct should be built on the 7nm+ process, and based on the names MI100 GPU whereas its successor should be the MI200.

4.1.3. Intel GPUs

As announced at SC19 in Denver, Intel plans to release a dedicated GPU sometime in 2021. The new Intel GPU, called Intel Xe HPC PVC [31][32][33], is built on a 7nm manufacturing process. It will be hosted by the Intel Sapphire Rapids CPU which facilitates Xe use through a unified memory architecture between the host and the device through an Xe link which should be based on CXL standards, layered on top of PCI-e Gen5. Intel plans to make the Xe GPUs as adaptable as needed to accommodate as many customers as possible. Hence, there could be several versions of GPU Xe either to accommodate HPC needs (double-precision performance FP64 & run high-performance libraries) or AI needs (equivalent of tensor accelerators for AI; flexible data-parallel vector matrix engine; BFLOAT16). Intel's Xe link should also be chosen to interconnect the Intel Xe HPC GPUs together, similarly like NVLINK does between NVIDIA GPUs.

It will feature an MCM package design based on the Foveros 3D packaging technology. Each MCM GPU will be connected to high-density HBM DRAM packages through EMIB (Embedded Multi-Die Interconnect) joining all chiplets together. The Xe HPC architecture should also include a very large unified cache known as Rambo cache which should connect several Xe HPC GPUs together on the same interposer using Foveros technology. This Rambo cache should offer a sustainable peak FP64 compute performance throughout double-precision workloads by delivering huge memory bandwidth. Similar to the Xeon CPUs, Intel's Xe HPC GPUs will come with ECC memory/cache correction and Xeon-Class RAS.

Intel has mentioned that its Xe HPC GPUs could feature 1000s of EUs, each capable of performing eight operations per clock and therefore sometimes seen as 8 cores. The EUs are connected with a new scalable memory fabric known as XEMF (Xe Memory Fabric) to several high-bandwidth memory channels. 16 EUs are grouped into a subslice within a Gen 12 GPU (the first generation of Xe GPUs), with the subslice being similar to the NVIDIA SM unit inside the GPC or an AMD CU (Compute Unit) within the Shader Engine. Intel currently features 8 EUs per subslice on its Gen 9.5 and Gen 11 GPUs. Each Gen 9.5 and Gen 11 EU also contain 8 ALUs which are expected to remain the same on Gen 12. A 1000 EU chip will hence consist of 8000 cores. However, this is just the base value and the actual core count should be much larger than that.

In terms of vector length, Intel Xe GPUs would feature variable vector width as mentioned below:

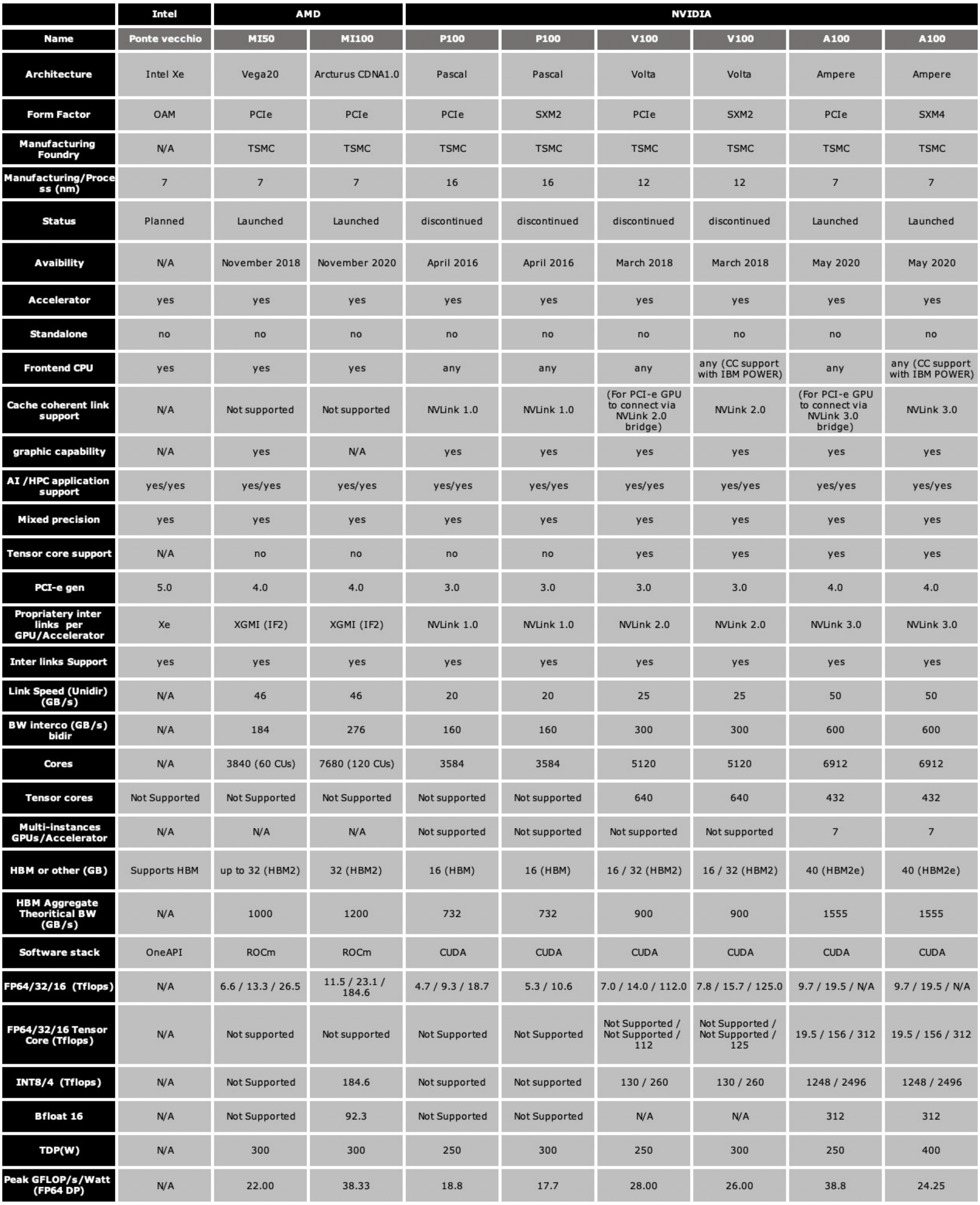
- SIMT (GPU Style)
- SIMD (CPU Style)
- SIMT + SIMD (Max Performance).

This architecture (Intel Sapphire Rapids + Intel Xe HPC Ponte Vecchio GPU) will power the future Aurora Supercomputer which will be launched sometime in 2021 at the Argonne National Laboratory and should be one of the first Exascale machines in the world.

Intel is backing the new hardware development with software support, aiming to provide a stack of hassle-free programming tools and increase their market share by ensuring a large user base. To that end, Intel is focusing their effort on OneAPI [34], an initiative that tries to combine many software projects under one roof in order to facilitate programming CPUs, GPUs, or FPGAs. OneAPI follows the core principle of a single-entry point into the ecosystem, no matter what the underlying hardware base is. OneAPI's [35] main player here is Distributed Parallel C++ (DPC++), which is essentially a mixed language of C++ and SYCL, enhanced with a few Intel flavours, targeting a single-source approach for programming multiple devices. Beyond DPC++, Intel is also working on OneAPI support in OpenMP in both their Fortran and C++ compilers, as any new programming language entails possibly large porting efforts of legacy code bases. Beyond that, Intel also intends to offer debugging and analysis tools with OneAPI, including existing solutions such as vTune, Trace Analyzer, and Intel Advisor, but also third-party tools such as GDB. Finally, Intel intends to offer migration tools that facilitate smooth porting of legacy codes that do require adaption to e.g. new hardware features – a crucial aspect, given that the issue of migration tools is traditionally a difficult one.

4.1.4. Summary of main technical characteristics of GPUs

The Table 4 summarises the main technical characteristics of GPUs.



	Intel	AMD		NVIDIA					
Name	Ponte vecchio	MI50	MI100	P100	P100	V100	V100	A100	A100
Architecture	Intel Xe	Vega20	Arcturus CDNA1.0	Pascal	Pascal	Volta	Volta	Ampere	Ampere
Form Factor	OAM	PCIe	PCIe	PCIe	SXM2	PCIe	SXM2	PCIe	SXM4
Manufacturing Foundry	N/A	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
Manufacturing/Process (nm)	7	7	7	16	16	12	12	7	7
Status	Planned	Launched	Launched	discontinued	discontinued	discontinued	discontinued	Launched	Launched
Avaibility	N/A	November 2018	November 2020	April 2016	April 2016	March 2018	March 2018	May 2020	May 2020
Accelerator	yes	yes	yes	yes	yes	yes	yes	yes	yes
Standalone	no	no	no	no	no	no	no	no	no
Frontend CPU	yes	yes	yes	any	any	any	any (CC support with IBM POWER)	any	any (CC support with IBM POWER)
Cache coherent link support	N/A	Not supported	Not supported	NVLink 1.0	NVLink 1.0	(For PCI-e GPU to connect via NVLink 2.0 bridge)	NVLink 2.0	(For PCI-e GPU to connect via NVLink 3.0 bridge)	NVLink 3.0
graphic capability	N/A	yes	N/A	yes	yes	yes	yes	yes	yes
AI /HPC application support	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes
Mixed precision	yes	yes	yes	yes	yes	yes	yes	yes	yes
Tensor core support	N/A	no	no	no	no	yes	yes	yes	yes
PCI-e gen	5.0	4.0	4.0	3.0	3.0	3.0	3.0	4.0	4.0
Propriatery inter links per GPU/Accelerator	Xe	XGMI (IF2)	XGMI (IF2)	NVLink 1.0	NVLink 1.0	NVLink 2.0	NVLink 2.0	NVLink 3.0	NVLink 3.0
Inter links Support	yes	yes	yes	yes	yes	yes	yes	yes	yes
Link Speed (Unidir) (GB/s)	N/A	46	46	20	20	25	25	50	50
BW interco (GB/s) bidir	N/A	184	276	160	160	300	300	600	600
Cores	N/A	3840 (60 CUs)	7680 (120 CUs)	3584	3584	5120	5120	6912	6912
Tensor cores	Not Supported	Not Supported	Not Supported	Not supported	Not supported	640	640	432	432
Multi-instances GPUs/Accelerator	N/A	N/A	N/A	Not supported	Not supported	Not supported	Not supported	7	7
HBM or other (GB)	Supports HBM	up to 32 (HBM2)	32 (HBM2)	16 (HBM)	16 (HBM)	16 / 32 (HBM2)	16 / 32 (HBM2)	40 (HBM2e)	40 (HBM2e)
HBM Aggregate Theoritical BW (GB/s)	N/A	1000	1200	732	732	900	900	1555	1555
Software stack	OneAPI	ROCm	ROCm	CUDA	CUDA	CUDA	CUDA	CUDA	CUDA
FP64/32/16 (Tflops)	N/A	6.6 / 13.3 / 26.5	11.5 / 23.1 / 184.6	4.7 / 9.3 / 18.7	5.3 / 10.6	7.0 / 14.0 / 112.0	7.8 / 15.7 / 125.0	9.7 / 19.5 / N/A	9.7 / 19.5 / N/A
FP64/32/16 Tensor Core (Tflops)	N/A	Not supported	Not supported	Not Supported	Not Supported	Not Supported / Not Supported / 112	Not Supported / Not Supported / 125	19.5 / 156 / 312	19.5 / 156 / 312
INT8/4 (Tflops)	N/A	Not Supported	184.6	Not supported	Not supported	130 / 260	130 / 260	1248 / 2496	1248 / 2496
Bfloat 16	N/A	Not Supported	92.3	Not Supported	Not Supported	N/A	N/A	312	312
TDP(W)	N/A	300	300	250	300	250	300	250	400
Peak GFLOP/s/Watt (FP64 DP)	N/A	22.00	38.33	18.8	17.7	28.00	26.00	38.8	24.25

Table 4: main technical characteristics of GPUs

N/A means Not Available.

4.2. Other Types of Accelerators

4.2.1. *EPI Titan*

The first generation of the EPI accelerator relying on RISC-V is called Titan (gen 1). It might support (but not be limited to) VPU (Vector Processing Unit), STX (Stencil/Tensor Accelerator - BF16) and VRP (Variable Precision accelerator).
EPI plans to design two different accelerators: one is based on RISC-V instruction set architecture and the other one is based on Kalray's intellectual property (IP) core. While the former will be used for HPC and Al, the latter will function in automotive computation.

4.2.2. *Graphcore IPU*

Graphcore is a young UK-based company which has designed a specific chip called IPU (Intelligence Processing Unit) dedicated to intensive ML algorithms. The IPU is a fine-grained parallel processor designed to deliver high performance for a wide range of computationally intensive algorithms in ML. The IPU design goal was to solve problems beyond the capabilities of current acceleration architectures found in most ASICs and GPUs. The first Graphcore product is the Colossus GC2 built upon a 16nm manufacturing process and is illustrated in Figure 1 [36].

One IPU is built with 1216 interconnected IPU-tiles. Each tile has one IPU core tightly coupled with 300 MB In-Processor-Memory (SRAM) local to each die to enable the model and the data to reside on the IPU to improve memory bandwidth and latency. The 1216 tiles are interconnected through an 8 TB/s on-die fabric (the "IPU-Exchange"), which also connects through "IPU-Links" running at 320 GB/s to create a chip-to-chip fabric. Each IPU core is capable of supporting 8 threads so one IPU can execute 7296 executions in parallel.

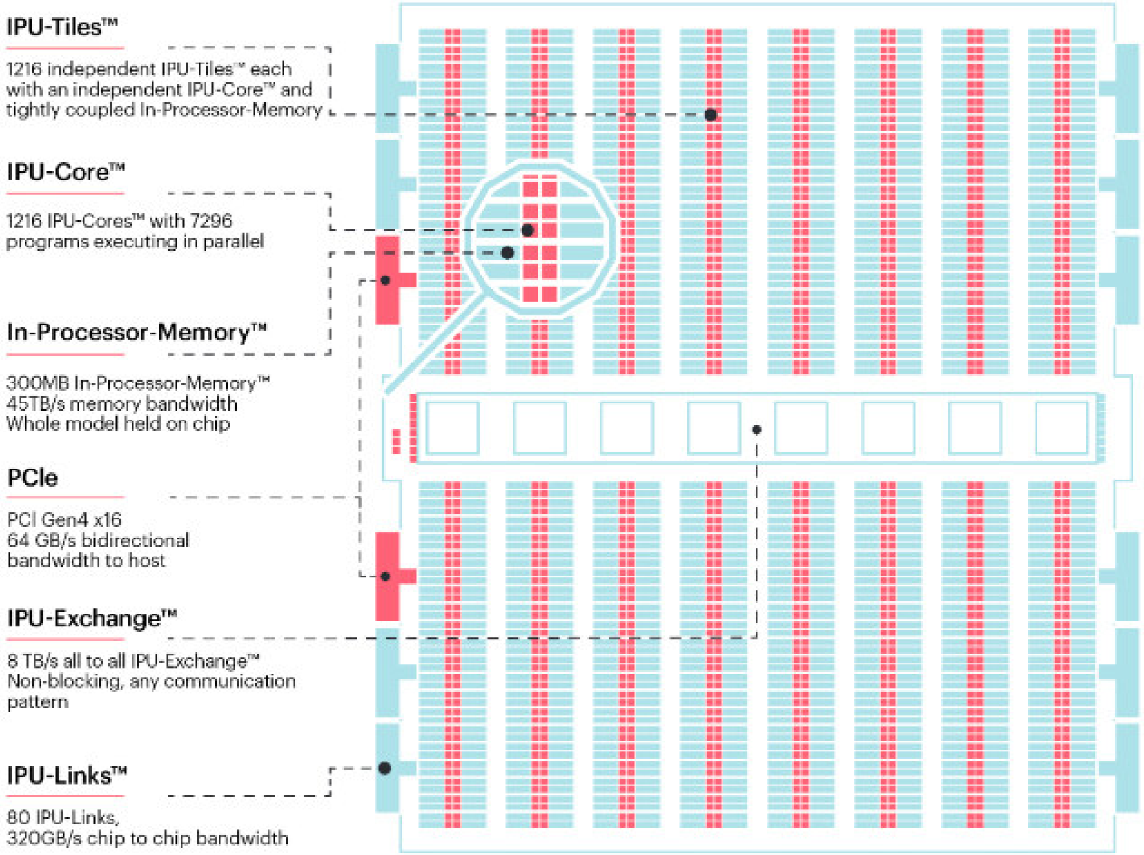

IPU-Tiles™
1216 independent IPU-Tiles™ each with an independent IPU-Core™ and tightly coupled In-Processor-Memory
IPU-Core™
1216 IPU-Cores™ with 7296 programs executing in parallel
In-Processor-Memory™
300MB In-Processor-Memory™
45TB/s memory bandwidth
Whole model held on chip
PCIe
PCI Gen4 x16
64 GB/s bidirectional bandwidth to host
IPU-Exchange™
8 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern
IPU-Links™
80 IPU-Links,
320GB/s chip to chip bandwidth

Figure 1 : Graphcore GC2 Intelligent Processor micro-architecture (https://moorinsightsstrategy.com/wp-content/uploads/2020/05/Graphcore-Software-Stack-Built-To-Scale-By-Moor-Insights-And-Strategy-2.pdf)

The GC2 is delivered in a Gen4 16x PCI-e card form factor, C2 PCI-e card. It embeds two GC2 IPU processors providing 600 MB aggregated In-Processor-Memory, 200 TFlops peak mixed precision FP16.32 IPU compute with a 315W TDP and an IPU link running at 2.5Tbps. The entire system (composed of many IPUS) executes in 2 synchronous phases: computation and communication. Applications that target the IPU are expressed as computational graphs. Computations are performed at the vertices of the graph, and the results are communicated to adjacent vertices according to the edges interconnecting the graph. The communication phase is implemented as a Bulk Synchronous Parallel (BSP) operation, which efficiently transfers data from each tile's on-die SRAM memory to connected tiles' memory. In addition to computation instructions, each IPU core features a dedicated tile-level instruction set for communication phases of the BSP model. The integrated exchange-communication fabric is designed to support BSP for both data and model parallelism — enabled by the graph compiler — potentially scaling to thousands of nodes. An important distinction for the IPU architecture, according to Graphcore, is the ability to efficiently process sparse data and graphs, which improves performance while reducing the total memory requirements. The Graphcore software stack is Poplar. Poplar supports users addressing the main challenges of ML, such as deep neural networks, providing the capability to optimise and efficiently run ML algorithms as research and development of entirely new fine-grained parallel workloads to run on an IPU infrastructure. The current ML frameworks supported by the Graphcore software platform are the most popular ones like TensorFlow, Pytorch, Mxnet, etc. Graphcore has also taken the next step in management software, providing containerisation, orchestration, security, and virtualisation. These strategic choices should ease the adoption as more applications are deployed on the Graphcore platform.

Graphcore has announced their new GC200 processor (Figure 2) in July 2020 built upon a TSMC 7nm FinFET manufacturing process. The GC200 processor features now 1472 independent IPU-tiles (+20% compared to the 1st GO2 generation), each IPU-Tile is built on top of an IPU-core coupled with 900 MB In-Processor-Memory (x3 times compared to the previous generation). The new processor is now capable to execute 8832 programs in parallel. The GC200 is delivered in a Gen4 16x PCI-e card form factor. The PCI-e card is as C200 PCI-e card and embeds two GC200 IPU processors providing 1.8 GB aggregated In-Processor-Memory, 600 TFlops (FP16) peak performance with a 425W TDP with an IPU link running at 2.5Tbps. It can also be provided as an IPU server with an x86 or Arm host CPU and 8 C200 PCI-e cards (16 GC200) in a 2D ring topology with a 256 GB/s card (C200) to card (C200), providing up to 4 PFlops (FP16) peak performance. The last form factor is an IPU-POD providing up to 32 PFlops (FP16) peak performance. The design relies then on the integration of several IPU-Machine (16 to 32), each one being based on 4x GC200 IPU + 1x IPU GATEWAY providing access to the other IPU-Machine through 2x 100Gbps links for inter-communication. The IPU-Machines are all connected through a 2D-Torus topology.

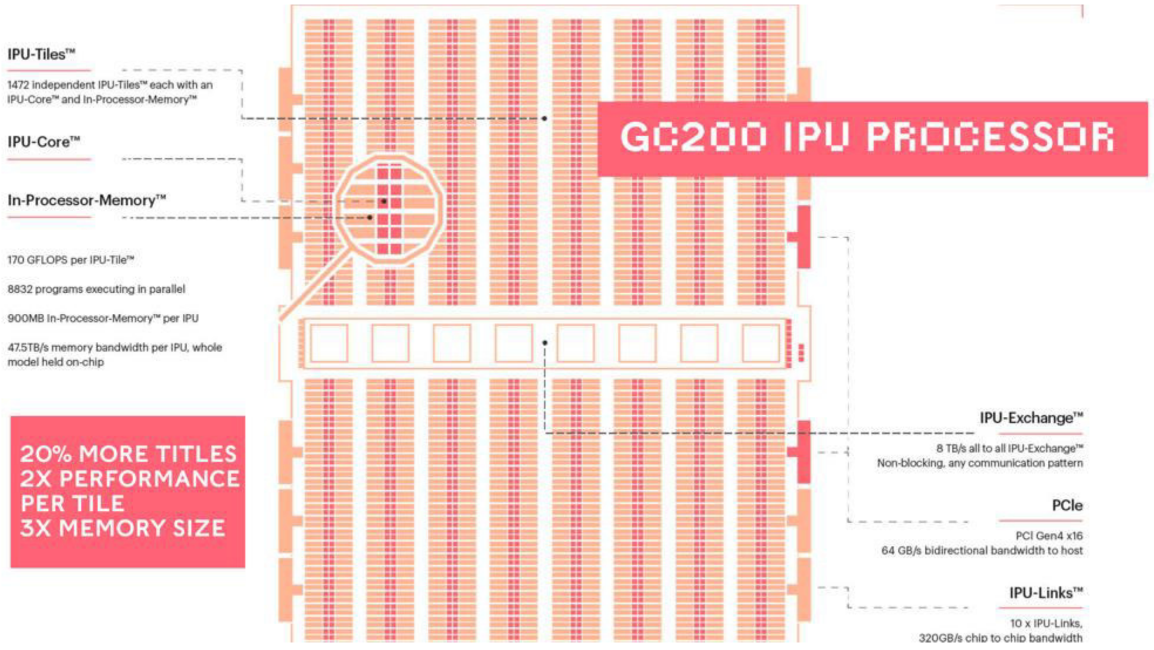
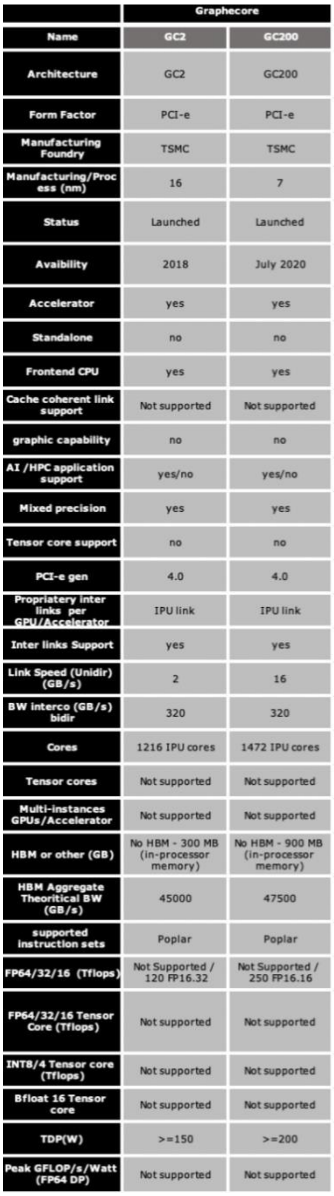

GC200 IPU PROCESSOR
IPU-Tiles™
1472 independent IPU-Tiles™ each with an IPU-Core™ and In-Processor-Memory™
IPU-Core™
In-Processor-Memory™
170 GFLOPS per IPU-Tile™
8832 programs executing in parallel
900MB In-Processor-Memory™ per IPU
47.5TB/s memory bandwidth per IPU, whole model held on-chip
20% MORE TITLES
2X PERFORMANCE
PER TILE
3X MEMORY SIZE
IPU-Exchange™
8 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern
PCIe
PCI Gen4 x16
64 GB/s bidirectional bandwidth to host
IPU-Links™
10 x IPU-Links,
320GB/s chip to chip bandwidth

Figure 2: Graphcore GC200 processor microarchitecture

The Table 5 summarises the Graphcore accelerators' main technical characteristics.



	Graphecore	
Name	GC2	GC200
Architecture	GC2	GC200
Form Factor	PCI-e	PCI-e
Manufacturing Foundry	TSMC	TSMC
Manufacturing/Process (nm)	16	7
Status	Launched	Launched
Avaibility	2018	July 2020
Accelerator	yes	yes
Standalone	no	no
Frontend CPU	yes	yes
Cache coherent link support	Not supported	Not supported
graphic capability	no	no
AI /HPC application support	yes/no	yes/no
Mixed precision	yes	yes
Tensor core support	no	no
PCI-e gen	4.0	4.0
Propriatery inter links per GPU/Accelerator	IPU link	IPU link
Inter links Support	yes	yes
Link Speed (Unidir) (GB/s)	2	16
BW interco (GB/s) bidir	320	320
Cores	1216 IPU cores	1472 IPU cores
Tensor cores	Not supported	Not supported
Multi-instances GPUs/Accelerator	Not supported	Not supported
HBM or other (GB)	No HBM - 300 MB (in-processor memory)	No HBM - 900 MB (in-processor memory)
HBM Aggregate Theoritical BW (GB/s)	45000	47500
supported instruction sets	Poplar	Poplar
FP64/32/16 (Tflops)	Not Supported / 120 FP16.32	Not Supported / 250 FP16.16
FP64/32/16 Tensor Core (Tflops)	Not supported	Not supported
INT8/4 Tensor core (Tflops)	Not supported	Not supported
Bfloat 16 Tensor core	Not supported	Not supported
TDP(W)	>=150	>=200
Peak GFLOP/s/Watt (FP64 DP)	Not supported	Not supported

Table 5: Graphcore Accelerators' technical characteristics

4.2.3. NEC SX Aurora

NEC has invested since the 80s in vector supercomputers and has recently innovated a new hybrid architecture called SX-Aurora TSUBASA. This hybrid architecture consists of a computation part and an OS function part. The heart of the new SX architecture is the vector engine (VE) contained in the vector host (VH) with the VE executing complete applications while the VH mainly provides OS functions for connected VEs.
The SX Aurora vector processor is based upon a 16 nm FinFET process technology and is available as a standard PCI-e card (Figure 3) which can be hosted on any x86 server host environment. Each vector CPU has access to six HBM2 memory modules, leading to a theoretical memory bandwidth of 1.53TB/s.

VE20, the second generation of SX vector processor, has two VE types, Vector Engine Type 20A and Type 20B. VE Type 20A is 3.07TF peak performance with 10 vector cores, and VE Type 20B is 2.45TF peak performance with eight vector cores. Each vector core and 16MB of shared cache are connected by a two-dimensional mesh network providing a bandwidth per vector core of 400GB/s maximum. The vector core on the VE achieves 307 GFlops peak performance per core and an average memory bandwidth of 150 GB/s per core for the 10 cores configuration per the VE processor. Each vector core mainly consists of 32 vector pipelines, and three FMA units are implemented into each vector pipeline. The vector lengths of 256 is processed by 32 vector pipelines with eight clock cycles. The 64 fully functional vector registers per core – with 256 entries of 8 bytes width each – can feed the functional units with data, or receive results from them, thus being able to handle double-precision data at full speed.

Depending on how much the application is vectorised, the VE card can be used in 2 modes. A native mode to run full application on VE card (for vector codes) or an offload mode to run part of the application on VE Card, the remaining on VH X86 scalar system (for scalar & vector codes) with a VE engine supporting standard programming languages and parallelisation paradigms (Fortran, C, C++, MPI, OpenMP, etc.).

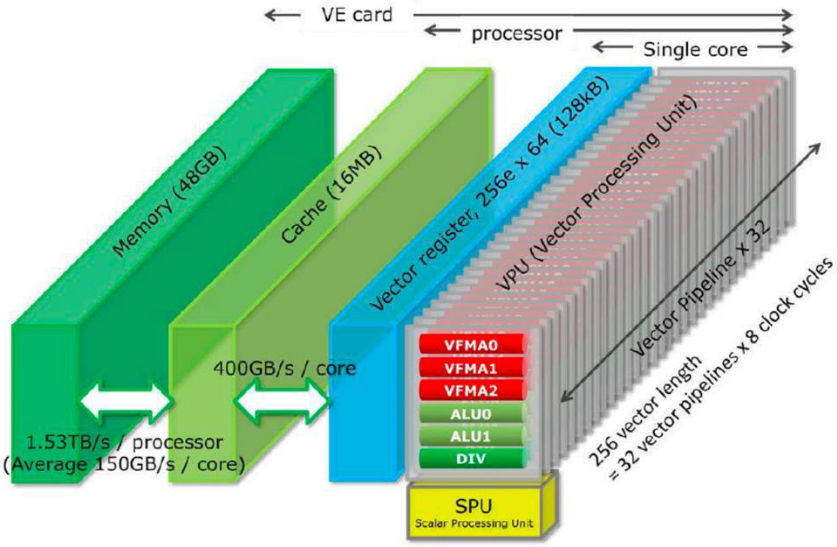


Figure 3 : VE card architecture and workflow

The next generation, named VE30, is planned to be released as the successor of the VE20 generation sometime in 2022. The main improvement from the predecessor is a memory bandwidth of 2+TB/s and a memory subsystem achieving higher sustained performance with lower power consumption.

4.3. FPGAs

Usage of FPGAs in HPC was very limited in the past, mainly due to the adoption barrier of porting applications (e.g. using VHDL – Very High-Speed Integrated Circuits Hardware Description Language) and a lack of support in domain-specific libraries (e.g. FFT, BLAS, LAPACK) or predominant parallelisation models (e.g. OpenMP). In addition, the relatively high performance of GPUs for large classes of floating-point-heavy applications provided a more feasible solution for increasing per-node computational power. For this reason, FPGAs were mainly used in business-centric applications such as high frequency trading and only selectively offered in data centres such as Amazon's Web Service (AWS) F1 instances (which employ Xilinx Virtex UltraScale+ VU9P devices). However, the recent rise in integer and fixed-point math fields such as deep learning, bioinformatics, cryptanalysis or data mining has opened a new market for FPGAs. The predominant company is Intel, having acquired both Omnitek and Altera, with their line of Agilex, Arria, and Stratix devices. With the recent acquisition of Omnitek, Intel plans to continue to offer devices tailored especially towards AI inference and visual applications [37]. Furthermore, Intel's recent advances in the programming software stack (e.g. OneAPI) will likely ensure programmability and sustainability of these devices. The Stratix 10 DX is specifically designed as a cache-coherent accelerator for servers and offers a logic capacity of up to 2.7 million elements along with four Arm Cortex-A53 cores on a 14nm monolithic die. Memory options include HBM2 up to 8 GB, Intel's persistent memory technology Optane up to 4 TB, or a combination of both. The interconnect to host systems will be established through Ultra Path Interconnect (UPI), the successor of QPI, which provides a peak bandwidth of 28 GB/s and enables adoption of future technologies such as CXL and PCI-e 5.0. Coupled with a 100 Gbps Ethernet network interface, these devices can be tailored to many application-specific use cases by end-users able to provide close to 10 TFlops in single precision IEEE 754 math and multiple 100 GBit Ethernet interfaces [38]. The next generation FPGAs is the

Agilex product line, with a number of samples already produced. It is manufactured at 10nm, increasing performance or decreasing power consumption by up to 40%, respectively. Specifically, the I and M series of these devices are intended as cache-coherent accelerators for Xeon processors and focus on high-performance interfaces and bandwidth-intensive use cases (I series) and compute-intensive applications (M series). These FPGAs are designed to deliver up to 40 TFlops of signal processing performance and include support for INT2 through INT8, FP16, and BFLOAT16 data types, improving their applicability to AI workloads. All these features are exposed through Intel's OneAPI initiative, facilitating fast adoption by end-users. Fabrication sizes are currently topping out at 10nm for their newest Agilex product [39], but are expected to shrink to keep up with the increasing energy efficiency of competing technologies.

4.4. OCP Acceleration Module

A new trend to consider in order to accelerate the integration of any kind of accelerator technologies emerging on the market within existing systems is the OAM (OCP Acceleration Module) specification - where OCP stands for Open Compute Project - which defines an open-hardware compute accelerator module form factor and its interconnect. New technologies frequently come with different sizes, different thermal characteristics, a variety of board wiring schemes, and, in some cases, unique sockets. This leads to many form factors that impact the whole system, that in turn need to be accommodated or redesigned for only a few add-ins, thus delaying time-to-market. Since many of these accelerators have similar design requirements and specifications - inter-module communication (to scale up) and high input/output bandwidth (to scale out) - enabling a common form factor specification that can be used in different types of hardware accelerators would definitely help the accelerators' integration. In that respect, the Open Compute Project has been created to reimagine hardware, making it more efficient, flexible, and scalable to support a common design on OAM available to those who want to use it. The OAM form factor could be a standard chosen by integrators and chip makers to ease integration of their new hardware in existing platforms reducing the amount of time needed to push for a new technology onto customer production.

5. Interconnects

Interconnects can be roughly divided into two categories: Inter-Node Connectivity (interconnects used between compute nodes) and Intra-Node Connectivity (interconnects used within a compute node).

5.1. Inter-Node Connectivity

The inter-node networks are one of the key building blocks for HPC systems, a fundamental requirement to connect any number of nodes to form a single, large system. Up to now, for medium and large-scale systems, the inter-node connectivity has usually been split into two physically independent network implementations: on the one hand a low-latency, high throughput network (e.g. Mellanox InfiniBand, Cray Aries, Intel Omni-Path Architecture (OPA), etc.) that is used for user traffic, i.e. MPI communication between nodes and I/O and, on the other hand, a management network (usually Ethernet) to support administrative traffic. This distinction is also made for security reasons. For some small-scale systems, Ethernet is also used for MPI communication. Over the years, the networks used in HPC systems have gradually increased in performance, both in terms of throughput and latency. Most of the network types currently used are switched networks deployed in different topologies in order to optimise the global bandwidth and keep the latencies low.

5.1.1. Ethernet

For a long time, the standard single lane multi-gigabit Ethernet was based on 10 Gbps, with 4 such links being used to create a 40 Gbps link. The corresponding standards were introduced well over 10 years ago and while they first saw slow adoption due to cost, they have nowadays become widespread. Some manufacturers, for instance Intel, have recently integrated 10 Gbps connectivity into chipsets and SoCs with 10 Gbps becoming available even in consumer devices.

5.1.1.1 100/25 Gigabit Ethernet

The 25G Ethernet Consortium was formed in 2014 to develop a single lane 25 Gbps Ethernet standard, approved as IEEE 802.3by in 2016. The 25 Gbps standard is based on one lane from the 100 Gbps 802.3bj approved in 2014, which uses 4 lanes running at 25 Gbps, similar to 10/40 Gbps Ethernet. This provides the ability to use 100 Gbps switches and fan-out cables to get a large number of 25 Gbps ports from a single switch.

5.1.1.2 200 and 400 Gigabit Ethernet

The next step for Ethernet will be 200 and 400 Gbps ports, ratified as a standard at the end of 2017. 200 Gbps Ethernet uses four 50 Gbps lanes per port while initial 400 Gbps standards will use eight 50 Gbps lanes and simply double the number of lanes in the port. Products are starting to become available now: Mellanox for instance has made both 200 Gbps network adapters and 400 Gbps switches available, with the 400 Gbps switches offering ports that can be split into two 200 Gbps ports.

5.1.1.3 RoCE

RoCE is a protocol for enabling RDMA accesses/transfers over regular Ethernet. With direct memory access (DMA) the network adapter can read and write from the host memory directly bypassing the CPU cores, thus lowering the CPU load. RoCE packets also bypass part of the regular ethernet stack and combined with DMA RoCE can have a significantly lower latency than traditional Ethernet. The basic idea for RoCE is to encapsulate an InfiniBand transport packet into a regular Ethernet packet on the link layer. RoCE comes as version 1 and version 2. V1 is a link layer protocol, allowing communication between two hosts in the same broadcast domain. V2 extends the RoCE packet with necessary IP headers to make it effectively a normal UDP (User Datagram Protocol) packet, making it routable.

RoCE capable hardware is available from multiple vendors and bandwidth ratings, from 10 Gbps up to 200 Gbps. While it is unlikely that RoCE will completely replace high performance interconnects such as InfiniBand, it does have some important use cases. For instance, for smaller or more cost optimised systems that need lower latency networks but not the extreme bandwidth, RoCE over 25 Gbps would be a good compromise offering lower latencies than regular Ethernet and reducing the network load on the CPUs of the system.

5.1.2. InfiniBand

InfiniBand is the most widely used HPC interconnect. It is based on a standard that is maintained by the InfiniBand trade association. While there have been multiple vendors manufacturing InfiniBand products, currently the only adapters and switches available are from NVIDIA Networking (formerly Mellanox). InfiniBand focuses on low latency and high bandwidth connections, providing RDMA capability to lower CPU overhead.

5.1.2.1 High Data Rate (HDR/HDR100)

The last generally available InfiniBand products are based on the HDR (High Data Rate) standard. With HDR, a regular 4-lane port theoretically reaches 200 Gbps with a physical port that can be split into two 100 Gbps ports, referred to as HDR100. HDR100 allows HDR to be efficiently used in servers providing only 16x PCI-e gen 3 ports while the 200 Gbps port requires 16x PCI-e gen4 to provide its full capability. Since HDR100 splits a single HDR port into two, it effectively doubles the port count per switch, building fat trees with hosts based on HDR100. One can thus use 200 Gbps HDR to connect and build the network reducing the cabling and switches needed for the network.

The updated version 2 of SHARP (Scalable Hierarchical Aggregation and Reduction Protocol) was introduced with HDR. Before that, SHARP version 1 was introduced in the Mellanox software stack with the previous version of the InfiniBand standard, EDR. SHARP allows some collective operations to be offloaded from the compute nodes to the network by the switches and adapters. SHARP v2 enables offloading to work with large sets of data. HDR also supports hardware-based congestion control and adaptive routing.

5.1.2.2 Next Data Rate (NDR)

The next InfiniBand standard will be NDR, offering a per port bandwidth of 400 Gbps. Availability of NDR products have not been announced yet. As a single 400 Gbps port would require 16x PCI-e gen5 in order not to be limited by the PCI-e bus. It is likely that NDR will be able to split the port into at least two 200 Gbps ports to allow it to be used optimally in systems before PCI-e gen5 processors become available.

5.1.3. Omnipath

In 2015, Intel introduced Omnipath, an HPC interconnect running at 100 Gbps per port. It was available both as standalone PCI-e adapters but also integrated into certain Skylake and Knights Landing CPUs, as an extra chip on the CPU substrate. However, the integrated version was not available on Cascade Lake CPUs and despite a strong roadmap for 200 Gbit Omnipath 2, Intel ceased development of Omnipath in 2019. End of September 2020, Intel has announced that the company spun off Omni-Path Architecture Business to Cornelis Network, an independent company from Intel [40].

5.1.4. Bull eXascale Interconnect (BXI)

The Bull eXascale Interconnect (BXI) is an HPC network designed by Bull, Atos' HPC division, integrated in the Atos BullSequana XH platforms and also available in standard form factors for rack mountable servers. One should notice that, while EPI is the only EU-processor, BXI is the only EU-interconnect network. Like several other interconnects, BXI supports multiple lanes per port, with currently available BXI adapters using four 25 Gbps lanes to provide access to 100 Gbps unidirectional bandwidth. Each BXI switch has forty-eight 100 Gbps ports. The BXI implementation relies on the Portals 4 architecture and has hardware features that map directly to many MPI and PGAS functions. While BXI initially only supported a fat-tree topology, it is now also capable of supporting DragonFly+ topology.

Despite 100 Gbps BXI having been available for quite some time, its deployment was mainly limited to Atos strategic partners. Atos has been selling XH2000 machines with InfiniBand network and not aggressively pushing their own alternative up to now. The largest BXI deployment has been the Tera-1000 machine (CEA - 8000 KNL nodes), a 23 PFlops peak performance supercomputer; in addition to this there are some smaller machines that also use BXI as low latency interconnect network to glue their supercomputing resources.

5.1.5. HPC Ethernet/Slingshot

Now part of HPE, Cray has been developing their own interconnects for a long time. Even after selling the interconnect division to Intel in 2012, Cray did not stop developing their own interconnects. The latest one introduced is branded Slingshot and it is based on Ethernet but uses features that make it more appropriate for HPC workloads, targeting lower latency and better scalability, while at the same time maintaining compatibility with commodity Ethernet. Slingshot is the network that will be used in the 3 US Exascale supercomputers announced recently.

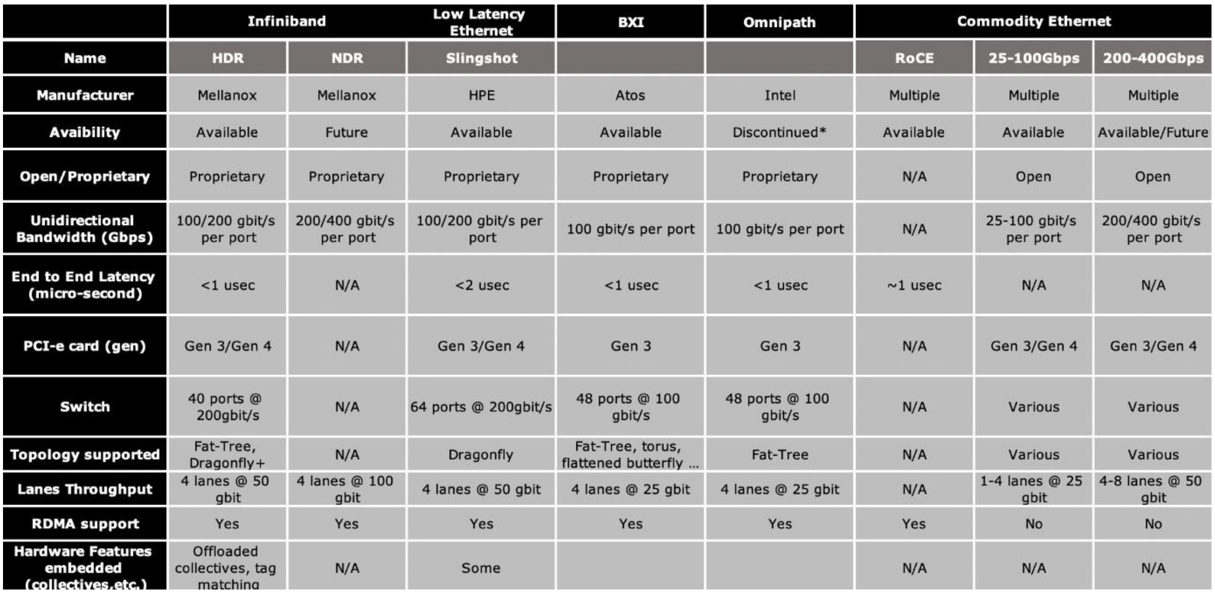
Slingshot switches have 64 ports and each port is running at 200 Gbps. Switches either come integrated into Shasta supercomputers or as regular 1U rack switches. Slingshot can use regular 100 Gbps Ethernet adapters to connect to the switches or dedicated Slingshot network adapters based on HPEs own NIC (Network Interface Card). The dedicated NICs will operate at full 200 Gbps speed with ports capable of some additional capabilities.

While the Shasta machines are built with a Dragonfly network topology, Slingshot does support other topologies such as the more traditional fat-tree as well. For large Dragonfly topologies, the first level switch uses 16 ports for connecting to compute nodes, 31 to do all-to-all connections between all the switches in the same group and then 17 to do the global network.

One of the more advertised features of the Slingshot network is its congestion control mechanisms which is supposed to provide significant improvement compared to the previous Aries network. These congestion control mechanisms should minimise the impact one job can have on other jobs running in the machine at the same time, especially trying to guarantee low latency for latency sensitive applications. This is done by identifying the sources of the congestion and throttling them at the source ports.

5.1.6. Summary of Inter-node interconnect main technical characteristics

The Table 6 summarises the main Inter-node interconnect main technical characteristics.



	Infiniband		Low Latency Ethernet	BXI	Omnipath	Commodity Ethernet		
Name	HDR	NDR	Slingshot			RoCE	25-100Gbps	200-400Gbps
Manufacturer	Mellanox	Mellanox	HPE	Atos	Intel	Multiple	Multiple	Multiple
Avaibility	Available	Future	Available	Available	Discontinued*	Available	Available	Available/Future
Open/Proprietary	Proprietary	Proprietary	Proprietary	Proprietary	Proprietary	N/A	Open	Open
Unidirectional Bandwidth (Gbps)	100/200 gbit/s per port	200/400 gbit/s per port	100/200 gbit/s per port	100 gbit/s per port	100 gbit/s per port	N/A	25-100 gbit/s per port	200/400 gbit/s per port
End to End Latency (micro-second)	<1 usec	N/A	<2 usec	<1 usec	<1 usec	~1 usec	N/A	N/A
PCI-e card (gen)	Gen 3/Gen 4	N/A	Gen 3/Gen 4	Gen 3	Gen 3	N/A	Gen 3/Gen 4	Gen 3/Gen 4
Switch	40 ports @ 200gbit/s	N/A	64 ports @ 200gbit/s	48 ports @ 100 gbit/s	48 ports @ 100 gbit/s	N/A	Various	Various
Topology supported	Fat-Tree, Dragonfly+	N/A	Dragonfly	Fat-Tree, torus, flattened butterfly …	Fat-Tree	N/A	Various	Various
Lanes Throughput	4 lanes @ 50 gbit	4 lanes @ 100 gbit	4 lanes @ 50 gbit	4 lanes @ 25 gbit	4 lanes @ 25 gbit	N/A	1-4 lanes @ 25 gbit	4-8 lanes @ 50 gbit
RDMA support	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Hardware Features embedded (collectives,etc.)	Offloaded collectives, tag matching	N/A	Some			N/A	N/A	N/A

Table 6: Inter-node Interconnect main technical characteristics

N/A means Not Available.

*Omnipath is discontinued by Intel. Business activities are ensured by Cornelis Network.

5.2. Intra-Node Connectivity

The intra-node connectivity is one of the major challenges to improve efficiency of heterogenous architectures (GPP combined to accelerators) since they became more widely adopted. A few years ago, the only open-standard was PCI-e, which does not feature cache coherency. This has led to the development of many new intra-node interconnects.

One could further divide the intra node interconnects between those used to connect different devices within the node and those used to connect different parts of the processor in silicon or on substrate. Modern servers include a large number of different devices that need to communicate with each other in order for the system to work. A typical HPC node may include multiple CPUs, GPUs, network adapters and SSD storage that all need to be connected to each other to make it work. A typical server will include multiple high-speed internal interconnects. For instance, a node with multiple Intel CPUs will use UPI to connect the CPUs together, PCI-e to connect storage, network and GPUs, while providing an additional high bandwidth link to connect all accelerators to each other such as NVIDIA NVLink.

5.2.1. PCI-e gen 3, 4, 5 and 6

PCI express (PCI-e) is the primary internal bus used today to connect devices to CPUs within a node. It is used to connect GPUs, network adapters, NVMe storage and other additional devices. Usually PCI-e starts out from the CPU and devices are directly connected to the CPU through PCI-e; however, there can also be switches introduced along the path allowing multiple devices to share a single connection to the CPU.

PCI-e is usually deployed in multi-lane configurations, 4 lanes (4x) being commonly used for NVMe SSDs and I/Os and 16x being used for more bandwidth intensive applications such as GPUs, and now - with the advent of 100 Gbps or multi 100 Gbps networks - also used for network adapters. CPUs come with a specific number of lanes: Intel server CPUs are using 56 lanes whereas the AMD CPUs come with 128 lanes. Using PCI-e switch chips servers provide more lanes than those that are available from the CPU. However, all devices connected to the same switch will share the same uplink connection to the CPU. While devices connected to the same switch can communicate at whatever speed they are connected, if multiple devices communicate with the CPU, they will share the link going to the CPU and thus may not get the full bandwidth to the CPU that the switch could provide. The PCI-e standard has been rapidly evolving since it has finally moved on from staying at Gen 3 for a long period. Currently the industry is quickly moving to PCI-e Gen 4 with a 16 GT/s (Giga Transfers/s) transfer rates, for a theoretical throughput of 31.5 GB/s for a 16x port with the same encoding as PCI-e Gen 3 (128/130bits). Both

AMD and IBM are currently offering server integrated CPUs with PCI-e Gen 4, and Intel is expected to introduce Gen 4 with the Ice Lake CPUs expected later in 2020.
The next evolutionary step is PCI-e Gen 5, the standard was finalised in 2019. The transfer rate is doubled from Gen4 to 32GT/s, yielding roughly 63 GB/s of bandwidth from a 16x port. CPUs and GPUs supporting gen 5 PCI-e are expected in between 2021 and 2022.
The final version of the PCI-e Gen 6 specification is expected to be released in 2021. It is again expected a doubling of the per lane transfer rate to 64GT/s, yielding roughly 124 GB/s of bandwidth for a 16x port. The new standard will switch to PAM4 to allow 2 bits to be transferred per transfer, and the new standard will also use forward error correction.

5.2.2. CXL

The Compute Express Link (CXL) is an open standard for a high-speed interconnect intended to be used as an interconnect between CPUs and devices and CPUs and memory. The initial consortium was made up of Alibaba, Cisco, Dell EMC, Facebook, Google, HPE, Huawei, Intel, and Microsoft; however, since then the consortium has grown to now include all major HPC hardware vendors, such as AMD, NVIDIA, Mellanox and Arm.
CXL is based on top of PCI-e, specifically CXL v1.0 is based on PCI-e Gen 5 using the same physical and electrical interface. CXL provides protocols in three areas: CXL.cache, CXL.memory and CXL.io. CXL.io is the simplest one, it is very PCI-e like and supports the same features as PCI-e. The more interesting protocols are CXL.memory and CXL.cache, which provide cache and memory semantics. CXL.cache is used for devices that want to cache data from the CPU memory locally, allowing for instance network drives to have their own cache. CXL.memory is used to provide processor access to the memory of attached devices.
The first practical HPC implementation of CXL is expected to be the CPU and GPU combination used in the US Aurora supercomputer that is to be installed in 2021, where it will be used to connect the accelerators to the host CPU system and create a coherent memory space between accelerators and the CPUs. So far, the only product supporting CXL has been some FPGA models from Intel.
With CXL being an open standard and with the participation of all of the large vendors there is hope that CXL could see widespread adoption. However, Intel has been the only major manufacturer that has been making commitments to create products using CXL so far.

5.2.3. Infinity fabric

Infinity Fabric (IF) is AMDs proprietary system interconnect. It is used on multiple levels within AMD systems, to connect different parts of the processor and even multiple GPUs together. The following will focus mostly on the variants not used within the processor.
AMD CPUs come with 128 PCI-e lanes, in dual socket configurations 48 to 64 of them are used to connect the CPUs to each other, and these lanes then switch over to running as IF lanes, providing a coherent shared memory space between the CPUs.
The MI50 and MI60 cards feature an IF connector allowing multiple GPUs to be connected to each other. Up to 4 GPU can be connected in a ring topology, providing up to 184 GB/s of peer to peer GPU bandwidth.

Considering that IF is already used to connect multiple CPUs and between multiple GPUs, the next logical step for AMD would be to extend it to work between CPUs and GPUs. Such a configuration is to be used in the Frontier supercomputer to provide a coherent memory space between accelerators and the CPU.

5.2.4. CAPI/OpenCAPI

The Coherent Accelerator Processor Interface (CAPI) is IBM's proprietary coherent interconnect for directly connecting accelerators to CPUs and forming a coherent shared memory space. CAPI was introduced with the POWER8 processors in 2014. It works over PCI-e, with version 1 using PCI-e gen 3 and version 2, introduced with the POWER9 processors, using PCI-e gen4. One practical implementation of CAPI has been Mellanox network cards in POWER9 systems, such as the ones used on the Summit and Sierra supercomputers.

OpenCAPI is the evolution of the CAPI protocol but instead of being a proprietary standard it is an open standard published by the OpenCAPI consortium. Unlike CAPI, OpenCAPI does not run on top of PCI-e but uses its own protocol. OpenCAPI continues the version numbering from CAPI, meaning the initial version is OpenCAPI 3.0. OpenCAPI 3.0 uses 25 GT/s transfer speeds, with the common deployment being 8 lanes, yielding a bandwidth of 25 GB/s. While OpenCAPI can be seen to use PCI-e slots, these are only used for power feed and as fixtures for the modules. The actual OpenCAPI connector is a slimline SAS cable carrying 8 lanes. Thus far the only support for OpenCAPI seen from a CPU is from IBMs POWER9 series of processors.

One interesting OpenCAPI product that has been shown is a serially attached memory through OpenCAPI. Here the additional system memory can be attached to the CAPI ports instead of the regular DIMM slots. This would alleviate the ever-increasing footprint and pin count needed for the increasing number of memory channels used in servers. This will be an open alternative to IBM's own Centaur memory controllers used until POWER9. The updated version of POWER9 has already supported this new Open Memory Interface, which will also be used for POWER10 [41].

5.2.5. CCIX

Cache coherent interconnect for accelerators (CCIX) is an open standard interconnect, designed to provide a cache coherent interconnect between the CPU and accelerators. CCIX operates on top of PCI-e. The 1.0 specification utilises the standard 16 GT/s transfer rate of PCI-e gen 4, but can also run in an extended speed mode of 25 GT/s. Version 1.1 of the specifications supports PCI-e gen 5 and the 32 GT/s transfer speeds that introduces.

CCIX allows devices to be connected in different flexible topologies: each CCIX device has at least one port that can be used either as a direct connection to another CCIX device or to a CCIX switch. Devices with multiple ports can be used to build more complex topologies, such as daisy changing multiple devices or creating mesh or all-to-all topologies.

With only some HPC vendors participating in the consortium (including Arm, AMD and Mellanox), and some (like Intel and NVIDIA) being absent, and considering also that all major vendors of the CCIX consortium are part of CXL, the future of CCIX is uncertain.

5.2.6. UPI/QPI

UPI (UltraPath Interconnect) is the evolution of QPI (QuickPath Interconnect), Intel's proprietary interconnect used to connect multiple processors in a node together. The previous QPI link was running at 9.6GT/s whereas the new UPI link, introduced with the Skylake architecture is running at 10.4 GT/s. Skylake and newer CPUs have 2 to 3 UPI links per processor. Intel has also used UPI to connect CPUs to FPGAs, creating a cache coherent domain between the CPU and FPGA.

Dual-socket systems, depending on motherboards and CPU types, can have either 2 or 3 links between the CPUs. Quad-socket systems can be built using some CPUs with just 2 UPI links, but in that case each CPU would only be directly connected to 2 other CPUs, and communication with the third one would have to traverse one of the other two. Most Xeon Scalable CPUs that support 4 sockets have the full 3 UPI links so a quad-socket system with just 2 links per CPUs is very uncommon. Since the current maximum number of lanes per CPUs is 3, systems with 8 sockets will not have direct connection between all CPUs in the system.

With Intel embracing CXL, the future of the UPI interconnect is uncertain. Intel will use CXL to connect the GPUs in the upcoming Aurora system to the CPUs, so the question is whether UPI will still be used between the CPUs or it will also be transitioned to CXL. With CXL being based on PCI-e Gen 5, it would potentially offer a higher bandwidth, but possibly a higher latency than solutions based on UPI.

5.2.7. NVLink

NVLink is NVIDIA's proprietary interconnect primarily used for connecting multiple GPUs together. The exceptions are POWER8+ and POWER9 processors that can use NVLink to connect the GPUs directly to the CPUs. NVLink is designed to offer a significantly faster link than what can be achieved with regular PCI-e in order to improve GPU to GPU communication in multi-GPU systems. NVLink allows multiple GPUs into one unified memory space, i.e. allowing a GPU to work on memory local to another GPU with support for atomic operations. With the second generation NVLink introduced with the Volta generation of GPUs, the NVLink connection between GPUs and CPUs was improved to support atomic operations, coherence operations, and allowing data reads from the GPUs memory to be cached in the CPUs cache hierarchy.

NVLink was introduced with the Pascal generation of GPUs, where each GPU had 4 NVLink ports, with 40 GB/s of bandwidth per port. This was expanded to 6 ports running at 50 GB/s for the Volta generation of GPUs. The latest Ampere generation keeps the same 50 GB/s bandwidth per port but increases the number of ports from 6 to 12, yielding a total of 600 GB/s of bandwidth for the GPU.

NVIDIA has also developed a separate NVLink switch chip, with multiple ports that can be used to connect GPUs with each other. The first generation was introduced for the Volta architecture, where the switch used 16 ports, and the NVSwitch was updated for the Ampere architecture to provide double the bandwidth.

GPU-to-GPU connectivity using NVLink can be implemented either as direct connection or using NVSwitch chips. Direct connection with NVLink is primarily used to connect 4 GPUs to each other, and with the Ampere

generation this means 4 links per GPU in an all-to-all configuration yielding 200 GB/s of bandwidth between each pair of GPUs. Initially the direct connection was also used to build 8 GPU systems, now completely replaced by 8 GPU systems featuring NVSwitches. NVSwitches can be used to build a 16 GPU system where each GPU can use its full bandwidth to communicate with any other GPU in the system.
The first system to employ NVSwitches was the Volta DGX-2 system. In these systems, 12 NVSwitches are used and the GPUs are in two planes, each using 6 switches with half the links going to the GPUs in that plane and the other half going to the switches of the other plane. For the Ampere generation, the 8 GPU DGX-1 systems also moved to feature NVSwitches. In this case, 6 switches are used, and the system baseboard is similar to one of the planes of a DGX-2 system.

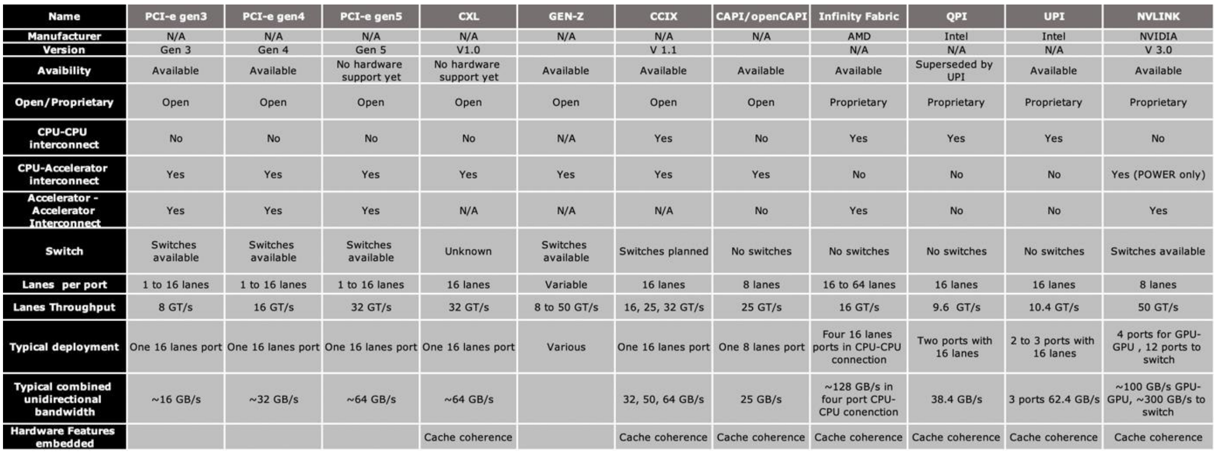
5.2.8. Gen-Z

Gen-Z is an open standard, built on memory semantic operation over a fabric. The goal is to be able to efficiently move data between memories located in different devices with low latency. The fabric supports basic operations such as loads and stores with the idea that all devices in a system will natively support the same operations, replacing both existing buses between GPUs and different devices, and possibly also between the CPU and its own memory. It lacks hardware coherency features, instead relying on software and atomic operations to avoid race conditions.

While not strictly an intra node interconnect, it provides similar functionality as the other intra node interconnects covered here. It has the capability of providing this feature also between hosts. The goal is to provide composable systems, for instance having a pool of memory that can be allocated to multiple different nodes based on their needs.
As of April 2020, the Gen-Z and CXL consortiums announced a Memorandum of Understanding describing how they are intending to cooperate in the future. In essence, CXL will focus on connectivity within the node whereas Gen-Z will focus on rack level fabric connectivity. Large HPC hardware manufacturers which are members in the Gen-Z consortium include NVIDIA, IBM and AMD, along with system integrators such as Dell and HPE. The latter vendors both having shown Gen-Z hardware used to connect multiple systems together, hinting that Gen-Z may evolve into a competitor to Ethernet and InfiniBand.

5.2.9. Summary of Intra-node interconnect main technical characteristics

The Table 7 summarises the main Intra-node interconnect main technical characteristics.



Name	PCI-e gen3	PCI-e gen4	PCI-e gen5	CXL	GEN-Z	CCIX	CAPI/openCAPI	Infinity Fabric	QPI	UPI	NVLINK
Manufacturer	N/A	N/A	N/A	N/A	N/A	N/A	N/A	AMD	Intel	Intel	NVIDIA
Version	Gen 3	Gen 4	Gen 5	V1.0		V 1.1		N/A	N/A	N/A	V 3.0
Avaibility	Available	Available	No hardware support yet	No hardware support yet	Available	Available	Available	Available	Superseded by UPI	Available	Available
Open/Proprietary	Open	Open	Open	Open	Open	Open	Open	Proprietary	Proprietary	Proprietary	Proprietary
CPU-CPU interconnect	No	No	No	No	N/A	Yes	No	Yes	Yes	Yes	No
CPU-Accelerator interconnect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes (POWER only)
Accelerator - Accelerator Interconnect	Yes	Yes	Yes	N/A	N/A	N/A	No	Yes	No	No	Yes
Switch	Switches available	Switches available	Switches available	Unknown	Switches available	Switches planned	No switches	No switches	No switches	No switches	Switches available
Lanes per port	1 to 16 lanes	1 to 16 lanes	1 to 16 lanes	16 lanes	Variable	16 lanes	8 lanes	16 to 64 lanes	16 lanes	16 lanes	8 lanes
Lanes Throughput	8 GT/s	16 GT/s	32 GT/s	32 GT/s	8 to 50 GT/s	16, 25, 32 GT/s	25 GT/s	16 GT/s	9.6 GT/s	10.4 GT/s	50 GT/s
Typical deployment	One 16 lanes port	One 16 lanes port	One 16 lanes port	One 16 lanes port	Various	One 16 lanes port	One 8 lanes port	Four 16 lanes ports in CPU-CPU connection	Two ports with 16 lanes	2 to 3 ports with 16 lanes	4 ports for GPU-GPU , 12 ports to switch
Typical combined unidirectional bandwidth	~16 GB/s	~32 GB/s	~64 GB/s	~64 GB/s		32, 50, 64 GB/s	25 GB/s	~128 GB/s in four port CPU-CPU conenction	38.4 GB/s	3 ports 62.4 GB/s	~100 GB/s GPU-GPU, ~300 GB/s to switch
Hardware Features embedded				Cache coherence		Cache coherence	Cache coherence	Cache coherence	Cache coherence	Cache coherence	Cache coherence

Table 7: Intra-node interconnect main technical characteristics

N/A means Not Available.

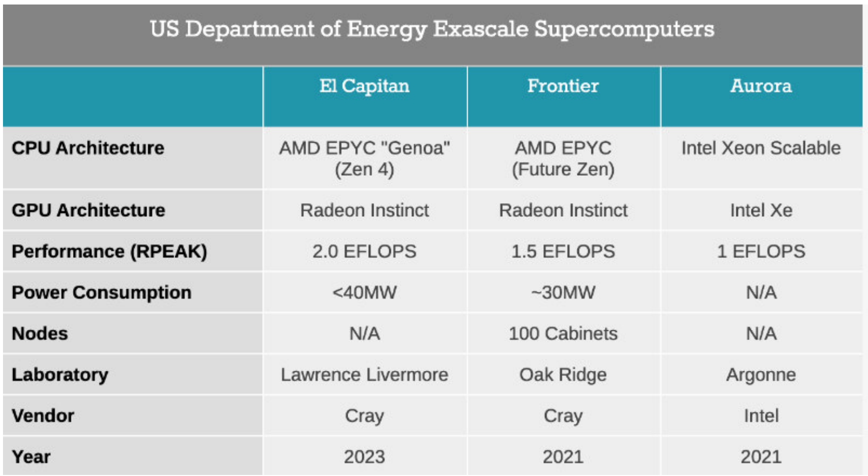
6. Power efficiency

Over the past 40 years, and until recently, the performance of HPC systems has grown following Moore's law. This was possible thanks to progress in semiconductor technology with continuously shrinking manufacturing processes as explained in Section 2 of this document.
While computer capacity has been evolving, the power consumption of supercomputers has also increased over the time leading to energy being one of the most expensive recurrent costs to run a supercomputing facility. As a reminder, the DOE power target was to run an Exascale machine at 20MW, which translates into 50 GFlops/W.

Looking back to 1997, the Sandia National Laboratory held the first TFlop/s computer, taking the No.1 spot on the 9th TOP500 list in June with a power consumption of 850kW resulting in 0.001 GFlops/W (HPL). Eight years later, in 2005, highest performing systems in Top10 were around 100 TFlops/s sustained performance with a power efficiency between 0.01 (CPU based) and 0.2 GFlops/W (Hybrid). In 2008, the first PFlops HPL performance was achieved on Jaguar (ORNL) and RoadRunner (Los Alamos National Laboratory - LANL) with an efficiency of 0.45 GFlops/W for the RoadRunner hybrid machine-based AMD Opteron processor and PowerXcell 8i, 45 times better than general purpose systems in 2005.

Over time, the compute capacity per watt kept increasing to reach around 17 GFlops/W in 2019 for the most energy efficient system, the Fujitsu prototype machine based only on general purpose A64FX processor, an Arm native SVE implementation. The other systems leading the Green500 list are mostly hybrid machines based on general purpose processors and NVIDIA GPUs. In June this year, the Fugaku machine was ranked number one in the Green500 and the GFlops/W was nearly 9% above the record achieved in November 19, reaching 21 GFlops/W.

Looking at the recent DOE announcements (Figure 4) or the US Exascale machines to be installed in 2021/2022 timeframe, the maximum GFlops per watt considering 100% HPL efficiency would be between 35 and 50 GFlops/W.



US Department of Energy Exascale Supercomputers

	El Capitan	Frontier	Aurora
CPU Architecture	AMD EPYC "Genoa" (Zen 4)	AMD EPYC (Future Zen)	Intel Xeon Scalable
GPU Architecture	Radeon Instinct	Radeon Instinct	Intel Xe
Performance (RPEAK)	2.0 EFLOPS	1.5 EFLOPS	1 EFLOPS
Power Consumption	<40MW	~30MW	N/A
Nodes	N/A	100 Cabinets	N/A
Laboratory	Lawrence Livermore	Oak Ridge	Argonne
Vendor	Cray	Cray	Intel
Year	2023	2021	2021

Figure 4: US Department of Energy Exascale Supercomputers main characteristics

Considering a more realistic approach based on a minimum of 68% efficiency, the energy efficiency would be minimum 35 GFlops/W with a target of 50 Gflops/W (sustainable) which could be reached somewhere in 2023 (Figure 5). GFlops/W data from 2015 up to 2020 have been collected from Green500 [42].

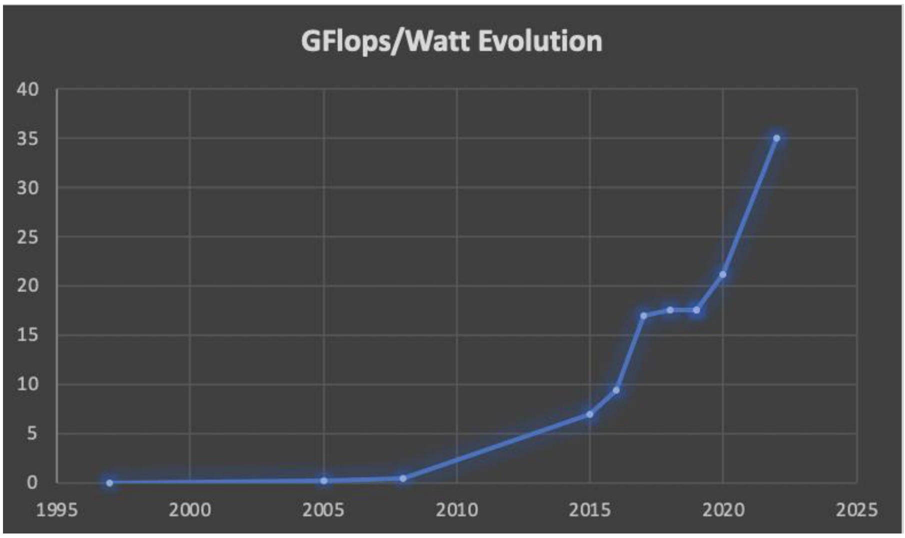


Figure 5: GFlops/Watt evolution based on collected data and Green500

Until end of 2019, there was no debate regarding which type of system could achieve the best performance-per-watt at the horizon of 2021/2022 (some systems are announced/chosen 2 years before they will be deployed): only a system built at least with a major special purpose partition (vs general purpose processor only) would have the capability to be as energy efficient as expected as both capable to support both AI and HPC workloads efficiently. In that respect, the three DOE Exascale supercomputers have already been announced based on hybrid infrastructure (AMD Genoa & ADM MI200 GPU - AMD Milan+ & AMD MI200 GPU – Intel SR & Intel Ponte Vecchio GPU) in the timeframe of 2021/2022.

End of 2019, the A64FX General Purpose Processor has demonstrated the highest Performance-Per watt (17 GFlops/W) on a prototype and mid-2020, A64FX efficiency was confirmed at large-scale on the Fugaku machine positioning Arm-based systems as serious candidates for Exascale, at least to challenge power efficiency.

7. Conclusion: Major Trends

The current section provides an outlook on future trends and summarises mid-terms projections about what users may expect in the coming years; The conclusions are based on all information gathered to build that report.

1. The big CPU market players are still Intel and AMD with their X86_64 processors with a strong competition between the two since AMD increased its market share at the end of 2019. While the X86_64 microarchitecture is still the most adopted in the HPC market, new players are taking the Arm path, while information on IBM POWER is scarce.
2. Market focus is to be part both of HPC and AI markets, with the capability to provide both CPU and GPU and to improve the CPU-GPU interconnect performance as well as the global memory bandwidth, either on pure DDR technology or by using HBM, in addition to DDR or in a DDR-less mode. As an example, upcoming Intel products feature the support of BFLOAT16 for Machine Learning applications and high-performance interfaces to their Xe GPU. The acquisition of Arm by Nvidia is also part of this trend.
3. Arm processors are continuing to expand their market share through Fujitsu (A64FX), Marvell (ThunderX – until their cancelation), Amazon (Graviton), Ampere (Altra) and other chip makers like SiPearl, the company which will design and commercialise the EPI Rhea processor, the first and only European Arm-based processor. While Amazon and Ampere might rather target the cloud sector with an extremely high core count per socket (up to 96), SiPearl (EPI Rhea) and Fujitsu (A64FX) both offer the most relevant features for HPC with less cores per socket but both support HBM and SVE capabilities.
4. It is more and more obvious that most of the high-end computing capabilities would, at least, partially, rely on accelerator technologies, with CPU technologies hosting either GPUs, accelerators (EPI Titan, NEC SX) and/or FPGAs. NVIDIA's new chip, the A100 optimised for AI and AMD GPUs will also implement machine-learning-specific optimisation with their new CDNA line, supporting AI-tailored data types. Intel will release their new Xe "Ponte Vecchio" GPU aimed at accommodating both HPC and AI users through flexible vector engines, supporting both GPU-style and CPU-style vector parallelism with enhanced CPU/GPU cache memory coherency.
5. FPGAs might see their use increased in the future for HPC and AI, mostly through Intel's acquisition of Omnitek and Altera and offer of their Stratix 10 DX product line, among others. AMD has recently also acquired Xilinx, the inventor of FGPA and adaptative SoC, to expand on the datacentre market. Similar to competitor technologies, it focuses on high interconnect and memory bandwidth and offers data types suitable for AI workloads.
6. While the landscape on the low-latency interconnect network is wider compared to a few years ago where most of the systems were InfiniBand-based, there are only a few players capable to power large scale supercomputers (> 5000 nodes): Mellanox (IB), Atos (BXI, the only European Inter-Node Interconnect), HPE Former Cray (Aries) and HPE Cray (Slingshot). Low latency Ethernet is a promising technology that will have to demonstrate its capabilities on the field.
7. Intra-node interconnect is an exciting area to follow in the future as it will allow building a tighter integration between CPU and GPU to ensure cache memory coherency and minimise data movement between CPU and GPU, also allowing to potentially rethink the design of an accelerated node with DDR-less CPU and memory consumed directly from GPU/accelerators' HBM. It will also help to support the adhesion of MCM design to go beyond the current process manufacturing limits and more powerful chips. Openness of intra-node interconnects will be key so it can see a wide adoption and ensure hardware interoperability as ease software programming.
8. Future Exascale system target efficiency of 50 GFlops/W system in order to sustain a high energy efficiency. While the first ½ Exascale class system is based on Arm architecture and has a very good power efficiency (Figure 5) due a balanced design, the future announced Exascale systems at 2021/2022 should reach around 35 sustainable GFlops per watt. 50 GFlops per watt should be achievable in the 2023 timeframe either through heterogeneous architectures based on accelerator computing capabilities combined with Arm processors or through a future well balanced Arm processor design with enhanced capabilities (AI, edge computing, etc.).
9. Further near-terms developments might include merging CPUs and GPUs onto a single die to build APU dedicated to HPC, both improving latency and memory coherency. Longer-term investment could be quantum computing: Lithography approaching the size of silicon atoms might entail incorporating

quantum computing for suitable workloads, giving birth, in a first approach, to another type of hybrid systems based on current computing technologies and quantum accelerators and/or simulator.

References

[1] A. Johansson D. Pleiter, C. Piechurski, K. Wadówka. *Data Management Services and Storage Infrastructures.* 2020. PRACE Technical Report.http://www.prace-ri.eu

[2] E. Krishnasamy S. Varrette, M. Mucciardi. *Edge Computing: An Overview of Framework and Applications.* 2020. PRACE Technical Report.

[3] [Online] https://www.top500.org.

[4] [Online] http://www.mooreslaw.org.

[5] IntelCFO. [Online] https://www.anandtech.com/show/15580/intel-cfo-our-10nm-will-be-less-profitable-than-22nm.

[6] [Online] https://www.nextplatform.com/2020/08/17/the-ticking-and-tocking-of-intels-ice-lake-xeon-sp/.

[7] [Online] https://adoredtv.com/exclusive-intel-sapphire-rapids-to-feature-chiplets-hbm2-400-watt-tdp-coming-in-2021/ .

[8] IntelAMD. [Online] https://www.techradar.com/news/intel-admits-it-wont-catch-up-with-amds-7nm-chips-until-2021.

[9] [Online] https://www.tomshardware.com/news/leaked-amd-epyc-milan-specifications-tease-possible-64-zen-3-cores-at-3-ghz.

[10] [Online] https://www.tomshardware.com/news/amd-zen-3-zen-4-epyc-rome-milan-genoa-architecture-microarchitecture,40561.html .

[11] AMD2022. [Online] https://www.tomshardware.com/news/amd-ddr5-usb-4-next-gen-cpus-2022.

[12] EuroHPC. [Online] https://eurohpc-ju.europa.eu/.

[13] EPI. [Online] https://www.european-processor-initiative.eu/project/epi/.

[14] [Online] https://www.anandtech.com/show/16072/sipearl-lets-rhea-design-leak-72x-zeus-cores-4x-hbm2e-46-ddr5.

[15] [Online] https://www.anandtech.com/show/15621/marvell-announces-thunderx3-96-cores-384-thread-3rd-gen-arm-server-processor .

[16] [Online] https://www.nextplatform.com/2020/08/18/taking-a-deeper-dive-into-marvells-triton-thunderx3/.

[17] HPCWIRE. [Online] https://www.hpcwire.com/2020/02/03/fujitsu-arm64fx-supercomputer-to-be-deployed-at-nagoya-university/.

[18] ANANDTECH. [Online] https://www.anandtech.com/show/15169/a-success-on-arm-for-hpc-we-found-a-fujitsu-a64fx-wafer.

[19] NEXTPLATFORM. [Online] https://www.nextplatform.com/2019/09/24/europeans-push-fpga-powered-exascale-prototype-out-the-door/.

[20] ANANDTECH-2. [Online] https://www.anandtech.com/show/15578/cloud-clash-amazon-graviton2-arm-against-intel-and-amd.

[21] EXTRTECH. [Online] https://www.extremetech.com/extreme/306951-ampere-altra-arm-cpus-launch-with-up-to-80-cores-to-challenge-xeon-epyc.

[22] AMPERECOMP. [Online] https://amperecomputing.com/ampere-altra-industrys-first-80-core-server-processor-unveiled/.

[23] VENTUREBEAT. [Online] https://venturebeat.com/2020/03/03/ampere-altra-is-the-first-80-core-arm-based-server-processor.

[24] ANANDAMP. [Online] https://www.anandtech.com/show/15575/amperes-altra-80-core-n1-soc-for-hyperscalers-against-rome-and-xeon.

[25] https://www.nextplatform.com/2019/08/06/talking-high-bandwidth-with-ibms-power10-architect/. [Online]

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

[27] [Online] https://www.aiforhumanity.fr/en/.

[28] NVIDIAAMP. [Online] https://devblogs.nvidia.com/nvidia-ampere-architecture-in-depth/.

[29] AMD. [Online] https://www.anandtech.com/show/15593/amd-unveils-cdna-gpu-architecture-a-dedicated-gpu-architecture-for-data-centers.

[30] AMDEpyc. [Online] https://www.servethehome.com/amd-cdna-gpu-compute-architecture-5nm-epyc/.

[31] [Online] https://www.hpcwire.com/2020/07/30/intels-7nm-slip-leaves-questions-about-ponte-vecchio-gpu-aurora-supercomputer/.

[32] [Online] https://www.anandtech.com/show/15119/intels-xe-for-hpc-ponte-vecchio-with-chiplets-emib-and-foveros-on-7nm-coming-2021.

[33] [Online] https://www.anandtech.com/show/15188/analyzing-intels-discrete-xe-hpc-graphics-disclosure-ponte-vecchio/2.

[34] oneAPI. [Online] https://software.intel.com/content/www/us/en/develop/tools/oneapi.html.
[35] oneAPI-2. [Online] https://www.anandtech.com/show/15188/analyzing-intels-discrete-xe-hpc-graphics-disclosure-ponte-vecchio/4.
[36] [Online] : https://moorinsightsstrategy.com/wp-content/uploads/2020/05/Graphcore-Software-Stack-Built-To-Scale-By-Moor-Insights-And-Strategy-2.pdf.
[37] FPGA. [Online] https://techcrunch.com/2019/04/16/intel-acquires-uks-omnitek-to-double-down-on-fpga-solutions-for-video-and-ai-applications/.
[38] Stratix. [Online] https://www.intel.com/content/www/us/en/products/programmable/fpga/stratix-10.html.
[39] Agilex. [Online] https://www.anandtech.com/show/14149/intel-agilex-10nm-fpgas-with-pcie-50-ddr5-and-cxl.
[40] Online] https://www.hpcwire.com/off-the-wire/intel-omni-path-business-spun-out-as-cornelis-networks/.
[41] *IBM's POWER10 processor.* W.Starke B. Thompto. brak miejsca : Hot Chips 32, 2020.
[42] [Online] https://www.top500.org/lists/green500/.

List of acronyms

2D	2-Dimension
AI	Artificial Intelligence
AI	Artificial Intelligence
AVX	Advanced Vector Extensions
AWS	Amazon Web Service
BFLOAT	Brain floating-point format
BLAS	Basic Linear Algebra Subprograms
BSP	Bulk-synchronous parallel
BXI	Bull eXascale Interconnect
CAPI	Coherent Accelerator Processor Interface
CCD	Compute Core Die
CCD	Compute Core Die
CCIX	Cache Coherent Interconnect for Accelerators
CCPI	Cavium Cache Coherent Interconnect
CDNA	GPU architecture for data centre compute
CEO	Chief Executive Officer
CFO	Chief Financial Officer
CISC	Complex Instruction Set Computer
CMG	Core Memory Group
CP	The EPI Common Platform
CPU	Central Processing Unit
CU	Compute Unit
CXL	Compute Express Link
DDR	Double Data Rate
DIMM	Dual in-line memory module
DMA	Direct Memory Access
DoE	Department of Energy
DP	Double Precision
DPC	DIMM Per Channel
DPC++	Distributed Parallel C++
DRAM	Dynamic Random Access Memory
EFLOPS	Exa Flops
eFPGA	embedded FPGA
EPI	European Processor Initiative
EU	European Union
EUs	Execution Units
EUV	Extreme-Ultraviolet Lithography
FFT	Fast Fourier Transform
FMA	Fused Multiply Add
FP16	Floating Points 16 bits
FP32	Floating Points 32 bits
FP64	Floating Points 64 bits
FPGA	Field Programmable Array
GB	Gigabyte
GB/s	Gigabyte per second
Gbits/s	Gibabits per second
Gbps	Gigabit per second
GCC	GNU Compiler Collection
GFS	Global File System
GHz	GigaHertz
GNA	Gaussian Neural Accelerator
GPP	General Purpose Processor
GPU	Graphics Processing Unit
GT/s	Giga Transfers per second
HBM	High Bandwidth Memory
HDR	High Data Rate
HIP	Heterogeneous-Computing Interface for Portability
HPC	High Performance Computing
HSL	High Speed Links
HSM	Hardware Security Modules

IDS	Intrusion Detection System
IF	Infinity Fabric
iGPU	Integrated Graphics Processing Unit
INT16	Integer 16 bits
INT8	Integer 8 bits
IO	I/O Input/Output
IP	Intellectual Property
IP	Internet Protocol
IPU	Intelligence Processing Unit
ISA	Instruction Set Architecture
KB	Kilobytes
KiB	Kibibyte
kW	Kilowatt
L	level
LANL	Los Alamos National Laboratory
LAPACK	Linear Algebra Package
LGA	Land Grid Array
LLNL	Lawrence Livermore National Laboratory
MB	Megabytes
MCM	Multi-Chip Module
MiB	Mebibyte
MIG	Multi-Instance GPU
ML	Machine Learning
MPI	Message Passing Interface
MPPA	Multi-Purpose Processing Array
MT/s	Mega Transfers per second
MW	Megawatt
N/A	Not Available
NDR	Next Data Rate
NIC	Network Interface Card
nm	Nanometre
NoC	Network-on-Chip
NUMA	Non-uniform Memory Architecture
NVDIMM	Non-volatile dual in-line memory module
NVLink	NVIDIA Link
NVMe	Non-Volatile Memory Express
NVSwitches	NVLink switches
OAM	OCP Acceleration Module
OCP	Open Compute Project
OMI	Open Memory Interface
OPA	Omni-Path Architecture
ORNL	Oak Ridge National Laboratory
OS	Operating System
PAM-4	Pulse Amplitude Modulation with 4 levels
PB	Petabyte
Pbps	Petabytes per second
PCI-e	Peripheral Component Interconnect Express
PFLOPS	Peta FLOPS
PUE	Power Usage Effectiveness
PVC	Ponte Vecchio
Q	Quarter
QoS	Quality of Service
QPI	QuickPath Interconnect
RDMA	Remote Direct Memory Access
RISC	Reduced Instruction Set Computer
RoCE	RDMA over Converged Ethernet
ROCm	Radeon Open Compute
SAS	Serial Attached SCSI
SATA	Serial ATA
SEV-ES	Secure Encrypted Virtualisation-Encrypted State
SHARP	Scalable Hierarchical Aggregation and Reduction Protocol
SIMD	Single Instruction Multiply Data

SKU	Stock Keeping Unit
SMT	Simultaneous Multithreading
SoC	System on Chip
SP	Scalable Platform
SPE	Synergistic Processing Element
SR-IOV	Single Root Input/Output Virtualisation
SSD	Solid State Drive
STX	Stencil/Tensor Accelerator
SVE	Scalable Vector Extension
TB	Terabyte
Tbps	Terabits per second
TDP	Thermal Design Power
TF32	TensorFloat-32
TFLOPS	Tera FLOPS
TSMC	Taiwan Semiconductor Manufacturing Company
UDP	User Datagram Protocol
UPI	Ultra Path Interconnect
VE	Vector Engine
VH	Vector Host
VHDL	Very High Speed Integrated Circuits Hardware Description Language
VM	Virtual Machine
VRP	Variable Precision accelerator
W	Watt
XEMF	XE Memory Fabric

Acknowledgements

This work was financially supported by the PRACE project funded in part by the EU's Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 823767.

Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

Quantum Computing – A European Perspective

Mikael P. Johansson [a*1], Ezhilmathi Krishnasamy [b*2], Norbert Meyer [c*3], Christelle Piechurski [d*4]

[a]CSC – IT Center for Science, Finland, [b]University of Luxembourg, Luxembourg, [c]Poznań Supercomputing and Networking Center, Poland, [d]GENCI, France

Abstract

Quantum computers have the potential to bring forth a major breakthrough in scientific computing. The foreseen increase in computational efficiency offered by quantum computing is of such magnitude that, despite being in its infancy, it is already being coupled with traditional high-performance computing technology. Here, we give an overview of quantum computing, the present state of affairs, and future scenarios. Europe has a unique opportunity to create world-leading supercomputing infrastructures incorporating quantum technology, by capitalising on the established expertise of European HPC centres in conjunction with the emerging European quantum ecosystem. This requires dedicated and sustained funding for quantum hardware and software developments, as well as for education. In addition, coordinated efforts and support for early adoption of quantum computing in academia and industry are essential.

[1] mikael.johansson@csc.fi

[2] ezhilmathi.krishnasamy@uni.lu

[3] meyer@man.poznan.pl

[4] christelle.piechurski@genci.fr

Table of Contents

This technical report is part of a series of reports published in the Work Package "HPC Planning and Commissioning" (WP5) of the PRACE-6IP project. The series aims to describe the state-of-the-art and mid-term trends of the technology and market landscape in the context of HPC and AI, edge-, cloud- and interactive computing, Big Data and other related technologies. It provides information and guidance useful for decision makers at different levels: PRACE aisbl, PRACE members, EuroHPC sites and the EuroHPC advisory groups "Infrastructure Advisory Group" (INFRAG) and "Research & Innovation Advisory Group" (RIAG) and other European HPC sites. Further reports published so far on the PRACE webpage [1] cover "State-of-the-Art and Trends for Computing and Network Solutions for HPC and AI", "Data Management Services and Storage", "Edge Computing: An Overview of Framework and Applications", "User Requirements influencing HPC Technologies" and "Security in an Evolving European HPC Ecosystem".

[1] https://prace-ri.eu/infrastructure-support/market-and-technology-watch

1. Introduction

Quantum computing (QC) is expected to bring a new revolutionary component to the high-performance computing palette. By directly exploiting quantum mechanical phenomena to an advantage, quantum computers may solve certain computational problems more efficiently than present day supercomputers and HPC algorithms. When sufficiently mature, quantum computers could tackle problems that due to their size and complexity will forever stay beyond the reach of conventional computing alone. Similar to the advent of transistor technology in 1947, the boost in computing power provided by this new computing resource is expected to dramatically increase the impact of research and accelerate problem solution, with a very promising effect on energy consumption.

Quantum computing is expected to have an impact on practically all fields of science, research, development and innovation that utilise, or *could* utilise computational modelling. Fields include artificial intelligence and machine learning, materials science and chemistry, pharmaceutical and medical research, finance and climate modelling, *etc*. This ground-breaking technology has the potential to provide solutions to some of the most pressing challenges of our society, from accurate modelling of complex weather systems and optimisation of resource usage, to the development of novel, sustainable materials as well as more efficient and personalised drugs.

The idea of quantum computers is already forty years old. Springing from the realisation that certain types of problems, for example simulation of physical processes at atomic scale, are by their very nature extremely difficult to model on classical computers, a new computing paradigm was born. After a long period of steady but arduous growth, the advances in quantum technology are now rapid, with the pace still accelerating. Presently, quantum computing is at a stage where its power has been demonstrated by performing actual calculations that are out of reach for classical computers [Arute 2019][Zhong 2020] [Wu 2021] . These *quantum supremacy* experiments serve as proof that there are no fundamental, physical limitations that would prohibit a quantum speed-up, although the actual problems that were solved are of little practical use. There is, however, still work to be done and challenges to overcome in order for quantum computers to show *quantum advantage*, and become integral components of workflows for solving real-world problems.

To harness the full potential of the upcoming quantum revolution, constructing the hardware alone is not sufficient. In order to utilise the hardware, tailor-made algorithms and software needs to be developed. Quantum programming requires fundamental rethinking on several levels. For example, quantum physical phenomena that are absent in classical computing, like superposition and entanglement have to be exploited. Problems have to be formulated properly, and in a novel manner, in order to be amenable to computation on quantum hardware.

For boosting and catalysing quantum computing and quantum software development, mature quantum computing infrastructures are crucial. The platforms have to provide a suitable level of abstraction, so that also users without deep expertise of quantum technology can utilise the new resources. Quantum experts will develop the required low-level software libraries and tools, while experts in other domains would use these tools for solving their respective research questions. In essence, the end-user should be given the most suitable tools possible for performing the actual research he or she is an expert in.

In order to increase quantum-literacy throughout Europe, the educational aspect of quantum computing requires attention. A prerequisite for this is that platforms that provide low-barrier adoption of the technology are made available. Then, students at various levels, including professionals in fields that could either utilise or further develop quantum computing can be reached.

In this report, quantum computing is introduced by defining the key concepts in order to familiarise the reader with the technology. Next, we focus on how to provide prompt access to quantum computing to the scientific community by coupling existing supercomputers to quantum systems while highlighting the algorithms that users can rely on to address different use cases. Existing and future synergies between industrials and academics follow. Europe should consider as key to support quantum adoption, as QC can be a decisive vector of green transition and digitalisation in a period where sustainability is at the heart of many discussions. The last section will be dedicated to the quantum market and future directions to keep an eye on. The conclusion focuses on what needs to be done to pursue the European efforts, together with individual European nations.

2. Bits and Qubits

A classical, digital computer uses bits to store and process information. A bit can be either 0 or 1, and can for example be represented by the absence or presence of an electrical signal, encoding “0” or “1”, respectively. When a computer performs an operation, the values either stay the same, or change: 0 becomes 1, or 1 becomes 0.

A quantum computer exploits the laws of quantum mechanics to enhance the capability of classical bits. Quantum bits, qubits, can, like classical bits, represent the states “0” and “1”. In addition, qubits can be “0” and “1” at the same time. This is known as a quantum mechanical superposition of states. To emphasise the quantum nature of qubits, the Dirac bra-ket notation is used, with states “0” and “1” represented by $|0\rangle$ and $|1\rangle$, respectively.

In general, the quantum state $|\psi\rangle$ of a qubit is a combination of the basis states $|0\rangle$ and $|1\rangle$, defined by the coefficients α and β: $|\psi\rangle = \alpha|0\rangle+\beta|1\rangle$. The coefficients, or amplitudes, are complex numbers, not simply real numbers between 0 and 1. As the square of the amplitude of a given state corresponds to the probability of that state (the Born rule), we get a constraint on the values of α and β. The sum $|\alpha|^2 + |\beta|^2$ must equal one, that is, 100%. The state of a qubit can therefore be represented as a point on the surface of a sphere, conventionally called the Bloch sphere. Each point on the sphere corresponds to specific amplitudes of $|0\rangle$ and $|1\rangle$, that is, specific superpositions of $|0\rangle$ and $|1\rangle$, see Figure 1.

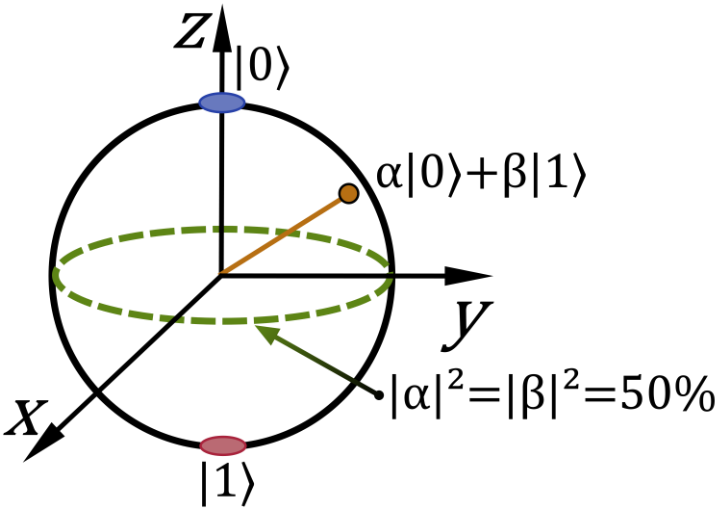


Figure 1: The Bloch Sphere. The blue dot along the z-axis represents the “north pole”, state $|0\rangle$, the red dot represents the “south pole”, state $|1\rangle$, and the dotted green line represents the “equator”, where states $|0\rangle$ and $|1\rangle$ are in equal superposition. The orange dot represents a general state of the qubit.

The Bloch sphere encapsulates the superior information content and operational flexibility of qubits with respect to classical bits. A classical bit can only take two values on the Bloch sphere, $|0\rangle$ or $|1\rangle$, and the only modification possible is to go from the north pole to the south pole, or vice versa. A qubit, on the other hand has access to the infinite set of points on the surface of the sphere; any combination of longitude and latitude, that is, a superposition of any amount of $|0\rangle$ and $|1\rangle$.

Just like ordinary bits, a qubit always returns the value “0” or “1” when read out, that is, measured, even if it would be in a superposition of both values. The result is probabilistic, with the probabilities dictated by the amplitudes. For example, a qubit at the equator of the Bloch sphere has a 50/50 chance of returning either 0 or 1 when measured; it will not return, say, 0.5. In addition, measurement destroys the superposition (collapses the wavefunction in the Copenhagen interpretation): the state of the qubit is fixed to either the north or south pole of the Bloch sphere, and all information about the amplitudes is lost. This means that unlike bits, the value of qubits cannot be read mid-calculation.

In addition to encoding any value between 0 and 1 for the qubit, the complex amplitudes make it possible to describe the phase of the wavefunction. This adds another powerful feature to qubits over ordinary bits: the possibility of constructive and destructive interference. Let us consider the equator of the Bloch sphere, specifically, two points on it, the “plus” state and the “minus” state:

$$|+\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle; \;\; |-\rangle = \frac{1}{\sqrt{2}}|0\rangle - \frac{1}{\sqrt{2}}|1\rangle \tag{1}$$

When measuring, both qubit states will return “0” or “1” with 50% probability. They are, however, located on opposite sides of the Bloch sphere, along the x-axis, and have opposite phases. Any operation (except measurement) will affect the two states differently. For example, a rotation around the y-axis will shift the probability of measuring “0” or “1” in an opposite manner for $|+\rangle$ and $|-\rangle$.

Qubits provide a third fundamental advantage over bits, in addition to superposition and interference: entanglement. When a pair of qubits are entangled, their states are connected so that for example measuring the state of one qubit immediately fixes the state of the second qubit. Consider the two-qubit Bell state:

$$|\psi\rangle = \frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle \quad (2)$$

Here, we have an equal superposition of two two-qubit states, one state where both qubits are "0": $|00\rangle$ and another where both qubits are "1": $|11\rangle$. While we do not know the value of either qubit before a measurement, we know that they must be equal. Thus, reading out one is sufficient for knowing also the value of the second qubit. If, for example, the first qubit is in state $|0\rangle$ after measurement, also the second qubit has to be $|0\rangle$. Further, performing an operation on just one of the qubits immediately affects the second qubit as well.

Physically, a qubit can be any system that can be in a superposition of two states. The ground and an excited state of an atom can, for example, represent $|0\rangle$ and $|1\rangle$. For switching the state from $|0\rangle$ to $|1\rangle$, one could use a laser light of specific frequency and duration, to provide the energy required for the atom to get excited from its ground state to an excited state (one could in principle implement ordinary bits with this scheme, too). By instead shortening the duration of the laser pulse to half of what is needed for the full flip of the qubit, we bring out the quantum: the result is an equal quantum mechanical superposition of both ground state and excited state, of both $|0\rangle$ and $|1\rangle$. It is not that the qubit *either* flipped or not, as it would be with classical objects: the qubit did both at the same time.

Several different physical implementations of qubits exist. In addition to the neutral atom scheme outlined above, for example superconducting loops, trapped ions, diamond vacancies, photonic and topological qubits are all actively developed. Presently, the different types of qubits have complementary strengths and drawbacks, and none of them are superior overall.

The main challenge for all qubit technologies is the effect of environmental noise [Cho 2020] . Noise sources, like temperature, vibrations, and cosmic radiation, interact with the qubits in an unwanted manner. This leads to decoherence, where the qubit loses its superposition, which in turn introduces errors into the calculation. In order to perform a useful calculation, the qubits need to stay in superposition for a sufficiently long time, so that enough computational operations can be performed on the qubits before measuring the result. The required coherence time depends on, among other things, the processing speed, "clock frequency" of the quantum processor. Gate operations are not perfect either, and gate errors will also affect the quality of the calculation.

Error correction schemes for mitigating the decoherence problem are actively being developed. The ultimate goal is to create perfectly functioning, so-called logical qubits. Logical qubits can be realised by combining several, on the order of a thousand, physical qubits [Google 2021]. Another approach is to employ cat qubits, named after Schrödinger's famous feline [Mirrahimi 2014]. At least in the near-term, the reality will be that qubits are noisy and prone to errors. Even if the longest coherence times for qubits already exceed an hour [Wang 2021], this still pales in comparison to the stability of bits in classical systems, where we have become accustomed to errors affecting *any* bit of a calculation to be rarer than the lifetime of the circuits. While the errors introduced by, for example, cosmic rays need to be considered in both classical and quantum computing, the fragility of qubits is in a class of its own [Wilen 2021] .

Today, most of the work on quantum computers uses two-state systems, as in classical computing. It is perfectly possible to use more states, however. Just as classical ternary computers are based on trits instead of bits, qutrits would use three-level quantum systems to represent information, $|0\rangle$, $|1\rangle$, and $|2\rangle$ (or $|-1\rangle$, $|0\rangle$, $|1\rangle$). In general, quantum information units with more than two levels are called qudits, and have some advantages over qubits, *e.g.*, due to their ability to encode information even more densely. Taken to the extreme, discrete variables can be discarded completely, in favour of continuous-variable quantum computing [Hillmann 2020] . In what follows, we will focus on implementations using qubits, as the main ideas of quantum computing remain the same regardless of the number of quantum levels used.

3. Architectures

A rather general definition of a quantum computer is, that it is a device, that directly exploits quantum mechanical phenomena to perform a calculation. This can be implemented in several ways, and quantum computers do come in many flavours and technical implementations. Quantum computers can be grouped into the following three main categories, in order of increasing practical generality and computational power:

1. Quantum annealers,
2. Quantum simulators,
3. Universal, or general-purpose quantum computers.

Quantum annealing exploits quantum tunnelling and entanglement in order to solve a limited set of minimisation or optimisation problems. First, the qubits of the annealer are initialised to their lowest energy state, which is an equal superposition of $|0\rangle$ and $|1\rangle$. Then, the annealer applies biases to each qubit to shift its probability towards either $|0\rangle$ or $|1\rangle$. In addition, couplings between qubits are introduced, which increases or decreases the probability of two qubits to have the same value. In the end, the quantum annealer returns configurations that are close to the "energy minimum" defined by the different biases and coupling strengths. The biases and couplings are problem specific, and defined by formulating an optimisation problem as an Ising problem or through a Quadratic Unconstrained Binary Optimisation (QUBO) model. The Canadian company D-Wave has been offering quantum annealers commercially since 2011 [D-Wave] [Merali 2011].

A quantum simulator is a device that exploits superposition and entanglement to simulate model systems of real systems. This is achieved by mimicking the Hamiltonian evolution of some specific quantum system of interest on the quantum processor. This requires that the problem under study is cast into a form of a model Hamiltonian , *i.e.*, an operator corresponding to the total energy of the system, and determining its time evolution. While the problems amenable to simulation are often physical in nature, also more general optimisation problems can be implemented. As the quantum interactions between quantum particles is a built-in feature of quantum simulators, near-term quantum advantage is expected for the specific class of problems that they can describe. The quantum simulators of the French company Pasqal [Pasqal] provide both digital and analog quantum simulation capability [Henriet 2020] .

Universal quantum computers are the most diverse and potentially the most powerful class of quantum computers. They directly exploit superposition, entanglement, and wave-function interference in order to perform a calculation. A universal quantum computer can, in principle, solve any computable problem, with the additional advantage of up to exponential speed-up over classical computers and algorithms [Deutsch 1985] . Complete universality would require a sufficient number of high-quality qubits for any given problem. The term "universal" is therefore commonly used for quantum computers that operate on the same principle of generality, even if their capacity would fall short of simulating everything imaginable. The term general-purpose quantum computers is also often used for this class. The first demonstration of quantum supremacy, that is, proof that a quantum computer can perform *some* calculation faster than a classical supercomputer, was performed on Google's general-purpose Sycamore processor [Arute 2019]. In Europe, for example the Austrian company AQT [AQT] and the Finnish company IQM [IQM] build general-purpose quantum computers based on ion traps and superconducting circuits, respectively.

Another division of QC technology can be based on the mode of operation: analog or digital. Quantum annealers are analog. Quantum simulators started out as fully analog, but, as mentioned above, can now combine digital computing elements as well. General-purpose quantum computers are digital, and use quantum gates, that is, basic logical operations for manipulating the qubits, and for achieving universality. Digital quantum computers can also benefit from performing parts of an algorithm in an analog manner, combining digital and analog blocks in quantum algorithms [Parra-Rodriguez 2020] .

Twenty years ago, DiVincenzo listed his now famous five criteria that a *general-purpose* quantum computer should fulfil [DiVincenzo 2000].

1. A scalable physical system with well characterized qubits,
2. The ability to initialize the state of the qubits to a simple fiducial state, such as $|000\ldots\rangle$,
3. Long relevant decoherence times, much longer than the gate operation time,
4. A "universal" set of quantum gates,
5. A qubit-specific measurement capability.

Note that criterion 4 cannot be fulfilled by analog quantum simulators that operate without gates. Quantum simulation without gates can in principle be universal, however [Aharonov 2007][Babbush 2014]. Also, continuous-variable quantum computing, which can be considered to be analog, comes with a universal set of quantum gates [Hillmann 2020] .

Constructing the part that performs the quantum computations, the quantum processing unit (QPU), is only the start of a full quantum hardware and software stack. A functioning architecture includes several layers above the QPU: interfaces between the classical and quantum parts; control logic and compilers that translate higher level operations or gates to specific quantum hardware; the actual quantum algorithms and quantum software; and finally, quantum computing theory [Van Meter 2013] [Fu 2016] [Bertels 2021] . Quantum error correction, when in use, is also part of the stack, both at hardware and software level. The quantum software and programming stack is just as crucial an ingredient as the actual QPU for the full stack. All components are needed in order for quantum computing to become a useful tool for doing science with. All of them are also highly non-trivial to implement.

4. Quantum Computer Emulators

It is important to have the full quantum software stack ready to take advantage of the physical QC's when they become generally available. Quantum computer emulators form an integral part of the initial stage of deploying quantum computing to a wide audience. Emulators provide access to quantum computing environments immediately, while access to real, physical quantum computers is still intermittent as physical quantum computing resources are scarce. Thus, emulators enable algorithm development ahead of access to the actual hardware.

A note on nomenclature: various definitions of a simulator and emulator in the context of quantum computing are in use in different communities. Here, we define a quantum *simulator* as a physical device used for simulating quantum mechanical systems and phenomena or problems otherwise beyond the capabilities of classical computers [Georgescu 2014]. While in principle a quantum simulator can be implemented using either universal quantum computers or with analogue devices, we follow the practice of considering quantum simulators to be analog, and in general, not universal or Turing complete. A classical device or software simulating a quantum computer, will in the text be referred to as an *emulator*. This definition will also apply in cases where the software does not necessarily model the actual physics taking place in a quantum computer.

Full emulation of a quantum computer on classical hardware is limited to a maximum of around fifty qubits, due to the exponentially increasing memory requirements of keeping track of the states of qubits. We are therefore already at the limit where the largest existing quantum computers cannot be fully simulated by classical computers anymore. Despite this limitation of using classical computers for emulating quantum computing, emulators have several advantages and features that will keep them relevant for the foreseeable future.

With sufficient hardware resources, emulators can give precision control of modelling the noise in a quantum computer. Thus, the effect of different types of noise can be studied, and bottlenecks in hardware specifications identified. Other hardware constraints, like qubit connectivity and readout errors can also be modelled, and individually assessed. By simulating the inner workings of a real quantum processing unit (QPU), the hardware itself can be improved. At the same time, algorithms can be made more resilient to noise, by, for example, optimising quantum gate operations. Debugging quantum algorithms is made simpler by the ability of reading out the full state of a qubit at any time during the execution of an algorithm. In a real quantum computer, reading the value of a qubit will return only either zero or one, and at the same time, destroy the superposition of the measured qubit. This makes debugging challenging on actual hardware. In general, new practices for code testing, augmenting existing procedures developed for classical software are needed.

Several advanced quantum computer emulators are already available and under continuous development. The Atos Quantum Learning Machine (QLM) [QLM] provides advanced simulation capabilities. From an HPC point-of-view, emulators developed for running on massively parallel architectures are of special interest. These include the Quantum Exact Simulation Toolkit (QuEST), the Jülich Universal Quantum Computer Simulator (JUQCS), and the Intel Quantum Simulator. In addition, toolkits for directly designing quantum hardware, like the open source Qiskit Metal and KQCircuits, fall in the broad category of quantum computer emulators.

Providing access and tuning the performance of quantum emulators on pre-exascale and upcoming exascale supercomputing infrastructures is crucial for extracting maximum synergy from combining HPC and QC. Having emulators play the part of actual quantum hardware in hybrid HPC+QC implementations (see next section) will speed up the development of the required interfaces and practices for connecting classical and quantum hardware into unified computing platforms. From an interconnectivity software point-of-view, whether the quantum processor is a physical device or emulated by software running on classical microchips is of little consequence.

5. Hybrid High-Performance Computing and Quantum Computing

Fault-tolerant, Large-Scale Quantum computers (LSQ) are still a technology of the future. In the present Noisy Intermediate-Scale Quantum (NISQ) era, the scientific research community can, however, already engage in quantum computing research, thanks to the recent availability of publicly accessible, physical quantum computers, in addition to the aforementioned emulators. This has enabled researchers to start developing future quantum algorithms on real hardware. There are several academic and industrial challenges that developers and application owners would face when adopting quantum computing. Such challenges are related (but not limited) to education, programmability and the availability of quantum computing systems. For decades, scientists have continued focussing on accelerating their applications through the adoption of new technological solutions, using CPU, GPU, and AI accelerators, *etc*. Accelerators in general refer to specialised processing units that handle certain computational tasks efficiently. In this sense, Quantum Processing Units (QPUs) follow an established route. The level of acceleration capabilities that quantum computing can bring in a 5 to 10-year time-frame cannot be ignored. A hybrid classical/quantum approach will allow application owners to benefit from the "best of both worlds".

There are also several challenges to overcome for quantum computers to run as separate appliances, namely user access (authentication, accessibility, environment, etc.), data access (input/output), workflow management, orchestration/allocation (batch scheduler), quantum resource management, to name a few. While these challenges need to be addressed also when coupling and properly integrating quantum systems and supercomputers, expertise and experience acquired in HPC during the last 40+ years will ease this integration. Coupling quantum simulators and computers with high-performance supercomputers through a unified cloud-mode access will allow a large part of the scientific community to become familiar with quantum computing, accelerating its adoption.

There are three main modes of integrating HPC and QC (hybrid HPC+QC), schematically represented in Figure 2: (1) stand-alone, (2) co-located, and (3) distributed. An actual implementation can use a combination of all three.

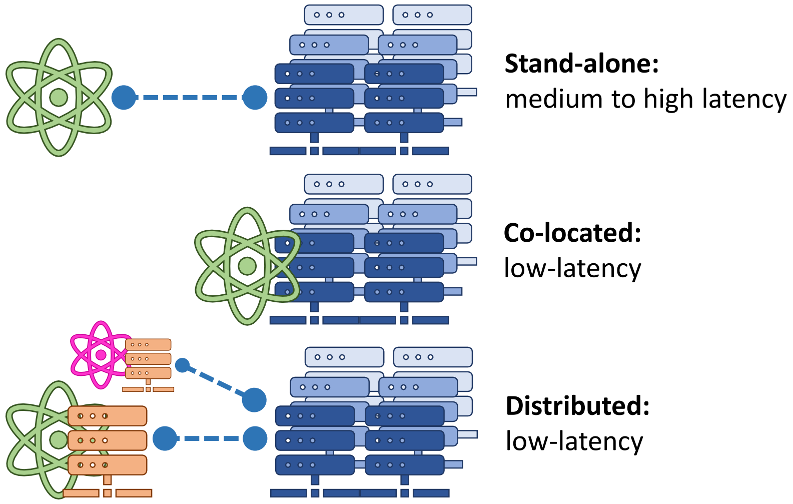


Figure 2: The three main modes of implementing hybrid HPC+QC.

The simplest manner of coupling HPC and quantum machines would be stand-alone quantum systems linked together through a communication network (Ethernet, Infiniband, etc.) to a supercomputer at an HPC centre. The HPC infrastructure then provides the programming environment and the orchestrator/resource scheduler.

For hybrid HPC+QC workflows where part of the calculation runs on classical HPC and part on the QC, latency between the classical and quantum processing units can become an issue. Algorithms that work by feedback loops, where the results of the quantum calculation are used as input for the classical part, and vice versa, can be latency-sensitive. If the execution time of the quantum and classical steps are on the scale of the network latency, then a significant part of the total execution time is spent idling. When QCs grow more powerful and allow for longer execution times for the quantum part of the algorithm, the issue of latency will diminish.

To avoid network-related latencies, an option would be to tightly couple a QPU with a CPU/GPU within a single server chassis. The host system's CPU would support QPU acceleration in, say, a PCI-e form factor, analogous to GPU accelerators, or embed a QPU within a single server. QPU and CPU could even be located on the same chip. While this could drastically reduce network latency, it would introduce additional physical design challenges related to cooling, noise suppression, electromagnetic shielding, ultra-high vacuum enclosures, etc. [Britt 2017] . This requires additional research, since this approach would require an entirely new stack of QC-specific technology that does not yet exist, rather than leveraging current, proven HPC technology.

The purpose of the classical server connected to the quantum computer or QPU, in addition to actually providing the interface between classical and quantum, is to keep the QC busy. For this, it needs the capacity for post-processing the output from the QC and preparing new input for the QC. For top-tier calculations, it will also need to connect to HPC infrastructure. It is worth noting, that the connection between the QC-connected server and the main HPC infrastructure is not expected to be latency-critical.

The server that connects the classical and quantum hardware can either be co-located with the HPC centre, or at a distance, providing a distributed infrastructure (see Figure 2). Both options enable execution of latency-sensitive algorithms, but cannot compete with the envisaged on-chip integration schemes, which would deliver even lower latencies. Co-location is somewhat simpler to implement, as all the components of the HPC+QC solutions are in close proximity. Synergies from research and system administration staff that maintain the two different architectures can also arise.

The distributed approach, where the classical server connected to the QC serves as a middleman between QC and HPC, has advantages. It for example allows QC and HPC servers to be located in spaces best suited for each architecture. The sensitivity to environmental noise requires that quantum computers are hosted in a more shielded space than classical computers. With the recent findings that cosmic rays have a larger than expected effect on qubits, it might become necessary to locate QC installations underground [Wilen 2021] .

In general, the distributed approach will enable higher modularity, increasing inclusiveness of different quantum technological solutions developed in Europe. Compared to the co-located approach, there are some additional complexities that need to be considered. For example, additional data security measures might need to be implemented. Further, scheduling has to consider that the classical server connected to the QC is separated from the HPC. When the distributed software stack is in place, connecting additional QCs into the workflow is rather straightforward. This allows QPUs in different locations to work in parallel on solving the same problem, or for cross-verification of the quantum results [Greganti 2021]

Today, there are on-going efforts in Europe to support quantum start-ups and the adoption of quantum technology by academic and industrial researchers. One such endeavour, the 4-year HPCQS pilot program, is the creation of the first (cloud-based) pan-European hybrid HPC+QC infrastructure that will integrate classical HPC infrastructure with several quantum nodes. HPCQS is an open and evolutionary infrastructure that would have the capability to grow and support diverse quantum computing solutions based on different technologies. The first building blocks of the HPCQS will rely on two 100-qubit quantum simulators from the French start-up Pasqal. By the end of 2022, these will be coupled to the Joliot-Curie (TGCC-CEA) and Juwels (JSC-FZJ) supercomputers through the Atos QLM, which will provide a hardware-agnostic programming environment for end-users.

Europe has thus started to leverage the accumulated knowledge and technological solutions found in the academic and industrial space, including HPC centres. This needs to continue, and different implementations of hybrid HPC+QC solutions need to be actively supported – as with qubit technology, the “winning” concept of combining classical and quantum computing is completely unknown.

6. Algorithms and Use Cases

As discussed in Section 2, qubits can encode much more information content than bits. The difference between bits and qubits grows more pronounced with increasing (qu)bit count. 2 bits can describe 4 different states (00, 01, 10, 11); 2 qubits can describe all 4 states *at the same time*. 3 bits can describe $2^3 = 8$ different states; 3 qubits can describe all 8 states at the same time. 20 qubits can already describe a million states simultaneously. The different states could, *e.g.*, represent inputs on which to perform a computation. With the qubits in superposition, all inputs can be processed in one run; in a classical computer, they need to be computed one by one.

There is a caveat to the massive parallelism that the qubits provide, however. In the end, a quantum computer will only provide one answer. Like ordinary bits, a qubit always returns the value 0 or 1 when read. Similarly, of the million states represented by 20 qubits in superposition, a measurement returns only one, say $|00000000011110111000\rangle$. This makes quantum computers well suited for tasks where one is interested in sifting out the “best” answer out of several possibilities. The travelling salesman problem exemplifies this: we have no interest in all the distances of all possible routes between a set of points, we only want to know the shortest route. If we *were* interested in all the routes, a quantum computer would not speed up the calculation.

Together with quantum entanglement and interference, the superposition of qubits potentially enables a very efficient solution of *certain* problems. Before getting results from a quantum computer, it has to be programmed; quantum algorithms have to be written. Quantum programming differs significantly from classical programming. The key differences are:

1. In classical computing, many basic operations are irreversible, where information is lost after the computation – *e.g.*, AND operands cannot be retrieved from the result of the operation. In quantum computing, all logical operations on the basic information units, qubits, must be reversible, where the input state can be inferred from the output, *e.g.*, NOT.
2. In classical computing, the basic bit operations are simple, either the state of the bit is unchanged or flipped. The basic operations of a quantum algorithm have to exploit the flexibility of qubits, by manipulating the qubits to move around the entire Bloch sphere.

3. In classical computing, the values of the bits can be read at any time. In quantum computing, measuring the value of a qubit collapses the superposition, thereby losing the information of its location on the Bloch sphere. For example, conditional branching inside the code (`if q0=1`) needs to account for this.
4. A classical computer is deterministic, the same input always gives the same output. A quantum computer is probabilistic by design: the results will in general be different, even with identical input values. Often, you need to run a quantum calculation several times to get statistically reliable results.
5. In classical programming, there is usually no need to consider computing errors caused by hardware. In quantum computing, the effect of noise has to be actively accounted for.

Essentially, quantum algorithms need to exploit quantum physical properties that are absent in classical computing, while considering the accompanying limitations. In order to be more efficient at a given task, a quantum algorithm has to explicitly use one or more of the quantum phenomena, superposition, entanglement, and wave interference. Then, polynomial to exponential speedups can be obtained [Montanaro 2016] Figure 3 outlines a general workflow for a quantum algorithm [Johnston 2019] .

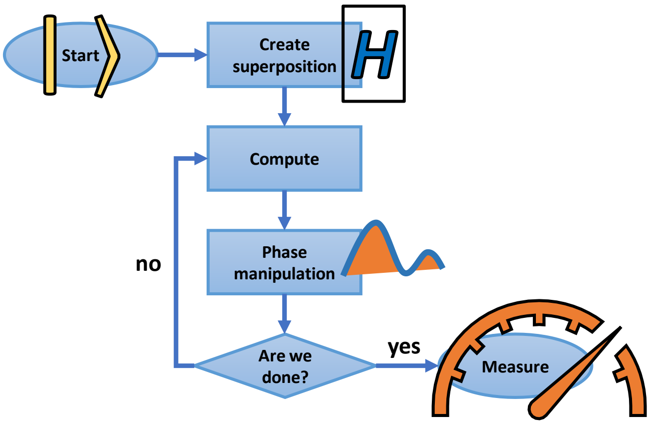


Figure 3: A general workflow for a quantum algorithm. First, an input state is prepared, and a superposition is created. After this, some computation is performed on the qubits, followed by phase manipulation. Generally, phase manipulation is used to increase the probability of the desired result. Finally, the states of the qubits are measured.

The operations for controlling the qubits depend on the quantum computer architecture. Presently, the quantum-gate programming model is the most common one for general-purpose quantum computers. Here, operations on the qubits are represented by operations, *i.e.*, gates on a virtual circuit diagram, see Figure 4. Various basic manipulations on qubits are performed in sequence, represented by symbols on the diagram. The gates can operate on several qubits at the same time, and quantum algorithms are often composed of one, two, and three-qubit operations. At the hardware level, usually only one and two qubit operations are used, while three-qubit gates can be represented as a string of one and two-qubit gates.

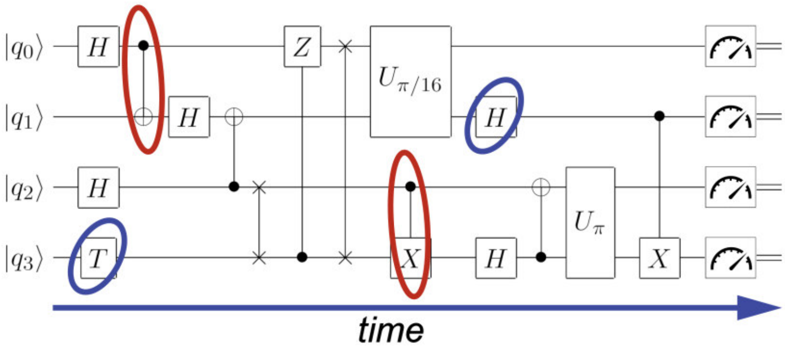


Figure 4: A fictitious quantum algorithm represented as a quantum circuit with quantum gates. The blue and red ovals encircle example one and two-qubit gates, respectively. Measurement of the qubits is at the end.

One or more of the following basic algorithmic building blocks are often present in quantum algorithms:

1. The Quantum Fourier Transform (QFT) [Camps 2021]

2. Quantum walks [Venegas-Andraca 2012]
3. Quantum search algorithms [Giri 2017],
4. Quantum algorithms for linear systems of equations [Harrow 2009] [Dervovic 2018] ,
5. Hybrid Variational Quantum Algorithms [Cerezo 2021] .

Other classes of algorithms include direct Hamiltonian simulation and optimization problems, often on quantum annealers or on quantum simulators. Derived algorithms then combine basic building blocks. The most famous example is probably Shor's algorithm for factorization of integers, a staple of public-key cryptography [Shor 1994]. The list of useful quantum algorithms with varying degrees of speedup grows continuously. The *Quantum Algorithm Zoo* maintains a comprehensive catalogue of existing quantum algorithms [Jordan 2021].

Quantum computers are excellently suited for some computational tasks, while performing poorly for other problem classes, relative to classical computers. Classical computers will always excel at certain tasks, even as universal quantum computers mature and become increasingly powerful. Tasks suitable for quantum computers can be divided into three main classes, which may overlap:

1. *simulating quantum systems*, the original driving force behind the idea of quantum computers — this class includes electronic structure theory, quantum chemistry, and materials science;
2. *optimisation problems* in general — *e.g.* logistics, portfolio optimisation, Monte-Carlo simulations;
3. *machine learning and artificial intelligence* — quantum acceleration is possible for several ML workflows, extending the applicability of quantum computing to a very diverse field of applications.

In addition to basic research efforts, there are various industrial sectors where quantum computing can solve real-world use cases: defence, aeronautics, automobile, finance, energy, travel and transportation, health, pharmacology, materials design, *etc*. Below, we highlight some typical use cases.

Solving the electronic structure of molecules and materials provides a prime example of a problem suited for quantum computers [McArdle 2020]. Today, electronic structure problems are tackled by approximating the underlying quantum-mechanical laws, and then solving these approximations on classical supercomputers. Even the approximate models become prohibitively time-consuming, when a large number of atoms or high accuracy is required. By treating the quantum mechanics of the electronic structure problem on a quantum computer, fully transforming an inherently quantum-mechanical problem to a classical computer can be avoided.

For quantum-chemistry problems, quantum algorithm development is an established and flourishing field [Cao 2019]. Electronic structure problems are central to several other fields of science as well. In the pharmaceutical sciences, the drug discovery phase is hampered by the inaccuracy of current models. With the aid of quantum computers, completely novel drug classes can be discovered. Energy storage and green energy catalysis are also based on quantum-mechanical interactions. Chemical catalysis involves the rearrangement of atoms and electrons to form new products from various reactants. Discovering new ways of synthesizing molecules and materials is of high industrial impact. Of equal importance is understanding natural processes. For example, understanding atmospheric chemistry in sufficient detail is pivotal for understanding and tackling climate change.

Even if simulating quantum systems was the original motivation for quantum computers, near-term quantum advantage is expected especially in modelling problems not directly related to quantum phenomena. Optimisation problems are abundant in, *e.g.*, finance modelling and risk assessment and analysis. In the energy sector, another use case is related to smart charging of electric vehicles to optimize the charging of thousands of electric vehicles combined with the charging-point network in an efficient way [Dalyac 2021]. The problem is a combinatorial optimization of several criteria, with increasing complexity with increased number of variables. The interest is to find the best order for the vehicles to be recharged according to users' priorities: the charging time required, charging-point availability, proximity of the terminals to vehicle location, and so forth. With increasing scale, quantum computers can excel in finding the optimal solution.

Modelling climate change at a larger scale, beyond molecules, can also benefit greatly from the increased optimisation capability of quantum algorithms. As mentioned, linear systems of equations, used for example in weather forecast models, can be solved much faster using quantum algorithms [Harrow 2009] . Accurate weather forecasts can have a direct, positive impact on food production and transport efficiency, for example.

Machine learning (ML) and artificial intelligence (AI) are used intensively for solving a large body of problems, from speech recognition, *via* targeted advertising, to optimising cancer treatments. An increasing portion of available computing capacity is spent on running AI algorithms. For this broad field, quantum computers are expected to have a large impact, and quantum machine learning (QML) is attracting much deserved academic and industrial attention [Dunjko 2020] [Huang 2021]. Quantum computers can be used both to enhance and speed up traditional machine learning procedures, as well as for studying truly quantum mechanical datasets [Uvarov 2020] .

In essence, *all* fields where computational modelling is used, will be able to benefit from quantum computing, from biology [Emani 2021] to digital humanities [Barzen 2020] At least some components of practically all complex modelling problems are amenable to quantum speed-up, if formulated in a suitable manner.

7. Academia and Industry in Europe

The European quantum ecosystem is globally strong, much due to academic researchers and start-ups. The key to set European future competitiveness resides in Europe's capability to enable research entities and industries to exploit these new technological and complex quantum solutions in real-world use cases. Industries know the issues their sectors are facing and own business data, while academic consortia and start-ups have the expertise and know-how to support industries to identify if a use case has a "quantum" solution and if so, translate the use cases into quantum algorithms. By uniting the efforts of academia, including research and technology organisations (RTOs), and industry, the impact of quantum computing can be maximised.

Start-ups focusing on quantum computing are at the intersection of both worlds, often having an academic background. It is important to create and support start-ups that would bridge together European academia, start-ups and industries. This would help established industries to understand resource needs such as R&D, education, hiring, and financial investment, that are required for investing in quantum technologies, and thereby assess whether the investments are worth it. Start-ups, on the other hand, would gain visibility to the European community, improve their technology readiness levels (TRLs) and ease knowledge transfer to and between industrial and academic scientists.

Several European countries are already supporting quantum technology start-ups. Next, we highlight a few of them; the list is in no way exhaustive, but serves to show that such initiatives are well-received and fruitful. A recent report by the Canadian Institute for Advanced Research provides a more exhaustive overview of policy measures taken by different countries to support quantum R&D [Kung 2021].

In the Paris Region of France (Région Ile-de-France), GENCI in conjunction with Le Lab Quantique, a non-profit organisation promoting quantum computing, has created the PAck Quantique (PAQ) initiative. PAQ funds projects tackling major industrial challenges by partnering start-ups and academic consortia. In 2020, this three-year program has funded three real-world use-case projects in the energy (EDF, Total) and pharmacology (Qubit Pharmaceuticals) sectors. The success is expected to lead to an expansion of the program at the French national level. The second initiative in France is the Teratec Quantum Computing Initiative (TQCI) which aims to foster synergies between academia and industry in order to rapidly build up skills and develop know-how in the field of quantum computing, bringing together future users, technology providers and research centres.

In Finland, a collaborative project between the VTT Technical Research Centre and the globally recognised Finnish start-up IQM are constructing a quantum computer, planning to reach 50 qubits by 2024. The Finnish Quantum ecosystem in general is emerging strongly. The recently formed InstituteQ: The Finnish Quantum Institute combines expertise in quantum computation, communication, sensing and metrology, and simulation. Another recent initiative in Finland is Business-Q, which connects industry, universities, research and technology organisations, and investors. Business Finland, the government organisation for innovation funding and Helsinki Business Hub have been supporting quantum industry and start-ups for several years.

In Germany, a recent initiative, Quantum Technology and Application Consortium (QUTAC) consists of ten of the most influential German companies, who will use their resources and know-how to accelerate German innovation in the field of quantum computing [QUTAC] .

The Netherlands have also begun a quantum computing initiative, Quantum Delta NL, which has already allocated 615 M€ in funding. The Quantum Delta NL is a private-public partnership tasked with organising the implementation of the QC projects from the Netherland's National Agenda for Quantum Technology (NAQT). The funding will be dedicated to train 2000 researchers, host three commercial R&D infrastructure by 2027 and support up to 100 start-up companies.

On a European scale, the Quantum Flagship initiative funds six projects on quantum computing and simulation [QT.EU]: AQTION – Advanced quantum computing with trapped ions; NEASQC – Next ApplicationS of Quantum Computing; OpenSuperQ – An Open Superconducting Quantum Computer; PASQuanS – Programmable Atomic Large-Scale Quantum Simulation; Qombs – Quantum simulation and entanglement engineering in quantum cascade laser frequency combs; and QLSI – Quantum Large Scale Integration in Silicon.

The European Quantum Industry Consortium (QuIC) has the objective of nurturing a fair and sustainable quantum technology business environment in Europe and ensure its global competitiveness [QuIC] .

It is of interest to identify the strengths of Europe in the field of quantum technologies and more precisely in quantum computing. Europe covers the entire domain of the quantum value-chain, from cryogenic cooling systems (*e.g.* Air Liquide, France; Bluefors, Finland), quantum communication (*e.g.* KETS, UK; SSH.COM, Finland; Cryptonext Security, France), quantum sensors (*e.g.* Qnami, Switzerland) up to quantum networks (*e.g.* QPHOX, The Netherlands; VeriQloud, France). Atos is heavily invested in quantum computing, with the Quantum Learning Machine (QLM) programming environment and QPU front-end. Academic institutions furthest along the path towards hundred-qubit quantum computers include the Wallenberg Centre for Quantum Technology (WACQT), directed from Chalmers University of Technology (Sweden) [WACQT] and the joint efforts of ETH Zürich and the Paul Scherrer Institute (Switzerland) [ETHZ 2021] . European quantum computing and quantum software start-ups and SMEs are numerous and visible on the international market. Table 1 lists a selection of these, showing the diversity of the European quantum computing domain.

European start-ups and SMEs	Country	Quantum computing	Quantum software
Algorithmiq	Finland		Life sciences
Alice & Bob	France	Cat qubits	
AQT	Austria	Trapped ion qubits	
C12 Quantum Electronics	France	Carbon nanotube qubits	
Cambridge Quantum Computing	UK	Quantum compiler	Chemistry, ML, finance
CEA-Leti	France	Silicon qubits	
HQS Quantum Simulations	Germany		Materials science
IQM	Finland, Germany	Superconducting qubits	HW/SW co-design, finance, KQCircuits
Multiverse Computing	Spain		Finance
ORCA Quantum Computing	UK	Photonic qubits	
Oxford Ionics	UK	Trapped ion qubits	
Oxford Quantum Computing	UK	Superconducting qubits	
Pasqal	France	Rydberg atomic qubits	Pulser
Phasecraft	UK		Materials
Qilimanjaro	Spain	Quantum annealer	
Qu & Co	The Netherlands		Chemistry & materials
Quandela	France	Photonic qubits	
Quantastica	Finland, Estonia, Serbia	Quantum emulators	Programming tools
Quantum Motion	UK	Silicon spin qubits	
Qubit Pharmaceuticals	France		Pharmacology
Rahko	UK		Drug discovery
Riverlane	UK		Deltaflow.OS quantum computer operating system
Terra Quantum	Switzerland		Cybersecurity

Table 1: Some European start-ups and SMEs actively developing quantum computing technologies.

Internationally, technology giants such as Alibaba, Amazon, Fujitsu, Google, Honeywell, IBM, Intel, Microsoft, and Huawei are investing significant resources into quantum computing. Compared to the US and Asia, the European quantum computing industrial landscape differs. At the end-user side, several large companies are involved, as noted above. Almost all of the development activity is due to start-ups and SMEs, however, with only a few incumbent firms involved. This makes the ecosystem vibrant and agile, but at the same time, vulnerable to challenges related to scaling capacity and financial stability. Therefore, pan-European efforts are needed to ensure that European start-ups have the necessary resources, incentives, and support available for continuing building their respective business cases *in Europe*. This would mitigate the risk of mergers with large non-European corporations. Incubation of a critical mass of successful start-ups is needed [Räsänen 2021]. The success of national support programmes has been demonstrated, now is the time to scale up.

8. Quantum Communication

In quantum communication, digital information is transferred using qubits, instead of bits; quantum states are transferred from one place to another [Gisin 2007]. This can be exploited in several ways. In this section, we briefly cover its application in securing digital communications, and in the context of a quantum internet.

Encryption of digital information is crucial for modern society. Online banking and e-commerce, state and company secrets, medical records and personal communication of individuals, all depend on reliable digital security. Digital encryption is based on decryption being a computationally hard task: by ensuring that breaking the encryption takes long enough, the information is practically secure. Shor's algorithm for factoring integers into prime factors [Shor 1994] showed that sufficiently advanced quantum computers can speed up the mathematical task of breaking encryption protocols. Thus, revision of commonly used encryption methods has become urgent. Quantum communication can mitigate some of the risks that quantum computing poses to digital security.

Digital encryption is based on so-called encryption keys, which are used to encrypt and decrypt information. Like all digital information, the keys can be represented as bit strings of zeroes and ones. From an encryption point-of-view, the advantage of using qubits during the transfer is, that is becomes possible to detect if someone is eavesdropping on the data transfer. This is because measurement of a quantum system, like a photonic qubit, necessarily disturbs the system. By using proper communication protocols, it thus becomes possible for two parties to exchange the secret encryption keys required for encrypting communications with the certainty that no-one has intercepted the key exchange. This is the basis for Quantum Key Distribution (QKD).

Since the Digital Assembly event in 2019, All 27 EU member states have signed the European Quantum Communication Infrastructure (EuroQCI) Declaration, committing to work together to build a pan-European, secure quantum communication infrastructure [EuroQCI] . This will be used for data transit and storage in a highly secure manner. It will ultimately link sensitive public and private communication assets all over the EU, such as banks and administrations. EuroQCI assumes both the construction of a network of terrestrial connections and the launch of a satellite segment - combining the technological potentials and capabilities of individual Member States. Although the communication mediums vary, the key motivation remains the creation of a common network that enables the use of quantum communication to improve the exchange of sensitive and classified data.

The first use of the EuroQCI infrastructure will be to provide QKD services. Work in this direction is underway in the OpenQKD EU pilot project [OpenQKD] , which connects European academia, industry and start-ups in the deployment of open QKD testbed sites around Europe. QKD is already being commercialised, with several QKD products appearing on the market. As a testament to the maturity, even satellite-based QKD is being deployed worldwide. For example, the European Space Agency (ESA) and the UK-based company ArQit are working on the QKDSat project, which aims for a satellite launch in 2023 [ESA 2021] .

While QKD offers enhanced security, it should not be considered *fully* secure. In any real-world implementation, security weaknesses will arise. For example, the French National Cybersecurity Agency [ANSSI 2020] and the National Cyber Security Centre of UK [NCSC 2020] discourage relying on QKD for ultra-secure applications. Instead of QKD, Post-Quantum Cryptography (PQC), that is, new encryption algorithms that are considered safe also against advanced quantum computers can be employed [ENISA 2021] A practical advantage of PQC over QKD is that PQC requires no new infrastructure to operate. Europe has the potential to be world-leaders in this field as well [Loesekrug-Pietri 2021]. For this, national programmes like PQC Finland [PQC.FI] need to be augmented by pan-European efforts.

Generation of truly random numbers is central to cryptography, including QKD and PQC, and is important for various computational simulation workflows and gaming. The inherent randomness of quantum systems lies at the heart of Quantum Random Number Generators (QRNGs). Several commercial QRNG systems are already available. Here too, has collaboration between academia and industry been fruitful, as exemplified by the secure, high-speed quantum random number generator developed in collaboration between the Universities of Sheffield and York, the Technical University of Denmark, and the Danish company Cryptomathic [Gehring 2021].

In quantum teleportation, information can be exchanged between two qubits over a long distance [Pirandola 2015] . First, two qubits (A and B) are entangled, by creating a Bell state (see Section 2). After this, the qubits can be transferred to separate locations. A third information qubit can then be made to interact with qubit A. This instantly affects qubit B, with information transferred through the quantum communication channel created by the entanglement. In order to make sense of the information, classical information needs to be sent from A to B as well (for example over the classical internet), which prevents superluminal information transfer.

Teleportation is the basis for a quantum internet [Castelvecchi 2018]. In addition to information transfer, a quantum internet could be used to combine the qubits of separate quantum computers. This type of parallel quantum computing would be revolutionary. By properly combining, say, two 50-qubit quantum computers, the result would in essence be a 100-qubit machine. Note the exponential increase in theoretical computing power for quantum computers: this would not *double* the computing capacity, the capacity would increase by a factor of 2^{50} = 10^{15}, a quadrillion. The long-term ambition of the European Quantum Internet Alliance (QIA), coordinated by QuTech in the Netherlands, is to build a quantum internet that enables quantum communication applications between any two points on Earth. This is pursued by developing, integrating and demonstrating all the functional hardware and software subsystems required [QIA] .

9. Green Transition and Digitalisation

From the global energy consumption point-of-view, quantum computing offers a promising augmentation to digitalization. The fundamental logical qubit operations use very little energy [Ikonen 2017, Chiribella 2021]. The major part of the energy consumption arises from cooling of the quantum computer. The most mature quantum computer technologies need to be cooled down close to absolute zero (–273 °C). The space that needs cooling is, however, small; there is no need to cool down the entire space, only the processor itself. This is in stark contrast to a traditional data centre. A quantum computer itself emits very little heat, which further decreases the need for cooling. For example, Google’s Sycamore quantum computer setup, which was used for demonstrating quantum supremacy, used less than 26 kW [Arute 2019].

As quantum computers operate on very different computing principles compared to classical computers, comparing the energy requirements of, say, qubit and bit operations is not meaningful. Instead, one has to compare the total power consumption for running specific calculations. For problems that the D-Wave quantum computer is suited for, the energy need has been estimated to be just 1% of what a digital supercomputer consumes [D-Wave 2017] . For general-purpose quantum computers, an advantage of many orders of magnitude in energy consumption over classical supercomputers is estimated [Villalonga 2020] As the speedup offered by quantum computers is superlinear, it is expected that energy consumption grows very modestly with increasing utility of a quantum computer. Thus, the ratio of computational power to energy use will become increasingly favourable in the future.

Direct benefits of quantum computing for the green transition include enabling highly accurate and reliable solutions to problems that are intractable using traditional high-performance computing. These areas include solving the electronic structure problem of atmospheric reactions, green catalysts for clean energy production and CO_2 sequestration, environmentally friendly alternatives to chemicals, and materials for solar cells and batteries, to name a few. Optimisation problems form another class relevant from an environmental perspective, including the logistics of travel and transport, flood prediction and crop optimisation, and in general, industrial processes. A sufficiently advanced quantum computer could model and solve these problems with unprecedented accuracy and speed, far surpassing the present and future capabilities of traditional supercomputers. Quantum computers can thus directly help decrease greenhouse gas emissions in the future.

The use of quantum computers for solving other problems, not directly related to sustainability, would have an indirect but notable effect on the energy consumption of high-performance computing. Parts of most modelling problems can be solved more efficiently using quantum computing. Some of the computationally most intensive problems running on current HPC infrastructure are well suited for quantum computing. A prime example are electronic structure calculations, which consume a major part of HPC resources: at CSC, almost half of the supercomputer capacity is spent on solving electronic structure problems; on GENCI’s supercomputers, combining all problems related to chemistry, including electronic structure, represents 30% of the available computing hours. Machine learning and artificial intelligence form another, still growing application field that uses increasing amounts of energy, but at the same time is amenable to significant acceleration by quantum computers. Thus, offloading suitable tasks to quantum computers decreases the load on traditional supercomputers. This opens up the possibility to curb the increasing energy-demand of data centres. The shift towards quantum computing in HPC will lead to the production of significantly higher-quality scientific returns per Watt consumed.

As discussed in detail in the section on hybrid HPC+QC, quantum computers are not separate appliances; to extract full advantage from quantum computing, powerful, traditional HPC infrastructures are required. In this scheme, the classical supercomputer capacity stands for the vast majority of the power consumption. A distributed HPC+QC approach, where the classical HPC and the QC resources can be geographically distant is ideal, also from a sustainability point-of-view. The traditional, power-hungry HPC centres can be set up in areas where, for example, cooling needs are lower, waste heat can be readily reused, and low-carbon power sources are available (as in, e.g., the LUMI pre-exascale EuroHPC data centre in Finland), while the low-power quantum computers can be located practically anywhere, provided a high-speed internet connection is in place.

Sufficient user support actions, discussed in the next section, have a direct impact on the green transition as catalysed by QC. The earlier users are able to adopt and utilise quantum computing for their specific problems, the sooner a transition towards more impactful science, research, and development can take place. Considering the high impact that utilising quantum computers can have on solving problems crucial for society, each additional year that the inauguration can be brought forward counts.

10. Competence Development and Education

Several highly interesting technologies that could have revolutionised the world turned out to be a failure. Despite large investments in both hardware and software, including subsequent maintenance costs, some technologies have found only niche use in few areas. This is partly due to failures in engaging end-users. This has resulted in low user uptake in fields that *could* have seen a benefit, but failed to do so due to insufficient support for non-experts to port their problems and codes to new platforms and technologies.

The introduction of quantum computing along with exascale systems within a short timeframe in Europe requires both reaching production stability of the computing systems, and sufficient literacy of users to program and harness such unique computing technology. We remember from the past several use cases of effective use of new technologies and the efforts in the preparation of algorithms for:

- Vector computers in the late1980's,
- Massively Parallel Processing supercomputers (MPP) in the 1990's,
- Accelerators (GPUs) at the beginning of 2000’s,
- And the less popular use of FPGA accelerators, starting in the 1980’s.

Open, standardised, and portable programming models are of high importance for end-users in order to, as far as possible, avoid re-writing their applications on different quantum hardware implementations and architectures. It is especially important to provide a flexible way to move applications from one to another architecture automatically or at least semi-automatically with minimal effort.

The introduction of QC in Europe in a rather short time-frame is challenging. At the same time, urgency is crucial for both the advances in computational modelling that QC provides for problems of high societal impact, and for ensuring European digital sovereignty in the field of quantum computing. Therefore, the EuroHPC programme has to support the development of new algorithms and novel applications. It is imperative to:

- Develop appropriate tools to facilitate the preparation of algorithms in specific application areas, from low-level libraries and tools to high-level portable programming languages;
- Establish competence centres for users and to reinforce existing HPC competence with HPC+QC support;
- Develop programming courses that focus on the effective use of quantum computers;
- Raise awareness among all user groups, starting from young scientists, even students at various levels, to professionals in the general field of computational modelling, to decision and policy makers.

As QC will not replace classical processing, but instead, augment it, hybridisation of existing HPC centres with new QC technology is a logical next step. This requires training and competence development. The extensive experience of the Partnership for Advanced Computing in Europe (PRACE) community in creating a training network of competence centres should be leveraged. This should be done in collaboration and concord with the training activities organised by the EuroCC programme, which aims to bring together the necessary expertise to set up a cross-European network of National Competence Centres (NCCs) in HPC-related topics with 33 participating members and associated states [EuroCC] .

The PATC (PRACE Advance Training Centres) and the PRACE portal of trainings [PRACE] could be used for training courses of quantum technology prior to the establishment of pan-European quantum computing systems. PRACE organises a programme of 90+ courses annually, via its 14 PRACE Training Centres (in Austria, Belgium, Czech Rep., Finland, France, Germany, Greece, Ireland, Italy, the Netherlands, UK, Slovenia, Spain, and Sweden). This training is complemented by Seasonal Schools and special on-demand events that are run in collaboration with other projects and European Centres of Excellence (CoEs).

Only a broad process of raising awareness and competences in the future academic and industrial communities will result in an increased level of use of QC methods and the creation of new algorithms, and thereby full leverage of the technology. At the same time, it is important to mitigate the risks involved with an emerging technology. The educational programmes and degrees need to be specific enough to enable the future graduates to successfully work on quantum technologies, and at the same time, diverse enough to ensure that the acquired skill set is not lost in case the job market growth falls short of predictions and expectations.

In Europe, career-affecting educational choices are often made at the end of elementary school, roughly around the age of 15. Thus, right now, many young minds are making decisions affecting the talent pool of the matured quantum computing era. The field can attract young talent by active engagement with students at high-school, even primary school level. HPC centres and higher-education institutions can help by actively promoting awareness of QC, and showing that the field provides a viable and exciting career path.

11. Future Directions and Challenges

Quantum computing is developing with accelerating speed, with no slow-down in sight. The technology has stepped out of the academic lab, and is growing commercially [MacQuarrie 2020] The Boston Consulting Group notes that equity investments in quantum computing almost tripled in 2020 [Bobier 2021] . Market data analyst Pitchbook reports that by early September, global venture capital investments in QC for the year 2021 surpassed USD 1 billion, more than during the previous three years combined [Temkin 2021] Estimates of the future trends on the market growth of quantum computing vary, but essentially all predict continued increases in investments, revenue, and job market size. For example, in their latest forecast on quantum computing, Inside Quantum Technology IQT predicts a total revenue of roughly EUR 2 billion by 2026, a ten-fold increase compared to 2020 [IQT 2020] . McKinsey predicts a whopping trillion-euro value potential by mid-2030s [Hazan 2020] . Expectations are thus exceptionally high.

To live up to the expectations, the technology needs to demonstrate constant progress. Keeping the momentum going can be considered to be the first challenge of quantum computing. As discussed, qubits are very sensitive to noise, and are still far from functioning in an ideal manner. In addition, the present qubit count is modest. The reportedly most powerful general-purpose quantum computer, the Chinese Zuchongzhi, has 66 superconducting qubits [Wu 2021] . We are therefore presently in what has been coined the Noisy Intermediate-Scale Quantum (NISQ) era [Preskill, 2018] .

Quantum advantage is possible also with NISQ devices, and incremental speed-up for HPC is to be expected, but NISQ cannot deliver on the promise of a quantum *revolution* in computational modelling. For this, fault-tolerant quantum computers with long coherence times and higher qubit count are required. Quantum computing needs to scale up, and achieving this is anything but trivial. Despite, or perhaps rather, *due* to the challenges, large-scale quantum (LSQ) has to be kept as the ultimate goal, and work on paving the way towards it needs to be resolute. It is worth noting that the transition from NISQ to LSQ will not be a binary event. Fittingly, we will see a period with a superposition of NISQ and LSQ, with larger and less noisy quantum computers.

Several qubit technologies have been implemented and shown to work, at small scale. The main challenge is not manufacturing the individual qubits anymore, even if advances in qubit quality are still required. Now, control and calibration of ever more densely packed qubits is becoming problematic. Connecting the qubits to create, say, a million-qubit entity is supremely challenging. The cabling work required for qubit control lines alone is massive. In the end, it might turn out that connecting several smaller quantum computers over a quantum network, for example the quantum internet, will be the method of choice for scaling up QC. Already, qubits 60 metres apart have been connected to perform a concerted calculation[Daiss 2021][Hunger 2021] .

The second major challenge on the road towards large-scale quantum computing is quantum error correction (QEC) [Devitt 2013][Terhal 2015]. QEC is needed due to qubit decoherence, gate inaccuracies, readout errors, *etc.* In classical systems, errors can be mitigated by redundancy, by having multiple copies of the same data. This is not applicable in quantum computing, due to the no-cloning theorem, which states that it is impossible to create an independent and identical copy of a qubit. Instead, the information content can be spread over several entangled qubits, thus creating a logical qubit out of several physical ones. To reach fault-tolerance, the ratio of physical to logical qubits is on the order of a thousand to one. Thus, present-day quantum computers have yet to reach the milestone of implementing *one* fully fault-tolerant qubit.

To reach LSQ, classical computing plays a large role. The importance of emulators was discussed previously, as was the prospect of quantum computers for accelerating machine learning. Conversely, machine learning can also speed up quantum computers. ML schemes have already been developed for calibration of qubits[Genois 2021] , error correction schemes [Nautrup 2019][Sweke 2021] , and algorithm optimisation [Cincio 2021][Fösel 2021] .

With the increasing complexity that comes with larger quantum computers, the role of ML/AI is expected to grow. It might, for instance become necessary to use supercomputing resources for optimising and compiling quantum algorithms for specific QPUs, resources that themselves could already be of the HPC+QC variety. The interplay between classical and quantum will continue for the foreseeable future.

The software development stack for quantum computers needs to evolve, in order to facilitate quantum programming. Without the software for solving specific classes and types of problems, quantum computers are rather useless. Ready-made algorithm and subroutine libraries are sorely needed, and significant resources should be dedicated for developing real-world scientific modelling packages that incorporate quantum algorithms. Quantum software development kits and ready-made application software needs to be made available to the end-users, alongside access to the hardware. This should be coupled to the introduction of HPC+QC solutions to R&D communities. HPC+QC deployment has already started at several European HPC centres: CINECA (Italy), CSC (Finland), ICHEC (Ireland), JSC (Germany), LRZ (Germany), SURF (The Netherlands), and TGCC (France).

User engagement and uptake of quantum computing is challenging. Especially in the near-term, when gains in computing power will be incremental rather than transformative, the barrier to adoption will be high. Even for users well versed in the art of writing software for traditional HPC systems, the leap to start developing quantum applications is a long one. From the perspective of an end-user comfortable with *using* HPC resources rather than developing them, QC can seem even more esoteric. Encouraging adoption at a stage when the technology readiness level of QC is low, can even be detrimental to the long-term prospects of quantum computing. It is, however, important to continuously build up the quantum software stack. Engaging end-users early on, even when the technology is far from being a polished product, enables a constant dialogue between those developing the technology and those that will take it into practical use. To ensure early real-world quantum advantage, engineering efforts need to align with user needs.

Large-scale quantum computers are often seen as special hardware that, like the mainframes of old, would exclusively be large-scale also by size. Already, this notion is being challenged. For example, Australian-German Quantum Brilliance is developing portable quantum accelerators, envisaged to be the size of the graphics card by 2026 [Doherty 2021] . Austrian AQT is already offering quantum computers in standard 19-inch rack format [Pogorelov 2021] . The most powerful QC solutions of the future will with all probability require tailor-made hosting spaces, however, as the classical supercomputers of today. The mere feasibility of consumer-grade QPUs does illustrate how far from the experimental lab-space QC has already ventured.

We are at the verge of witnessing whether QC will begin delivering on its promises in earnest. The public roadmaps of both IBM and Google foresee reaching a million qubits by 2030. More interestingly, IBM has disclosed its shorter-term goals, including reaching the 1.000-qubit milestone by the end of 2023, with steady increase in qubit count in-between now and then [Gambetta 2020] . Reaching the milestones will certainly be challenging, and at the same time, serve as a motivation and driving force for QC development in general. In Europe, the global IT player Atos is visibly investing in QC, and many talented teams are developing QC solutions, industrialised via various start-ups. As discussed in Section 7, the European challenge is to ensure the competitiveness of the largely start-up driven quantum ecosystem on the global scene. The EU Quantum Flagship project has a plan for global stimulation of the ecosystem, which augments and complements national efforts. Sustained, sufficiently long-term EU-level support for the QC industry will be needed to ensure that European competence and capacity in the field reaches the critical mass of self-sustainability.

12. Conclusions and Summary

Quantum computing has begun to deliver on its promise of revolutionising computational modelling and supercomputing. We are still at the very early stages of this new computing era, however. In order to become a usable tool for end-users, several unsolved challenges must be overcome.

From the hardware point-of-view, the most significant issues to be addressed are scaling up the qubit count and increasing the error-resilience of quantum computers. All bets are off when it comes to which specific technologies will emerge as champions in the race towards real quantum advantage. Several different physical implementations of qubits and the controlling machinery are being explored. Most likely, some technologies will turn out to be ideal for some types of calculations, while others are better at solving other types of problems.

Although the leading quantum computing architectures and systems of the future are unknown, what we *do* understand is that the future of quantum computing is hybrid. Not only in the sense of combining classical and quantum computing in an HPC+QC platform, but also when it comes to the implementation of the quantum part. We foresee a future where several quantum technologies that presently are developed rather independently, will ultimately merge: digital and analog quantum computing; superconducting, trapped ion, neutral atom and photonic qubits; qubits, qutrits, and continuous variable data representations; quantum annealers and universal quantum computers; task-specific and general-purpose hardware; the list goes on. A comparison to classical platforms is appropriate. In the 1970s and 1980s, many supercomputers were based on vector instructions. Around 1990, microprocessors based on scalar instructions took over, and pure vector machines were mostly abandoned. By the end of the 20th century, vector instructions were reintroduced, and the microprocessors of today are hybrids of both scalar and vector architectures.

As no single technological solution is expected to dominate even in the long-term, it is imperative to explore various options and to bet on as many European Schrödinger horses as possible. This can ensure a successful build-up of a homegrown and independent European quantum technology ecosystem. The risks associated with locking down on only a few approaches at this stage are enormous: Europe could create a future situation where neither the hardware nor the know-how of the most powerful supercomputer solutions can be found within the continent.

In addition to the hardware, also the software and user interfaces for the upcoming quantum systems need to be developed. Due to a very different programming model, quantum software development requires significant effort and support. In contrast to what has been the norm for some time now, it is not only a question of porting existing codes to quantum computers, but rather fully rewriting applications and changing our mindset to "Think Quantum". A multi-tiered, don't-place-all-your-eggs-in-one-basket approach needs to be taken that allows for technological progress without having to wait for which quantum technologies will be available or become dominant in the future. Europe must forge forward in the interim in a way that allows for the development of new algorithms exploiting computational features that have never before been used in programming. This requires high-level portable programming environments. In addition, the current computational problems themselves have to be recast to a form suitable for quantum algorithms. This requires an immediate and massive educational program and support for industrial adoption. Exploiting the synergies between academia and industry is far more important than in traditional HPC. The development of widely accepted standards for programming and tools is equally important. This will be a cornerstone of the future, and Europe can be pro-active in this area – Europe must not lag behind nor depend on closed solutions, and actively support open-source efforts.

For catalysing the uptake of quantum computing, incorporating QC into existing supercomputing infrastructure is essential. For real-world problems, quantum computers will never *replace* classical computers, but rather become an integral part of HPC. Europe has the opportunity to set up unique hybrid supercomputing infrastructures by seamlessly incorporating quantum technology, and by utilising the expertise of HPC centres built up over decades. Europe is well on track in setting up individual HPC+QC hybrid infrastructures. A federated user access approach, where a login to any of the EuroHPC supercomputers would give access to a large selection of quantum computing resources, resources will allow end users to find the most suitable technologies for their specific applications. This would provide European scientists an advantage over international competition.

Efforts should be pan-European, following the already proven success of the EuroHPC Joint Undertaking for setting up petascale and pre-exascale computing facilities, and the plans for going to exascale and beyond.

As with the traditional supercomputing infrastructure, diversity is key also when setting up a distributed and federated European QC backbone. Endeavours should be concerted, but not overly concentrated. Initially, setting up at least two, and ideally three European HPC+QC hubs would represent a good balance between joint effort and risk-mitigating diversity.

The turning point for science, when research *about* quantum computers shifts towards research *using* quantum computers draws closer. Still, we need to actively remind ourselves that quantum computing is in its infancy. It needs to be nurtured by a sustained and significant support for fundamental and basic research into quantum technologies, both on the hardware and software side. Only then can European competence and competitiveness in the fields of quantum technology in general, and quantum computing in particular be ensured. The basic research on quantum technologies performed for decades in partnership between all European countries, whether members of the European Union or not, has laid the seeds for the fruits we reap today. In order to push these boundaries further, this collaboration needs to continue, with minimum restriction.

IBM and Google aim for a million qubits by 2030. Europe too needs to set its goals for quantum computing sufficiently high to keep up with global developments. Maintaining an open and inclusive spirit of international cooperation is what keeps Europe an appealing target for international investments and talent also in the future. To reach the number of qubits needed for truly disruptive quantum computing, the roughly one hundred noisy qubits that we have today have to be scaled up by four to five orders of magnitude. Achieving this feat requires dedication, sustained funding, and a worldwide pooling of talent and technology.

13. References

[Aharonov 2007] D. Aharonov *et al.*, "Adiabatic Quantum Computation is Equivalent to Standard Quantum Computation", *SIAM J. Comput.* **37** (2007) 166. https://doi.org/10.1137/S0097539705447323

[ANSSI 2020] "Should Quantum Key Distribution be Used for Secure Communications?", Agence nationale de la sécurité des systèmes d'information (2020). https://www.ssi.gouv.fr/en/publication/should-quantum-key-distribution-be-used-for-secure-communications/

[AQT] Alpine Quantum Technologies. https://www.aqt.eu/

[Arute 2019] F. Arute *et al.*, "Quantum supremacy using a programmable superconducting processor", *Nature* **574** (2019) 505. https://doi.org/10.1038/s41586-019-1666-5

[Babbush 2014] R. Babbush, P. Love, A. Aspuru-Guzik, "Adiabatic Quantum Simulation of Quantum Chemistry", *Sci. Rep.* **4** (2014) 660. https://doi.org/10.1038/srep06603

[Barzen 2020] J. Barzen and F. Leymann, "Quantum humanities: a vision for quantum computing in digital humanities", *ICS Softw.-Inensiv. Cyber-Phys. Syst.* **35** (2020) 153. https://doi.org/10.1007/s00450-019-00419-4

[Bertels 2021] K. Bertels *et al.*, "Quantum Accelerator Stack: A Research Roadmap", *arXiv* 2102:02035 (2021). https://arxiv.org/abs/2102.02035

[Bobier 2021] J.-F. Bobier *et al.*, "What Happens When 'If' Turns to 'When' in Quantum Computing?", *Boston Consulting Group* report (2021). https://www.bcg.com/publications/2021/building-quantum-advantage

[Britt 2017] K.A. Britt and T.S. Humble, "High-Performance Computing with Quantum Processing Units", *J. Emerg. Technol. Comput. Syst.* **13** (2017) 39. https://doi.org/10.1145/3007651

[Cao 2019] Y. Cao *et al.*, "Quantum Chemistry in the Age of Quantum Computing", *Chem. Rev.* **119** (2019) 10856. https://doi.org/10.1021/acs.chemrev.8b00803

[Camps 2021] D. Camps, R. Van Beeumen, C. Yang, "Quantum Fourier Transform Revisited", *Numer. Linear Algebra Appl.* **28** (2021) e2331. https://doi.org/10.1002/nla.2331

[Castelvecchi 2018]. D. Castelvecchi, "The quantum internet has arrived (and it hasn't)", *Nature* **554** (2018) 289. https://doi.org/10.1038/d41586-018-01835-3

[Cerezo 2021] M. Cerezo *et al.*, "Variational quantum algorithms", *Nat. Rev. Phys.* **3** (2021) 625. https://doi.org/10.1038/s42254-021-00348-9

[Chiribella 2021] G. Chiribella, Y. Yang, R. Renner, "Fundamental Energy Requirement of Reversible Quantum Operations", *Phys. Rev. X* **11** (2021) 021014. https://doi.org/10.1103/PhysRevX.11.021014

[Cho 2020] A. Cho, "The biggest flipping challenge in quantum computing", *Science* news feature. https://doi.org/10.1126/science.abd7332

[Cincio 2021] L. Cincio *et al.*, "Machine Learning of Noise-Resilient Quantum Circuits", *PRX Quantum* **2** (2021) 010324. https://doi.org/10.1103/PRXQuantum.2.010324

[D-Wave] https://www.dwavesys.com/

[D-Wave 2017] "Computational Power Consumption and Speedup", *D-Wave White Paper Series* 14-1005A-D (2017) https://www.dwavesys.com/media/ivelyjij/14-1005a_d_wp_computational_power_consumption_and_speedup.pdf

[Daiss 2021] S. Daiss *et al.*, "A quantum-logic gate between distant quantum-network modules", *Science* **371** (2021) 614. https://doi.org/10.1126/science.abe3150

[Dalyac 2021] C. Dalyac *et al.*, "Qualifying quantum approaches for hard industrial optimization problems. A case study in the field of smart-charging of electric vehicles", *EPJ Quantum Technol.* **8** (2021) 12. https://doi.org/10.1140/epjqt/s40507-021-00100-3

[Dervovic 2018] D. Dervovic *et al.*, "Quantum linear systems algorithms: a primer", *arXiv* 1802.08227 (2018) https://arxiv.org/abs/1802.08227

[Deutsch 1985] D. Deutsch, "Quantum theory, the Church–Turing principle and the universal quantum computer", *Proc. R. Soc. Lond. A* **400** (1985) 97. http://doi.org/10.1098/rspa.1985.0070

[Devitt 2013] S.J, Devitt, W.J. Munro, K. Nemoto, "Quantum error correction for beginners." *Rep. Prog. Phys.***76** (2013) 076001. https://doi.org/10.1088/0034-4885/76/7/076001

[DiVincenzo 2000] D.P. DiVincenzo, "The physical implementation of quantum computation." *Fortschr. Phys.* **48** (2000) 771-783. DOI: https://doi.org/10.1002/1521-3978(200009)48:9/11%3C771::AID-PROP771%3E3.0.CO;2-E

[Doherty 2021] M. Doherty, "Quantum accelerators: a new trajectory of quantum computers", *Digitale Welt* **5** (2021) 74. https://doi.org/10.1007/s42354-021-0342-8

[Dunjko 2020] V. Dunjko, P. Wittek, "A non-review of Quantum Machine Learning: trends and explorations", *Quantum Views* **4** (2020) 32. https://doi.org/10.22331/qv-2020-03-17-32

[Emani 2021] P.S. Emani *et al.*, "Quantum computing at the frontiers of biological sciences", *Nat. Methods* **18** (2021) 701. https://doi.org/10.1038/s41592-020-01004-3

[ENISA 2021] "Post-Quantum Cryptography: Current state and quantum mitigation", European Union Agency for Cybersecurity (2021). https://www.enisa.europa.eu/publications/post-quantum-cryptography-current-state-and-quantum-mitigation

[ETHZ 2021] "ETH Zurich and PSI found Quantum Computing Hub", ETH Zürich press release (2021). https://ethz.ch/en/news-and-events/eth-news/news/2021/05/eth-zurich-and-psi-found-quantum-computing-hub.html

[EuroCC] https://www.eurocc-access.eu/

[EuroQCI] The European Quantum Communication Infrastructure (EuroQCI) Initiative. https://digital-strategy.ec.europa.eu/en/policies/european-quantum-communication-infrastructure-euroqci

[ESA 2021] "Quantum communication in space moves ahead", *European Space Agency.* https://www.esa.int/Applications/Telecommunications_Integrated_Applications/Quantum_communication_in_space_moves_ahead

[Fu 2016] X. Fu *et al.*, "A Heterogeneous Quantum Computer Architecture", *Proceedings of the ACM International Conference on Computing Frontiers* (2016). https://doi.org/10.1145/2903150.2906827

[Fösel 2021] T. Fösel *et al.*, "Quantum circuit optimization with deep reinforcement learning", *arXiv* 2103.07585 (2021). https://arxiv.org/abs/2103.07585

[Gambetta 2020] J. Gambetta, "IBM's roadmap for scaling quantum technology", *IBM Research* blog (2020). https://www.ibm.com/blogs/research/2020/09/ibm-quantum-roadmap

[Gehring 2021]. T. Gehring *et al.*, "Homodyne-based quantum random number generator at 2.9 Gbps secure against quantum side-information", *Nat. Commun.* **12** (2021) 605. https://doi.org/10.1038/s41467-020-20813-w

[Genois 2021] E. Genois *et al.*, "Quantum-tailored machine-learning characterization of a superconducting qubit", *arXiv* 2106.13126 (2021). https://arxiv.org/abs/2106.13126

[Georgescu 2014] I.M. Georgescu, S. Ashhab, F. Nori, "Quantum simulation", *Rev. Mod. Phys.* **86** (2014) 153. https://doi.org/10.1103/RevModPhys.86.153

[Gisin 2007] N. Gisin, R. Thew, "Quantum communication", *Nature Photon.* **1** (2007) 165. https://doi.org/10.1038/nphoton.2007.22

[Google 2021] Google Quantum AI, "Exponential suppression of bit or phase errors with cyclic error correction", *Nature* **595** (2021) 383. https://doi.org/10.1038/s41586-021-03588-y

[Giri 2017] P.R. Giri, V.E. Korepin, "A review on quantum search algorithms", *Quantum Inf. Process.* **16** (2017) 315. https://doi.org/10.1007/s11128-017-1768-7

[Greganti 2021] C. Greganti *et al.*, "Cross-Verification of Independent Quantum Devices", *Phys. Rev. X* **11** (2021) 031049. https://doi.org/10.1103/PhysRevX.11.031049

[Harrow 2009] A.W. Harrow, A. Hassidim, S. Lloyd, "Quantum Algorithm for Linear Systems of Equations", *Phys. Rev. Lett.* **103** (2009) 150502. https://doi.org/10.1103/PhysRevLett.103.150502

[Hazan 2020] E. Hazan *et al.*, "The next tech revolution: quantum computing", *McKinsey & Company* report (2020). https://www.mckinsey.com/fr/our-insights/the-next-tech-revolution-quantum-computing

[Henriet 2020] L. Henriet *et al.*, "Quantum computing with neutral atoms", *Quantum* **4** (2020) 327. https://doi.org/10.22331/q-2020-09-21-327

[Hillmann 2020] T. Hillmann *et al*., “Universal Gate Set for Continuous-Variable Quantum Computation with Microwave Circuits”, *Phys. Rev. Lett.* **125** (2020) 160501. https://doi.org/10.1103/PhysRevLett.125.160501

[Huang 2021] H.-Y. Huang *et al.*, “Power of data in quantum machine learning”, *Nat. Commun.* **12** (2021) 2631. https://doi.org/10.1038/s41467-021-22539-9

[Hunger 2021] D. Hunger, “Quantum logic at a distance”, *Science* **371** (2021) 576. https://doi.org/10.1126/science.abg1536

[Ikonen 2017] J. Ikonen, J. Salmilehto, M. Möttönen, “Energy-efficient quantum computing”, *npj Quantum Inf.* **3** (2017) 17. https://doi.org/10.1038/s41534-017-0015-5

[IQM] IQM Quantum Computers. https://www.meetiqm.com/

[IQT 2020] “Quantum Computing: A Seven-year Market Forecast”, *Inside Quantum Technology* report IQT-QCM-1020 (2020). https://www.insidequantumtechnology.com/product/quantum-computing-a-seven-year-market-forecast/

[Johnston 2019] E.R. Johnston, N. Harrigan, M. Gimeno-Segovia, “Programming quantum computers: essential algorithms and code samples”, O’Reilly Media (2019).

[Jordan 2021] S. Jordan, “Quantum Algorithm Zoo”, https://quantumalgorithmzoo.org

[Kung 2021] J. Kung, M. Fancy, “A Quantum Revolution: Report on Global Policies for Quantum Technology”, *CIFAR* report (2021). https://cifar.ca/cifarnews/2021/04/07/a-quantum-revolution-report-on-global-policies-for-quantum-technology/

[Loesekrug-Pietri 2021] A. Loesekrug-Pietri, “It’s not too late for Europe to lead the post-quantum cryptography race”, *Sifted* (2021). https://sifted.eu/articles/europe-post-quantum-cryptography/

[McArdle 2020] S. McArdle, S- Endo, A. Aspuru-Guzik, S.C. Benjamin, X Yuan, “Quantum computational chemistry”, *Rev. Mod. Phys.* **92** (2020) 015003. https://doi.org/10.1103/RevModPhys.92.015003

[MacQuarrie 2020] E.R. MacQuarrie *et al.*, “The emerging commercial landscape of quantum computing”, *Nat. Rev. Phys.* **2** (2020) 596. https://doi.org/10.1038/s42254-020-00247-5

[Merali 2011] Z. Merali, “First sale for quantum computing”, *Nature* **474** (2011) 18. https://doi.org/10.1038/474018a

[Mirrahimi 2014] M. Mirrahimi *et al.*, “Dynamically protected cat-qubits: a new paradigm for universal quantum computation”, *New J. Phys.* **16** (2014) 045014. https://doi.org/10.1088/1367-2630/16/4/045014

[Montanaro 2016] A. Montanaro, “Quantum algorithms: an overview”, *npj Quantum Inf.* **2** (2016) 15023. https://doi.org/10.1038/npjqi.2015.23

[Nautrup 2019] H.P. Nautrup *et al.*, “Optimizing Quantum Error Correction Codes with Reinforcement Learning”, *Quantum* **3** (2019) 215. https://doi.org/10.22331/q-2019-12-16-215

[NCSC 2020] “Quantum security technologies”, *National Cyber Security Centre* white paper (2020). https://www.ncsc.gov.uk/whitepaper/quantum-security-technologies

[OpenQKD] https://openqkd.eu/

[Parra-Rodriguez 2020] A. Parra-Rodriguez *et al.*, “Digital-Analog Quantum Computation”. *Phys. Rev. A* **101** (2020) 022305. https://doi.org/10.1103/PhysRevA.101.022305

[Pasqal] https://pasqal.io/

[Pirandola 2015] S. Pirandola *et al.*, “Advances in quantum teleportation”, *Nature Photon.* **9** (2015) 641. https://doi.org/10.1038/nphoton.2015.154

[Pogorelov 2021] I. Pogorelov *et al.*, “Compact Ion-Trap Quantum Computing Demonstrator”, *PRX Quantum* **2** (2021) 020343. https://doi.org/10.1103/PRXQuantum.2.020343

[PRACE] PRACE Training Portal. https://training.prace-ri.eu/

[Preskill, 2018] J. Preskill. "Quantum computing in the NISQ era and beyond." *Quantum* **2** (2018) 79. https://doi.org/10.22331/q-2018-08-06-79

[PQC.FI] Post-Quantum Cryptography Finland. https://www.pqc.fi/

[QIA] Quantum Internet Alliance. https://quantum-internet.team/

[QLM] Atos Quantum Learning Machine. https://atos.net/en/solutions/quantum-learning-machine

[QT.EU] European Quantum Flagship, https://qt.eu/about-quantum-flagship/projects/

[QuIC] The European Quantum Industry Consortium. https://quantumindustry.eu

[QUTAC] QUTAC - Quantum Technology & Application Consortium. https://www.qutac.de/

[Räsänen 2021] M. Räsänen *et al.*, "Path to European quantum unicorns", *EPJ Quantum Technol.* **8** (2021) 5. https://doi.org/10.1140/epjqt/s40507-021-00095-x

[Shor 1994] P.W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring", *Proceedings 35th Annual Symposium on Foundations of Computer Science* (1994) pp. 124-134. https://doi.org/10.1109/SFCS.1994.365700

[Sweke 2021] R. Sweke *et al.*, "Reinforcement learning decoders for fault-tolerant quantum computation", *Mach. Learn.: Sci. Technol.* **2** (2021) 025005. https://doi.org/10.1088/2632-2153/abc609

[Temkin 2021] M. Temkin, "Investors bet on the technologically unproven field of quantum computing", *Pitchbook* news & analysis (2021). https://pitchbook.com/news/articles/quantum-computing-venture-capital-funding

[Terhal 2015] B.M. Terhal, "Quantum error correction for quantum memories", *Rev. Mod. Phys.* **87** (2015) 307. https://doi.org/10.1103/RevModPhys.87.307

[Uvarov 2020] A.V. Uvarov, A.S. Kardashin, J.D. Biamonte, "Machine learning phase transitions with a quantum processor", *Phys. Rev. A* **102** (2020) 012415. https://doi.org/10.1103/PhysRevA.102.012415

[Van Meter 2013] R. Van Meter, C. Horsman, "A Blueprint for Building a Quantum Computer", *Commun. ACM* **56** (2013) 84. https://doi.org/10.1145/2494568

[Venegas-Andraca 2012] S.E. Venegas-Andraca. "Quantum walks: a comprehensive review", *Quantum Inf. Process.* **11** (2012) 1015–1106. https://doi.org/10.1007/s11128-012-0432-5

[Villalonga 2020] B. Villalonga *et al.*, "Establishing the quantum supremacy frontier with a 281 Pflop/s simulation", *Quantum Sci. Technol.* **5** (2020) 034003. https://doi.org/10.1088/2058-9565/ab7eeb

[WACQT] Wallenberg Centre for Quantum Technology. https://wacqt.se

[Wang 2021] P. Wang *et al.*, "Single ion qubit with estimated coherence time exceeding one hour", *Nat. Commun.* **12** (2021) 233. https://doi.org/10.1038/s41467-020-20330-w

[Wilen 2021] C.D. Wilen *et al.*, "Correlated charge noise and relaxation errors in superconducting qubits", *Nature* **594** (2021) 369. DOI: https://doi.org/10.1038/s41586-021-03557-5

[Wu 2021] Y. Wu *et al.* "Strong quantum computational advantage using a superconducting quantum processor", *arXiv* 2106.14734 (2021). https://arxiv.org/abs/2106.14734

[Zhong 2020] H.-S. Zhong *et al*. "Quantum computational advantage using photons", *Science* **370** (2020) 1460. https://dx.doi.org/10.1126/science.abe8770

14. List of acronyms

AI	Artificial Intelligence
CPU	Central Processing Unit
EU	European Union
GPU	Graphics Processing Unit
HPC	High Performance Computing
HPC+QC	(Hybrid) High Performance Computing and Quantum Computing
I/O	Input/Output
IP	Intellectual Property
kW	Kilowatt
LSQ	Large-Scale Quantum
ML	Machine Learning
MPP	Massively Parallel Processing
NISQ	Noisy Intermediate-Scale Quantum
OS	Operating System
PQC	Post-Quantum Cryptography
QC	Quantum Computing
QEC	Quantum Error Correction
QFT	Quantum Fourier Transform
QKD	Quantum Key Distribution
QLM	Quantum Learning Machine
QML	Quantum Machine Learning
QUBO	Quadratic Unconstrained Binary Optimisation
QPU	Quantum Processing Unit
QRNG	Quantum Random Number Generator
SME	Small and Medium-sized Enterprise
TRL	Technology Readiness Level

Acknowledgements

This work was financially supported by the PRACE project funded in part by the EU's Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 823767.

Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

Security in an evolving European HPC Ecosystem

Dirk Pleiter[a*], Sebastien Varrette[b], Ezhilmathi Krishnasamy[b],
Enver Özdemir[c], Michał Pilc[d]

[a]PDC Center for High Performance Computing, KTH, Sweden
[b]Department of Computer Science, Université du Luxembour, Luxembourg
[c]Informatics Institute, Istanbul Technical University, Turkey
[d]Poznan Supercomputing and Networking Center, Poznan, Poland

Abstract

The goal of this technical report is to analyse challenges and requirements related to security in the context of an evolving European HPC ecosystem, to provide selected strategies on how to address them, and to come up with a set of forward-looking recommendations. A key assumption made in this technical report is that we are in a transition period from a setup, where HPC resources are operated in a rather independent manner, to centres providing a variety of e-infrastructure services, which are not exclusively based on HPC resources and are increasingly part of federated infrastructures.

1 Introduction

In recent years, security incidents did have a severe impact on the availability of several high-end HPC systems throughout Europe. Securing HPC systems can be expected to become more challenging due to increasing complexity of the ecosystem, in which these systems are being operated. HPC systems are increasingly often operated in conjunction with other compute resources, e.g. private cloud instances, as well as different storage resources. Additionally, the e-infrastructure services, through which these resources are accessed, are becoming federated. This evolution of the HPC ecosystem is one of the reasons for rethinking the approaches to managing security and putting information security management systems in place. Also, the changing risk assessment due to an increased risk of cyber-attacks as well as the need for a higher level of protection, e.g. in the context of the processing of sensitive data, has to be taken into account.

We follow in this technical report the language used in the EU regulation 2019/881 [EU2019] and use the term security (or cybersecurity) to refer to any activities necessary to protect network and information systems, the users of such systems, and other persons affected by cyber threats.

This technical report is organised as follows: In Section 2 we describe the expected evolution of the European HPC ecosystem and provide a reminder of the main types of threats. We continue with a discussion of a selected set of approaches to improve security in Section 3 and an analysis of how security standards can be leveraged in Section 4. This is followed by a summary and a set of recommendations in Section 5.

This technical report is part of a series of reports published in the Work Package "HPC Planning and Commissioning" (WP5) of the PRACE-6IP project. The series aims to describe the state-of-the-art and mid-term trends of the technology and market landscape in the context of HPC and AI, edge-, cloud- and interactive computing, Big Data and other related technologies. It provides information and guidance useful for decision makers at different levels: PRACE aisbl, PRACE members, EuroHPC sites and the EuroHPC advisory groups "Infrastructure Advisory Group" (INFRAG) and "Research & Innovation Advisory Group" (RIAG) and other European HPC sites. Further reports published so far on the PRACE webpage[1] cover "State-of-the-Art and Trends

[*] Corresponding author email address: pleiter@kth.se

[1] https://prace-ri.eu/infrastructure-support/market-and-technology-watch/

for Computing and Network Solutions for HPC and AI", "Data Management Services and Storage", "Edge Computing: An Overview of Framework and Applications", "Quantum Computing - A European Perspective" and "User Requirements influencing HPC Technologies".

2 Background

2.1 Evolving HPC infrastructures

In this subsection, we consider some trends in HPC that have a strong impact on future approaches to security.

In the past, the primary role of HPC centres was to operate one or more supercomputers and to provide access to this resource. In the future, we expect these centres to transform to service providers, where the latter are based on different types of underlying compute and storage resources. While this will continue to include supercomputers, an increasing number of centres have also deployed, e.g. on-premise private cloud instances. Most of the new EuroHPC pre-exascale and petascale systems do include such instances. These services started to become harmonised and federated, e.g. in the context of the Fenix initiative[2]. Meanwhile the trends sketched here have meanwhile become part of the future EuroHPC strategy, as EuroHPC added "HPC Federation and Services" as one of the five pillars of future activities.[3]

With these architectural changes, HPC centres respond to changing user needs as well as those of emerging new science and engineering domains, which do need HPC resources for their research. A first trend to highlight are efforts towards establishing domain-specific platform services that allow end-users to run HPC workflows through open portal services. Such platform services are being implemented among others by projects like the Human Brain Project[4] or different ESFRIs like those organised in the European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (ESCAPE)[5] or the ESFRI for life-science information Elixir[6]. Furthermore, we observe a growing demand for using HPC infrastructures for storing and processing sensitive data like personal data, where compliance with the EU General Data Protection Regulation (GDPR) is mandatory [EU2016]. Finally, HPC infrastructures have to become ready for more support of users coming from private organisations, e.g. industry or SMEs.

These trends require rethinking our approach to security in the context of HPC infrastructures for multiple reasons. Firstly, workflows will run on HPC resources, where access is restricted, but are connected to and triggered through relatively openly accessible platform services like web portals. Secondly, these workloads and platform services need to be able to run on top of federated e-infrastructure services provided by different sites. Therefore, a harmonisation of the security levels provided by different sites is required. Finally, processing and storing of sensitive data as well as service offerings for users from private organisations lead to higher demand for security to ensure the protection of data and to ensure a high level of confidentiality for instance to protect trade secrets.

The HPC community can benefit from the important role that security is playing for commercial suppliers of Cloud e-infrastructures, where the commercial impact of security breaches is more obvious. Security incidents may not only result in a loss of customers, but liability obligations may have severe financial consequences.

2.1 Main concepts and threats

HPC facilities are exposed to threats inherent to digital communications when interacting with such large-scale computing systems. The formalization and mitigation of these threats are widely addressed in the literature tied to the cryptology and network security domains [Dumas2015]. The main functionalities aimed to be guaranteed in this context are traditionally summarized using the acronym CAIN: Confidentiality, Authentication, Integrity, Non-repudiation.

In digital communication systems, data ***confidentiality*** means that data is only accessible to a specific set of users. Confidentiality has become more important as more sensitive data is being processed, e.g. in the context of life sciences. Typical measures to support confidentiality are to provide mechanisms for restricting access to data, e.g. by means of access control lists (ACL), or by encrypting data using keys. The encryption can be symmetric or

[2] https://www.fenix-ri.eu

[3] European Commission, "Equipping Europe for world-class High Performance Computing in the next decade", SWD(2020) 179 final, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020SC0179&rid=9.

[4] See, e.g. EBRAINS Simulation services offering (https://ebrains.eu/services#category2)

[5] https://projectescape.eu/

[6] https://elixir-europe.org/

asymmetric. Symmetric key algorithms employ a single secret key which is used for both encryption and decryption, while asymmetric key algorithms, like public-key algorithms, use different keys for encryption and decryption. In the context of HPC, encryption can be challenging for different reasons. The performance impact of data encryption and decryption can be significant, particularly in extreme-scale data sets. Furthermore, there are no established and widely deployed solutions for distributing keys.

Authentication is the process where an entity (the Principal) is proving its identity to another entity (the System) that must be able to detect identity theft. This is a challenging property for federated authentication services foreseen to become standard within the European digital ecosystem. Any service provider has to be able to manage the identity of individuals (staff members, platform users, business clients) or systems and cope with the threat of identities being compromised.

Integrity of data is about ensuring data to be never altered or partially removed. With the increased amount of data stored in HPC centres even the risks related to unintended data corruption due to failures of the storage technologies is increasing. For being protected against data integrity threats it might be necessary to keep hash functions in a separate safe location. Also providing the option of keeping redundant copies of data may help preserve integrity of data.

Non-repudiation is required to provide protection against an individual falsely denying having performed a particular action. It may, in particular, involve the capability to determine whether a given individual took a particular action such as creating information, sending a message, approving information, or receiving a message [NICCS2021].

The Cloud Security Alliance (CSA), which is an organisation that works on defining and raising awareness of best practices related to security in the area of cloud computing, performed a survey a few years ago among experts to compile opinions on the most important security issues within cloud computing [CSA2016]. The identified threats do also apply to HPC data centres. As the most important security concern, the CSA report lists data breaches, i.e. incidents that lead to sensitive, protected or confidential information to be released or stolen. Next, insufficient identity, credential, and access management are listed, followed by insecure interfaces and APIs.

3 Approaches to improving security

In this section, we collect a selected set of approaches to improve security in the context of infrastructures that comprise HPC resources.

3.1 Access and identity management

3.1.1 Use of cryptographic protocols and SSH

The users of HPC centres access and use supercomputer facilities remotely. Securing such communications over an unsecured network relies on cryptographic protocols. The fundamental objective of cryptography is to enable two entities, traditionally called Alice and Bob, to communicate through an insecure channel in such a way that any opponent, Oscar, having access to the information circulating on the communication channel is not able to understand what is exchanged [Dumas2015].

There exist two main types of cypher schemes ensuring the confidentiality of the exchanged messages: secret key (or symmetric) cryptography which mimics perfect secrecy where the same key is a shared secret between the sender and the recipient and is used for both encryption and decryption of messages. In such a scheme, there is only one key that each user utilizes; therefore, this key has to be already known to all users. The messages to be sent to the other party are first encrypted by this key. When the encrypted text or cypher text is transmitted to other users, they decrypt the cypher text with the same key and recover the original messages. Symmetric key cryptosystems can be almost secure and efficient in terms of computation time. However, before using such a scheme, how can the key be distributed to all communicating parties?

This is the place where public key (or asymmetric) cryptography is needed: it eases key exchanges (typically through key exchange protocols) since no secret information is shared a priori among the involved protagonists. In the public key encryption method, unlike symmetric key algorithms, each communicating party has 2 different keys; one public key and one private key. Encryption is done with the user's public key and decryption is only performed with the same user's private key. The sender encrypts the message to be transmitted with the public key of the receiver. The public key is known to everyone but it does not violate the security of communication as the sent text can only be decrypted with the recipient's private key. A symmetric key algorithm is much less costly than a public key algorithm.

These concepts are implemented to secure the interactions with HPC centres, from the Secure Shell protocol (SSH) guiding the remote access to the supercomputers and described in the next section, to SSL protecting the web portal and HPC management services. For instance, SSH employs a public key algorithm to authenticate users and to distribute a shared key for symmetric key encryption. Thus the user's side SSH application encrypts the username and (the hash value of) the password with the public key of remote servers and in this case, the only way to decrypt the cypher requires having the private key of the remote server. Once the remote user is authenticated by the host, the distribution of the shared symmetric key is also performed via a public key algorithm. In other words, the shared key is encrypted with the recipient's public key then the cypher text is sent to the intended receiver.

3.1.2 Improving security of SSH-based access

By default, access to HPC facilities is enabled through Secure Shell (SSH) communication and encryption protocol (version 2). The PRACE service catalogue does foresee SSH as a "core service". This encrypted network protocol is used to log into another computer over an unsecured network, to execute commands in a remote machine, and to move files from one machine to another in a secure way. It is also used as the default medium to secure communications to remote servers with OpenSSH[7] , the most popular implementation used on nearly all systems. Most servers related to an HPC system have an SSH daemon running, for instance to allow system administrators to connect and manage the system.

Since SSH has such an important function on the HPC ecosystem, and as firewalls are often opened up to allow traffic, proper hardening of the service is needed. Nowadays, this protection is handled according to three main axes:

- Proper security hardening of the SSH configuration to reduce known weaknesses, which can, in particular be achieved through updates of /etc/ssh/sshd_config. Among the traditional recommended configuration settings, the following changes should be enforced:
    - DNS hostname checking
    - Disabling the password-based authentication and allow public key authentication (as using public key authentication is considered much safer and less prone to brute-force attacks). The way SSH handles the keys and the configuration files is illustrated in Figure 1:

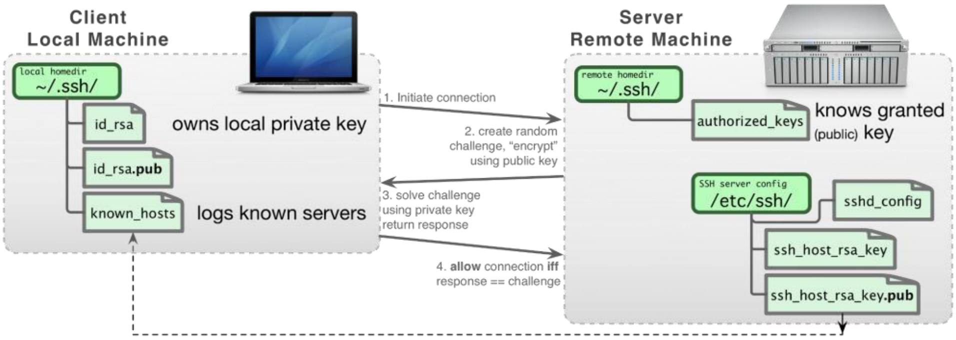


Figure 1: Illustration of the way SSH handles the keys and the configuration files

    - Disabling the root login (direct root logins may result in bad accountability of the actions performed by this user account). If possible, i.e. when not all users should have access to the system, it is advised to explicitly white-list the allowed users and groups of users, using the "default deny" principle
    - Restricting the number of authentication attempts
    - Restricting the cryptographic cyphers, key exchange algorithms and Message Authentication Code (MAC) functions to only the most secure ones recommended by NIST
    - If possible, the default listening port for the SSH service should be changed

[7] https://www.openssh.com/

References and resources that include examples to define highly secured configurations are provided in the appendix.

Additional traditional restrictions should be implemented. This can be facilitated by traffic filtering. An easy mechanism provided by the SSH protocol is to make the use of a from="pattern-list" clause in a list of authorised keys mandatory.

Furthermore, the usage of SSH configuration scanners and security tools are encouraged to be performed in a regular (and automated) way. This includes:

- Lynis[8], an open source security tool designed as a security scanner and compliance auditing tool. It can detect vulnerabilities and configuration flaws, but also provides recommendations for an in-depth audit and continuous improvement
- ssh_scan[9], a SSH configuration and policy scanner designed by the Mozilla foundation and thus kept up-to-date with the latest security hardening best practices.
- ssh-audit[10], while slightly outdated, allows to perform a test on a selected set of remote targets to analyse the responses it receives.

- Enabling Firewall to restrict and monitor SSH traffic, coupled with protective tools against brute-force attacks such as Fail2ban[11].
- Consider general system hardening by enabling Security-Enhanced Linux (SELinux) [SELinux14] is also recommended. SELinux is a security architecture for Linux systems designed to finely control and define access controls for the applications, processes, and files on a system. It uses security policies, which are a set of rules that tell SELinux what can or can’t be accessed, to enforce the access allowed by a policy.

Finally, instead of allowing file-based authorised keys under the control of users, using a centralized Identity Management (IdM) service (yet potentially distributed and replicated across multiple servers to allow for high availability) could also be considered. Such a service is used to create identity stores, centralized authentication, domain control for Kerberos and DNS services, and authorization policies. It is a central component within HPC sites to allow for federated authentication services as expected in reference European projects and initiatives as PRACE, EuroHPC, Fenix[12], EOSC or even Eduroam. IdM services are traditionally handled by middleware systems such as FreeIPA and Redhat IdM, 389 Directory Server, Microsoft Active Directory, OpenLDAP. Within a HPC facility, Redhat IdM[13] or IPA[14] are native and reference Linux-based frameworks addressing this need. It offers a more secure and robust environment for handling user identity, roles and credentials (including SSH key pairs) than what was possible in the past with directory services and LDAP.

In practice, the System Security Services Daemon (SSSD) interacts as a client component of the centralized IdM service deployed within the HPC facility to handle in a transparent, consistent and secure way the authentication service on a given host based on (Host-Based Access Control) HBAC rules. It also caches the information stored in the remote directory server to provide identity, authentication and authorization services to the host machine (login server, HPC head or compute server, web or cloud portal etc.). This comes with several advantages:

- Novel secure authentication methods and schemes (Multi-Factor authentication (MFA), OICD etc.) can be transparently enabled to serve the full user community
- The potential overhead induced by federated authentication services when dealing with the trust delegation protocols and verifications can be offloaded to dedicated servers

3.2 Improving the network security architecture

HPC encompasses advanced computation on parallel systems, enabling faster execution of highly compute-intensive tasks which heavily rely on interconnect performance. For this reason, the main high-bandwidth low-latency network of an HPC facility relies on the dominant interconnect technology in the HPC market, i.e. InfiniBand (IB) or HPE/Cray Slingshot. However, as these networks are internal they are less affected by security

[8] https://cisofy.com/downloads/lynis/

[9] https://github.com/mozilla/ssh_scan

[10] https://github.com/arthepsy/ssh-audit

[11] https://github.com/fail2ban/fail2ban

[12] https://www.fenix-ri.eu

[13] Redhat Idm: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/linux_domain_identity_authentication_and_policy_guide/introduction

[14] FreeIPA: https://www.freeipa.org

threats. Additionally, technology solution providers offer advanced monitoring infrastructures (e.g. Nvidia/Mellanox Unified Fabric Manager (UFM)[15]) that allow to assess the security and the performance of these networks.

The main security concerns are tied to the IP networks available at all HPC facilities. To limit the risk of network intrusion and to control the access to sensitive information, a network segmentation and segregation policy must be defined and implemented. The key features expected by the network organization are:

1. Scalability, i.e. support of
    - Thousands of computing and data storage elements which may be physically distributed across different racks and/or server rooms
    - Many virtualised systems per computing element
    - Hundreds of services, both external (Internet-facing) and internal
    - Direct connection to hosted HPC high-throughput equipment
    - Random user data access and throughput patterns during the day (majority) and night (minority)
    - Constant administrative data throughput patterns during the night (on site / off site backups)
    - Service level agreements (SLA) with internal/external users
2. Isolation from the hosting entity's internal network. This is realised by a separation of the different flows and streams within dedicated Virtual Local Area Network (VLAN), with strict policies enforced at all levels:
    - Access network, aiming for DMZ (demilitarized zone) services i.e. accessible by users on one side, linking to the data/Infiniband networks on another side
    - Production network: meant for user-level data transfer and Internet access, in-band management
    - Management network: meant for out of band management of base hardware by the HPC operation team
    - Infiniband network: user-level very-high bandwidth, very-low latency data transfer

This assumes systematic rules for IP addressing being in a place such that there is no overlap with a range reserved by the hosting entity. In the physical implementation of the network, a systematic network cabling policy should also be enforced, which foresees identical and unique labels on both ends of each cable.

Figure 2 illustrates the implementation of an IP network for a large-scale HPC and Big Data analytics infrastructure following secure best-practices in terms of VLAN structure.

It is recommended to organize such a network as a 2-layer topology: one upper level (Gateway Layer) with routing, switching features, network isolation and filtering (ACL) rules and meant to aggregate only switches and routers. As can be seen, this is where the interfaces to the local organization network, the outside world (i.e. public internet), as well as external public or private partners (including HPC centres or other members of a federating structure) takes place. The bottom level (Switching Layer) is typically composed of core switches as well as the TOR (Top-of-rack) network equipment, meant to interface the HPC servers and compute nodes. Both layers define at least three isolated VLANs that allow structuring the different flows and streams and thus with different security ACLs: the sensitive DMZ VLAN for external accesses, either to the login nodes or to the storage equipment, the production VLAN serving user applications with the computing facility, the Management VLAN for the corresponding tasks of the operational team responsible for maintaining and administering the system. In addition, non-routed VLANs are recommended: one to serve applications that do not support natively the fast interconnect technology in place (Infiniband for instance) through an IP emulation layer in addition to the production network. Then cloud modules would benefit from a dedicated VLAN offering an overlay network isolating the virtual machines and services deployed within this module. Secure implementations as KVlan within Grid5000[16] or Chameleon[17] illustrate the capabilities of such VLANs.

[15] https://www.nvidia.com/en-us/networking/infiniband/ufm/

[16] https://www.grid5000.fr/

[17] https://chameleoncloud.readthedocs.io/

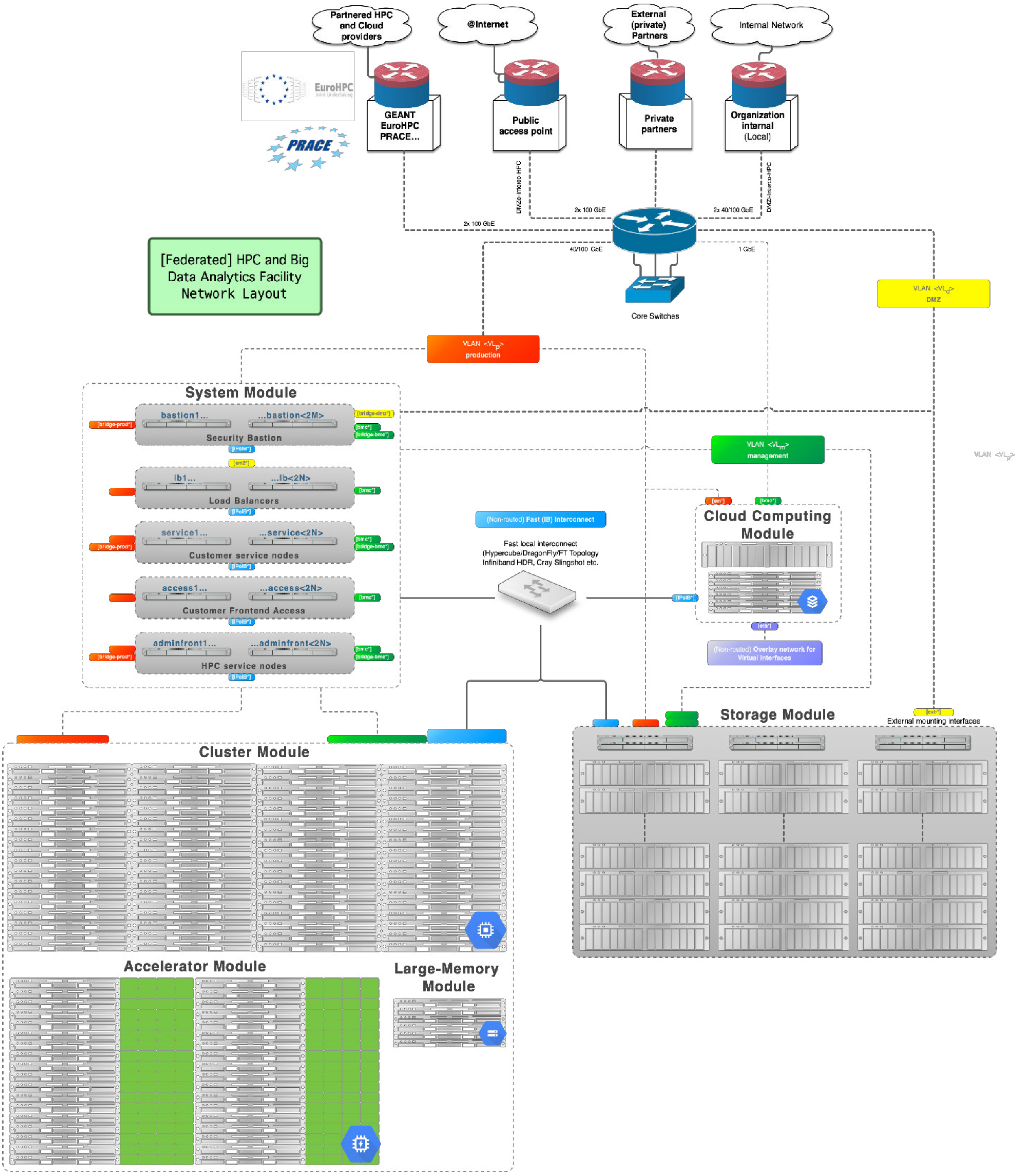


Figure 2: Schematic view of IP network for a large-scale HPC and Big Data analytics

3.3 Cybersecurity and Big Data Storage protection

With the embedded storage capacities of large-scale computing facilities, novel security challenges appear with, on the one hand the emerging paradigm of Open Science enabling an easier access to expert knowledge and material, and on the other hand the necessary compliance to the EU's General Data Protection Regulation (GDPR) [EU2016].

The interactions occurring during data processing activities are covered in the literature, for instance in [Paseri2021] and illustrated in Figure 3.

This illustrates the different types of data processing steps to be protected:

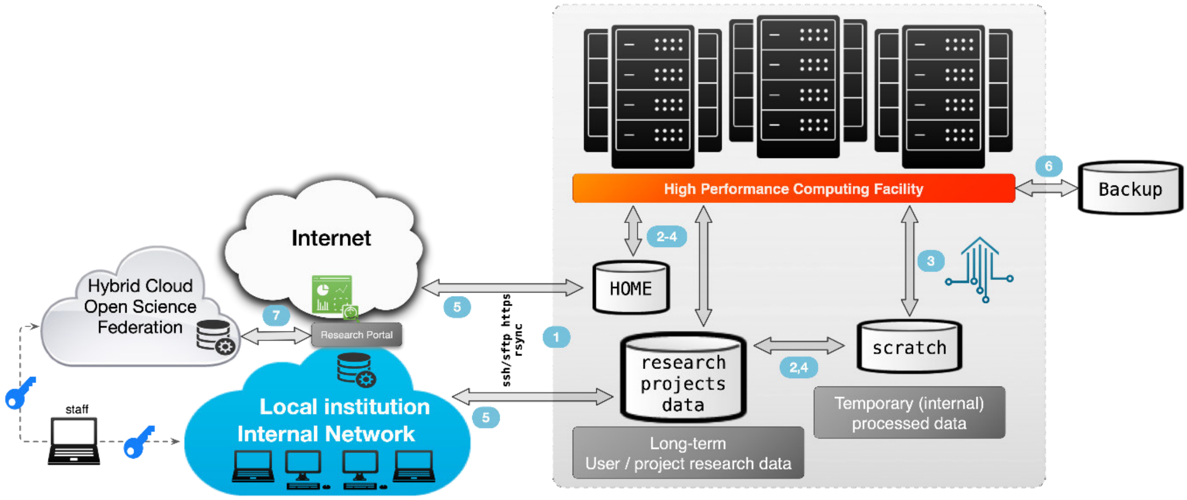


Figure 3: Illustration of data analytics workflow in a federated environment.

1. Data transfer of the input data towards the long-term storage area
2. Pre-processing phase to prepare the data for research analysis (this may include partial or total data transfer towards the internal scratch area)
3. Job processing on the computing facility, generating both intermediate and final data components
4. Post processing phase to derive scientific results from the processed data: this typically includes the creation of a metadata catalogue allowing to index and quickly recover scientific data
5. Data transfer of the output data towards external resources (for instance a laptop of an external server).
6. Archiving of the results, and backup of the long-term storage area
7. Data replication and synchronization may happen in a federated environment; in parallel, data sharing can be performed in the context of research collaborations; this includes live processing and access by external stakeholders

One of the major security challenges foreseen is represented by the complex tracking of data movements done in the steps 2 to 4 of the above figure. Parallel and distributed file systems used in HPC environments are not yet fully able to account and log internal data movements. More precisely, changelogs-based auditing capabilities relevant for the GDPR compliance are only featured starting with recently released versions of Lustre (2.11) and IBM Spectrum Scale (5.0). With regards to the other type of data transfer performed in the considered workflows (i.e. steps 1, 5 and 7), other accountability and monitoring mechanisms can be implemented to fulfil this constraint.

3.4 Introducing monitoring and intrusion detection systems

An intrusion detection system (IDS) is a device or software application that monitors a network or system for malicious activity or policy violations. Intrusion detection systems fall into one of three categories: Host Based Intrusion Detection Systems (HIDS) which typically monitors the integrity of important operating system files, Network Based Intrusion Detection Systems (NIDS) which analyses incoming network traffic to detect suspicious activities, and hybrids of the two. It is also possible to classify an IDS by detection approach. The most well-known variants are signature-based detection (recognizing bad patterns, such as malware) and anomaly-based detection (detecting deviations from a model of "good" traffic, which often relies on machine learning).

IDS in HPC environments is rather challenging since such large-scale facilities tend to have very distinctive modes of operation or be used for very distinctive purposes depending on the user profile. However, they are necessary to detect cyber-attacks against these systems, i.e. unauthorized access with a malicious intent to steal sensitive documents, compromise networks, vandalize the resources or use the resources for further malicious actions. Supercomputers across Europe were for instance recently infected (in May 2020) with cryptocurrency mining malware, forcing operators to shut the systems down to investigate the attack. Details were made public in an advisory from the European Grid Infrastructure (EGI) CSIRT about these cases, where it is claimed that compromised servers in Poland, Canada and China were used in these attacks[18]. The CERN Computer Security

[18] https://csirt.egi.eu/attacks-on-multiple-hpc-sites/

Team and other organizations identified the usage of the Kobalos malware which predates these incidents[19]. It is thus crucial to at least detect such malware within the implemented IDS systems.

To detect whether a system is compromised, also the use of machine learning methods for automated analysis of system behaviour has been explored [Peisert2017]. The idea was to search for indications where users were running unusual workloads.

To assess the current use of IDS solutions in European HPC centres, the three questions shown below were distributed among those centres that are organised in PRACE. Five HPC centres responded. For obvious reasons their identity is not disclosed here.

Question 1: What kind of IDS tools do you use?

Two out of five centres reported that they use a dedicated IDS. Both use AIDE[20], but in one case it is supplemented with additional dedicated scripts to detect unusual spikes in resource usage.

In the remaining three HPC centres alternative solutions are applied, namely:

1. IP address whitelisting on nodes connected to the public Internet implemented with a hardware firewall; in addition, all incoming and outgoing connections are logged and sent via rsyslog to a centralized logging server
2. No dedicated IDS; a hardware-based solution from Fortinet that is mainly used for establishing VPN connections but has built-in IDS functionalities.

Question 2: What is a typical configuration of your IDS? What parameters are monitored?

All responding sites that use an IDS system reported that malicious software signatures are monitored and tracked instantaneously. One of the HPC centres uses Ganglia for monitoring and Nagios to receive alerts and when CPU/memory/disk consumption exceeds predefined limits. Two sites reported that their systems are configured such that changes in the file system of management master nodes are observed. One site reported that they enabled tracking of login details (unsuccessful attempts, source IP addresses of successful and unsuccessful login).

Question 3: Does the IDS tool used at your site impact new services like OpenStack?

One response did not address this question in a clear manner. In all other responses no negative impact of IDS on services like OpenStack is reported. Two of them do not provide any explanation for this observation. In one case, the site believes that the lacking negative impact of using an IDS is because neither virtualisation nor any sophisticated network setups are being used. One site did highlight that their IDS solution is limited by the rate at which the data can be processed, which is currently 11 Gbps. They expect to have 52 Gbps in the future.

4. Leveraging security standards

In the previous sections it was shown that security issues have become more important and need to be addressed at a level that exceeds a single data centre as sites are starting to offer federated e-infrastructure services. One strategy to enhance security and establish common security levels is to leverage standards. In this section a number of such standards is reviewed.

4.1 NIST guides for conducting risk assessment

Dealing with security in the context of operating e-infrastructure services requires balancing the possible impact of threats on the one hand and the needs for openness and usability of these services as well as the costs associated with security measures on the other hand. This can be framed as a risk management process and the need for creating an environment, where decisions can be made based on an assessment of the risks. NIST [NIST2012] developed a guide for systematising the process of risk assessment. The latter should lead to an identification of all relevant threats, vulnerabilities to organizations, the impact that may occur and the probability that risks will materialise. More technical guidance on enhancing security of network and information systems is provided by NIST in a technical guide [NIST2008]. Here a number of approaches to security testing and examination are described. This includes, in particular, vulnerability validation techniques like penetration testing.

[19] https://www.welivesecurity.com/2021/02/02/kobalos-complex-linux-threat-high-performance-computing-infrastructure/

[20] https://aide.github.io/

4.2 ISO/IEC 27000 standards on information security management

The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) published a set of standards related to information security management, which are known as the ISO/IEC 27000 family of standards. Starting with the ISO/IEC 27001:2018 standard a general framework for defining information security management systems (ISMS) including a vocabulary [ISOIEC2018].

In this context the most relevant standard is the ISO/IEC 27001:2013 standard [ISOIEC2013]. The standard specifies the requirements for implementing and maintaining an effective information security management system (ISMS). As such it focuses on organisational aspects and does not prescribe specific technical measures that need to be taken. For example, the standard describes how leadership is to be involved, how planning should happen, what resources to allocate for support and how to regularly evaluate the implemented ISMS.

An important aspect of this standard is the option for certification. This allows organisations to document compliance with this standard based on a successful audit by an accredited certification body.

4.3 EU Cybersecurity Directives and Regulations

In the future, stipulations provided by the European Commission related to protection of network and information systems operated in EU member states will become more important. In 2016 the EU Parliament and Council confirmed the goal of achieving a high common level of security of network and information systems within the EU [EU2016b]. It considers the member states as the actors and focuses on services like online marketplaces, search engines and cloud computing services. In 2019 the EU Parliament and Council adopted a regulation on the establishment of a European Union Agency for Cybersecurity (ENISA)[21] [EU2019], i.e. the EU is starting to become an actor. ENISA is mandated with the creation and maintenance of a European cybersecurity certification framework and to work on the necessary technical ground for more specific standards and schemes for certification based on the framework provided by the aforementioned regulation. At this point, ENISA has not published standards yet that are of particular relevance in the context of this technical report.

4.4 German Federal Office for Information Security's catalogue C5

The Federal Office for Information Security in Germany (BSI)[22] introduced in 2016 a Cloud Computing Compliance Criteria Catalogue (C5). In 2020 an updated version of C5 was published [BSI2020]. BSI is an agency of the German federal government, which is in charge of managing computer and communication security for the German government. The C5 catalogue is intended to be used by different stakeholders to use the proposed criteria for a risk assessment. It is intended to be used by cloud service providers, customers and auditors.

The catalogue defines a set of base criteria that can be used as a checklist by relevant stakeholders including the operators of network and information systems. While the C5 catalogue mainly targets (commercial) cloud providers, many of these criteria can also be used for current and upcoming HPC infrastructures. For this reason, Fenix created the "Fenix Security Measures Catalogue" [Fenix2020] based on the C5 catalogue.

The criteria are grouped with an object defined for each of these groups of criteria. The following table shows a selection of such groups of criteria that are more important in this context:

Group of C5 criteria	Objective
Organisation of information security (OIS)	Plan, implement, maintain and continuously improve the information security framework within the organisation
Security policies and instructions (SP)	Provide policies and instructions regarding security requirements and to support business requirements
Personnel (HR)	Ensure that employees understand their responsibilities, are aware of their responsibilities with regard to information security, and that the organisation's assets are protected in the event of changes in responsibilities or termination.
Asset management (AM)	Identify the organisation's own assets and ensure an appropriate level of protection throughout their lifecycle.

[21] https://www.enisa.europa.eu/

[22] https://www.bsi.bund.de

Physical security (PS)	Prevent unauthorised physical access and protect against theft, damage, loss and outage of operations.
Identity and Access Management (IDM)	Secure the authorisation and authentication of users of the Cloud Service Provider (typically privileged users) to prevent unauthorised access.
Communication Security (COS)	Ensure the protection of information in networks and the corresponding information processing systems.
Security Incident Management (SIM)	Ensure a consistent and comprehensive approach to the capture, assessment, communication and escalation of security incidents.

4.5 French National Information Systems Security Agency's catalogue

The Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI)[23] is the French counterpart of BSI. In cooperation with BSI they published in 2018 "Prestataires de services d'informatique en nuage (SecNumCloud)" [ANSSI2018]. As it is rather similar to the C5 catalogue, no specific analysis of this security catalogue has been performed.

5. Summary and Recommendations

Starting from the observation that the European HPC is evolving towards a provisioning of a federated service portfolio for serving changing and emerging user needs, current security threats and different approaches for improving security have been reviewed. Some of these approaches like the use of intrusion detection systems would not affect the openness and usability of the provided e-infrastructure services. We argued that security cannot be addressed by assuming data centres to operate in isolation. Security standards could be leveraged to harmonise security levels and security-related restrictions to the benefit of the users of the future European HPC infrastructure.

Based on our analysis, we make the following recommendations:

1. Data centres should review their security strategies within the evolving European HPC ecosystem, where an increased number of services are provided beyond providing access to a supercomputer.
2. With SSH being currently the most widely used network protocol for connecting to an HPC system, SSH configurations should be hardened for security, e.g. by enforcing DNS hostname checking, disabling password-based authentication, restricting access through white-listing methods, or by using SSH configuration scanners.
3. Use of Intrusion Detection System should be further explored as relatively few sites seem to use these today.
4. Security stands for HPC centres should be adopted at European level and should be leveraged (or even be established) to realise a common security level within a European infrastructure where services start to be federated, which would improve the usability of this infrastructure by users with specific security requirements, e.g. in the context of processing of sensitive data. The C5 catalogue from the German Federal Office for Information Security is a promising starting point, because it prescribes concrete measures.
5. The collaboration between HPC centres in Europe should be strengthened to improve the response to security incidents, which in future are even more likely to affect more than one site. Such collaboration would also allow to harmonise security measures to avoid users having to deal with different security restrictions.

[23] https://www.ssi.gouv.fr/en/

Glossary

ACL	Access Control List
CAIN	Confidentiality, Authentication, Integrity, Non-repudiation
DNS	Domain Name Service
DMZ	Demilitarized Zone
ESFRI	European Strategy Forum on Research Infrastructures
GDPR	General Data Protection Regulation
HPC	High-Performance Computing
IDS	Intrusion Detection System
ISMS	Information Security Management System
VPN	Virtual Private Network

References

[ANSSI2018] L'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI), "Prestataires de services d'informatique en nuage (SecNumCloud)", June 2018, https://www.ssi.gouv.fr/uploads/2014/12/secnumcloud_referentiel_v3.1_anssi.pdf

[BSI2020] BSI, "Cloud Computing Compliance Criteria Catalogue – C5:2020", 2020, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/ComplianceControlsCatalogue/2020/C5_2020.pdf

[CSA2016] Cloud Security Alliance, "'The Treacherous Twelve' Cloud Computing Top Threats in 2016", 2016, https://cloudsecurityalliance.org/artifacts/the-treacherous-twelve-cloud-computing-top-threats-in-2016/

[Dumas2015] J.-G. Dumas, J.-L. Roch, E. Tannier, and S. Varrette, "Foundations of Coding: Compression, Encryption, Error-Correction", Wiley & Sons, 376 pages, ISBN 978-1-118-88144-6, 2015

[Fenix2020] Fenix Consortium, "Fenix Security Measures Catalogue", August 2020, https://fenix-ri.eu/

[EU2016] EU, "Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data", 2016, https://data.europa.eu/eli/reg/2016/679/oj

[EU2016b] EU, "Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union", 2016, http://data.europa.eu/eli/dir/2016/1148/oj

[EU2019] EU, "Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act)", 2019, https://eur-lex.europa.eu/eli/reg/2019/881/oj

[ISOIEC2013] ISO, IEC, "ISO/IEC 27001:2013: Information technology — Security techniques — Information security management systems — Requirements", 2013, https://www.iso.org/standard/54534.html

[ISOIEC2018] ISO, IEC, "ISO/IEC 27000:2018: Information technology — Security techniques — Information security management systems — Overview and vocabulary", 2018, https://www.iso.org/standard/73906.html

[NICCS2021] NICCS, "Cybersecurity Glossary", https://niccs.cisa.gov/about-niccs/cybersecurity-glossary (accessed on 28.09.2021)

[NIST2008] NIST, "Technical Guide to Information Security Testing and Assessment", SP 800-115, September 2008, https://doi.org/10.6028/NIST.SP.800-115

[NIST2012] NIST, "Guide for Conducting Risk Assessments", SP 800-30 Rev. 1, September 2012, https://doi.org/10.6028/NIST.SP.800-30r1

[Paseri2021] L. Paseri, S. Varrette, and P. Bouvry, "Protection of Personal Data in High Performance Computing Platform for Scientific Research Purposes," in Proc. of the EU Annual Privacy Forum (APF) 2021, 2021, vol. 12703, pp. 123–142.

[Peisert2017] Peisert, Sean. "Security in high-performance computing environments" Communications of the ACM 60.9 (2017): 72-80.

[SELinux14] Bill McCarty. 2004. “SELinux: NSA's Open Source Security Enhanced Linux”. O'Reilly Media, Inc. https://dl.acm.org/doi/10.5555/1096126

Acknowledgements

This work was financially supported by the PRACE project funded in part by the EU’s Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 823767.

Appendix

OpenSSH Server and Client secure configuration

Although most default OpenSSH settings that already good from a security perspective, we encourage further efforts for securing OpenSSH servers configurations (/etc/ssh/sshd_config) and adapt the settings according to best security practices. More specifically, we suggest HPC operational teams to follow the recommendations of the Security Assurance and Security Operations team of the Mozilla Foundation[24].

Security/Server Side TLS configuration

The configuration of TLS within all web-enabled services is a challenging task for HPC operational teams. To that end, the same team from the Mozilla Foundation propose a reference guide[25] which includes an SSL Configuration Generator[26].

[24] https://infosec.mozilla.org/guidelines/openssh.html

[25] Security/Server Side TLS: https://wiki.mozilla.org/Security/Server_Side_TLS#

[26] https://ssl-config.mozilla.org/